Predictive Models

مدل های پیش بینانه

dayche.com . گروه دایکه

دای‌کِه DAYCHE

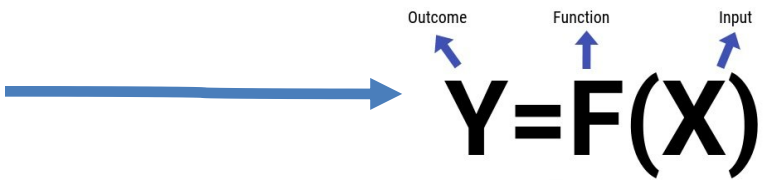
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ مدل های پیش بینانه

با رویکرد یادگیری با نظارت به دنبال یافتن تابعی از ویژگی های ورودی هستند تا مقدار فیلد هدف را با کمترین میزان خطا برآورد نمایند.

هریک از الگوریتم های مورد استفاده با استفاده از روش های خاص خود برای یافتن پاسخ به این هدف، نقاط قوت و ضعف خود را دارند.

X1	X2	X3	X4	Target
.	.	.	.	Y1
.	.	.	.	Y2
.	.	.	.	Y3
.	.	.	.	Y4
.	.	.	.	Y5
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.



فرآیند داده کاوی

مدل های پیش بینانه

- انواع الگوریتم های پیش بینانه
  - بر اساس توزیع آماری فیلد هدف

for Classification | for Regression | for Both

  - بر اساس نوع رابطه بین ویژگی ها و فیلد هدف

for Linear | for Non-Linear

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه

- انواع الگوریتم های پیش بینانه
  - بر اساس تفسیرپذیری الگوها

White-Box

Black-Box

  - بر اساس ساختار و فرمت خروجی

Rule Based

Bayesian Prob.

Statistical Models

Similarity Based

Complex Computation

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه

❑ انواع الگوریتم های پیش بینانه

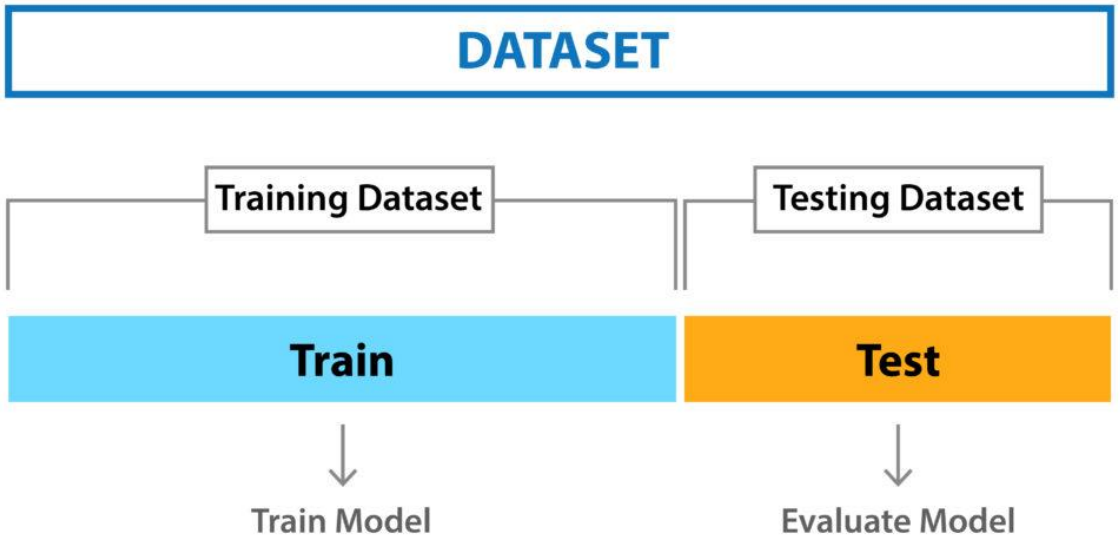
Algorithms	Reg. / Class.	Lin. / Non-Lin.	Interpretability	Output Structure
C5.0	Classification	Non-Linear	White-Box	Rule Based (DT)
Linear Regression	Regression	Linear	White-Box	Statistical Models
SVM	Classification	Both	Black-Box	Complex Computation
KNN	Both	Non-Linear	Black-Box	Similarity Based
MLP Neural Network	Both	Non-Linear	Black-Box	Complex Computation
Naïve Bayes	Classification	Non-Linear	White-Box	Bayesian Probability
Logistic Regression	Classification	Linear	White-Box	Statistical Models
LDA	Classification	Linear	White-Box	Statistical Models
CART	Both	Non-Linear	White-Box	Rule Based (DT)

فرآیند داده کاوی

مدل های پیش بینانه

❑ طرح آزمون برای مدلسازی

رایج ترین طرح آزمون، افراز کاملا تصادفی مجموعه داده ها به دو بخش داده های آموزشی و داده های آزمایشی می باشد.



روش Holdout

- **داده های آموزشی**

برای آموزش و ساخت مدل بکار می رود.

- **داده های آزمایشی**

برای ارزیابی کیفیت مدل ساخته شده بکار می رود.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ طرح آزمون برای مدلسازی

دو ایراد مهم به طرح Holdout وارد است:

- امکان ناپایداری نتایج ارزیابی و وابستگی مدل به انتخاب داده های آموزشی و آزمایشی

راهکار: تکرار روش Holdout و استفاده از برآیند آنها / استفاده از روش اعتبارسنجی متقابل

- عدم قابلیت تنظیم پارامتر مستقل از داده های آزمایشی

راهکار: استفاده از تقسیم بندی سه تایی داده ها (آموزش، اعتبارسنجی و آزمایش)

فرآیند داده کاوی

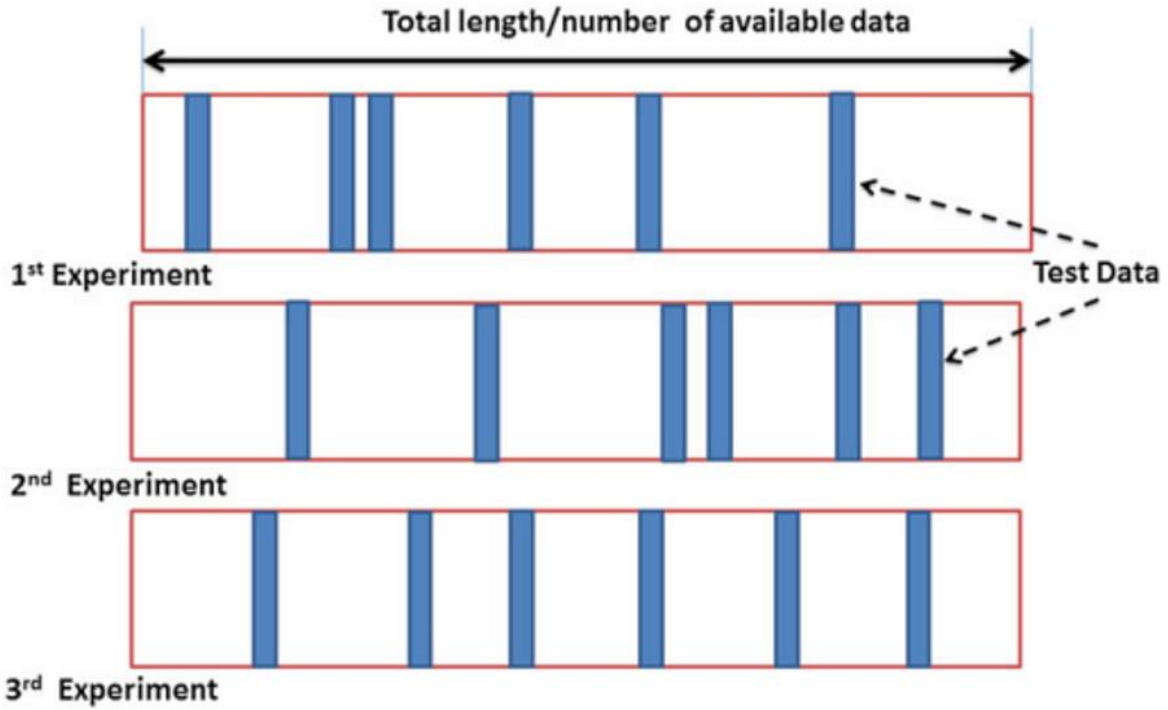
مدل های پیش بینانه

❑ طرح آزمون برای مدلسازی

یکی از ایرادانی که به روش Holdout وارد بود، امکان ناپایداری نتایج به علت انتخاب تصادفی نمونه های آموزشی و آزمایشی بوده است.

- روش Random Subsampling

در این روش به تعداد K بار، روش Holdout تکرار می شود و سپس میزان دقت مدل از میانگین دقت K مدل بدست آمده محاسبه می شود.



ارزیابی مدل با این روش می تواند اطمینان بیشتری از کیفیت مدل در جهت رسیدن به الگوهای اصلی داده ها در اختیار قرار دهد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

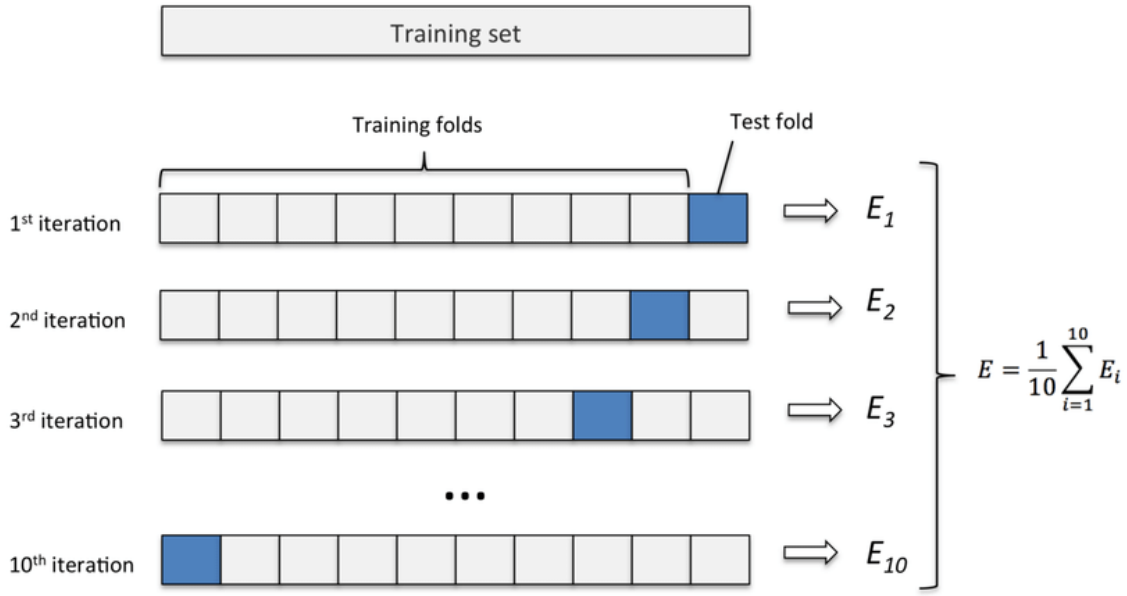
طرح آزمون برای مدلسازی

به علت انتخاب تصادفی داده های آموزشی و آزمایشی در روش Random Subsampling، از ظرفیت کامل داده ها در آموزش، ساخت و ارزیابی مدل ها استفاده نمی شود.

- **روش اعتبارسنجی متقابل Cross Validation**

در این روش با تقسیم مجموعه داده ها به K قسمت برابر، در هر مرتبه یکی از آنها به عنوان داده آزمایشی و مابقی به عنوان داده های آموزشی برای ساخت مدل استفاده می شوند.

ارزیابی مدل با این روش از تمامی داده های در دسترس جهت ساخت مدل و ارزیابی استفاده می کند؛ ولی در داده های زیاد نیاز به زمان و محاسبات بیشتری خواهد داشت.



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

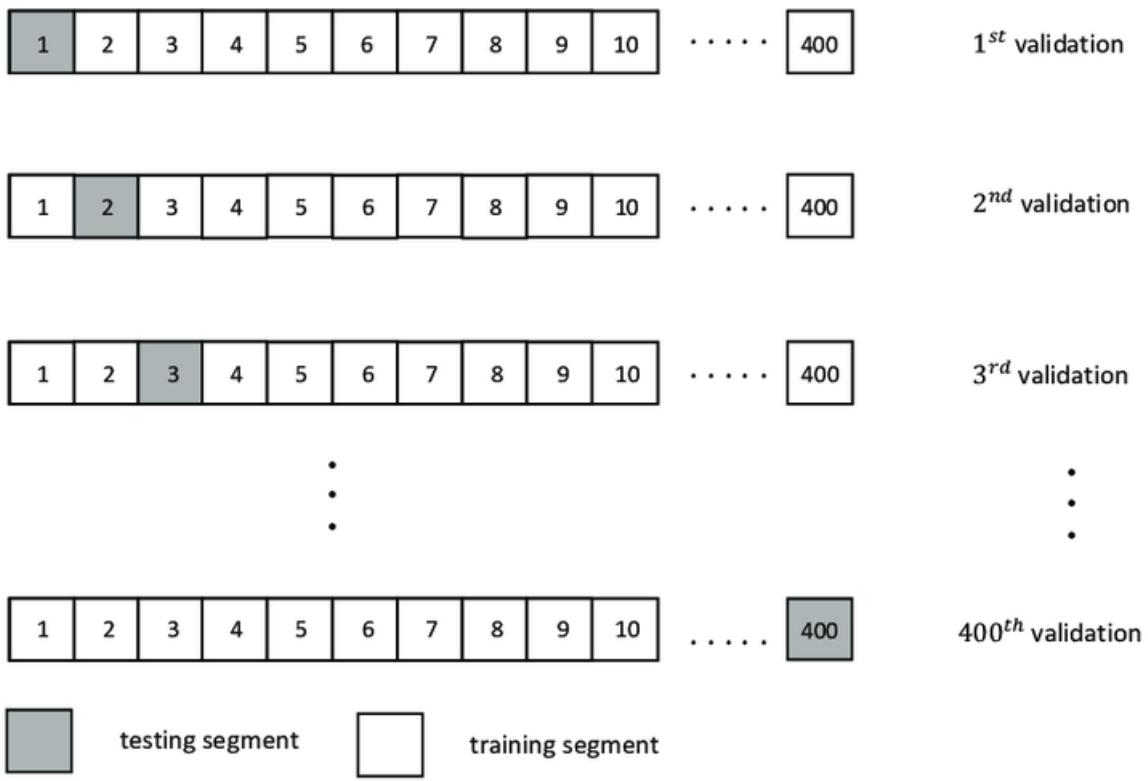
فرآیند داده کاوی

مدل های پیش بینانه

- طرح آزمون برای مدلسازی
  - روش اعتبارسنجی متقابل Leave One Out (LOO)

نوع خاصی از روش اعتبارسنجی متقابل می باشد که در آن تعداد K برابر با تعداد نمونه ها می باشد. در این حالت به تعداد نمونه های در مجموعه داده ها بایستی مدل ساخته شود و هر بار با n-1 رکورد مدل آموزش داده شده و با یک رکورد تست می شود.

این رویکرد در مواردی که تعداد رکوردها کم باشد، گزینه خوبی برای ارزیابی مدل هست تا علاوه بر اینکه همه داده ها در ارزیابی مدل مشارکت داشته باشند، بلکه آموزش و ساخت مدل نیز از حداکثر داده های موجود استفاده کند.



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه

❑ طرح آزمون برای مدلسازی

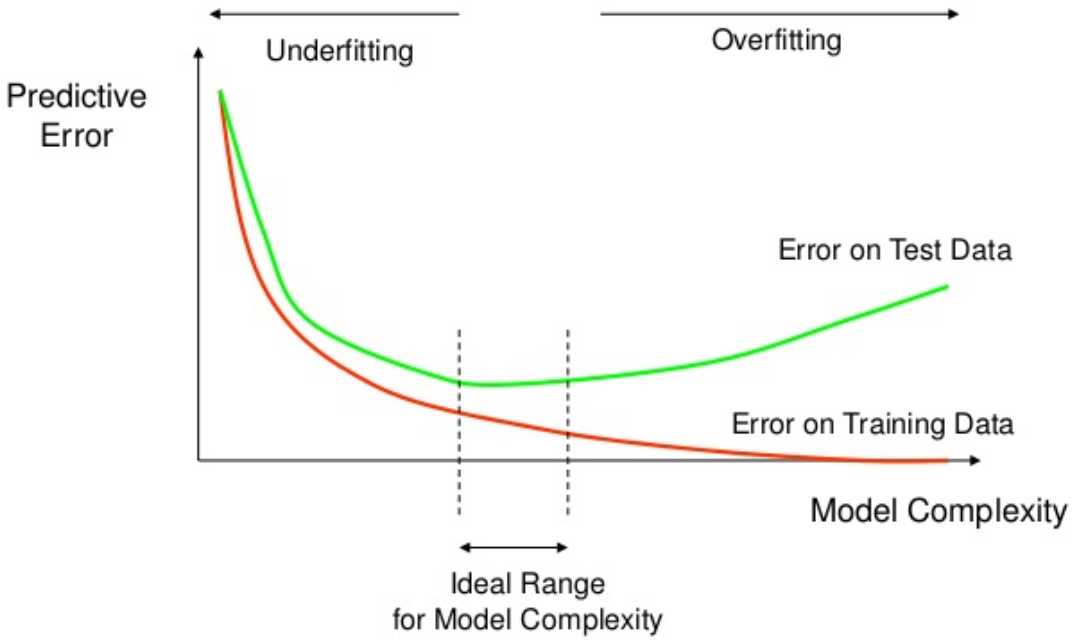
ایراد دوم به طرح آزمون Holdout تنظیم پارامترهای مدل بر اساس نتایج ارزیابی روی داده های آزمایشی می باشد. میزان پیچیدگی یک مدل با کنترل و تنظیم پارامترهای آن، بر اساس مقایسه میزان خطای مدل در مجموعه داده های آموزشی و آزمایشی تعیین می گردد.

- **بیش برازشی (Overfitting)**

پیچیدگی زیاد مدل، منجر به حفظ کردن داده های آموزشی و عدم شناسایی الگوهای تعمیم پذیر می گردد.

- **کم برازشی (Underfitting)**

سادگی بیش از حد مدل، منجر به کاهش صحت نتایج شده و ارزش مدلسازی را کم می کند.



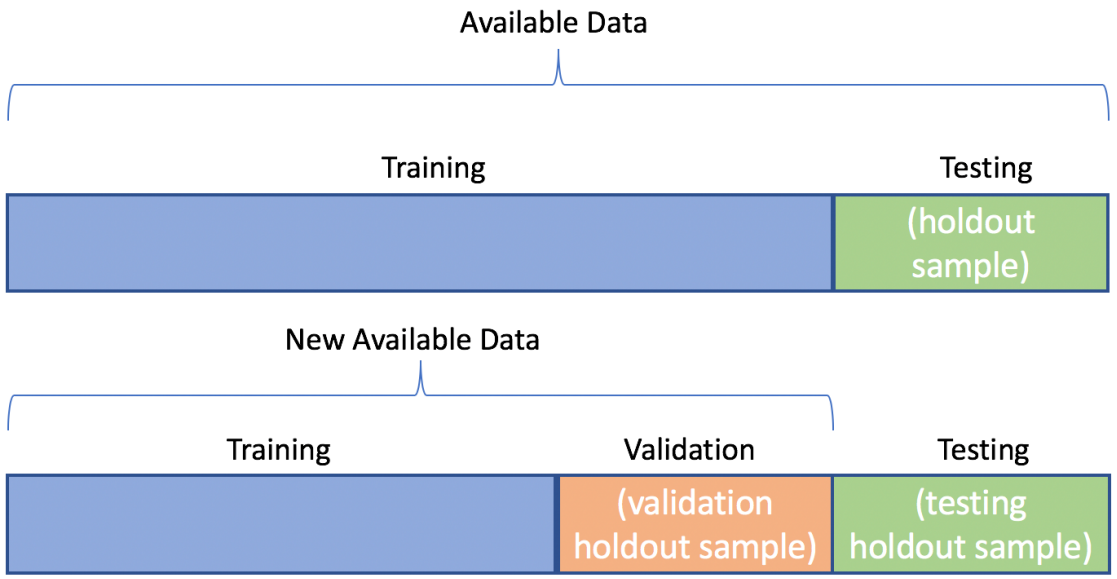
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه

❑ طرح آزمون برای مدلسازی

افراز مجموعه داده ها به سه بخش داده های آموزشی، داده ها اعتبارسنجی و داده های آزمایشی.



- **داده های آموزشی**

برای آموزش و ساخت مدل بکار می رود.

- **داده ها اعتبارسنجی**

برای تنظیم پارامترهای مدل بکار می رود.

- **داده های آزمایشی**

برای ارزیابی کیفیت مدل ساخته شده بکار می رود.

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

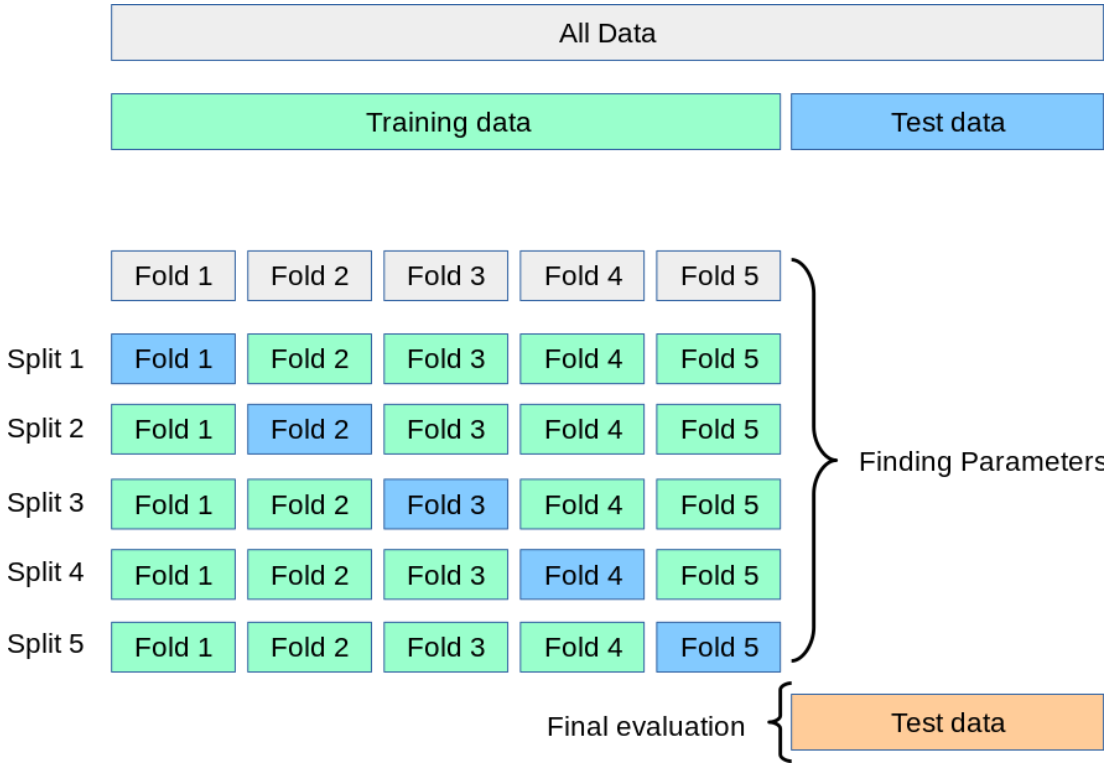
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه

- طرح آزمون برای مدلسازی
  - روش اعتبارسنجی متقابل Cross Validation

کارکرد دیگر این روش در تنظیم پارامترهای مدل می باشد. با این هدف، به جای انتخاب مجموعه داده اعتبارسنجی بصورت Holdout می توان با روش اعتبارسنجی متقابل داده های آموزشی، تنظیمات پارامترهای مدل را انجام داد.



ارزیابی مدل با این روش از تمامی داده های در دسترس جهت ساخت مدل و ارزیابی استفاده می کند؛ ولی در داده های زیاد نیاز به زمان و محاسبات بیشتری خواهد داشت.

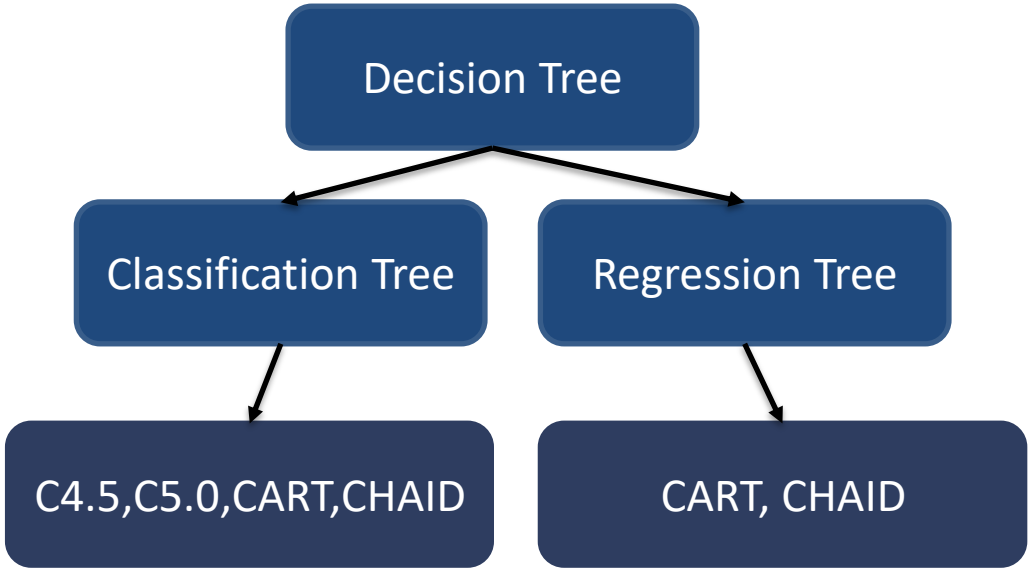
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم

❑ درخت تصمیم (Decision Tree)

این الگوریتم یکی از پرکاربردترین الگوریتم های مبتنی بر قانون است که با ایجاد مرزهای تصمیم گیری بصورت شفاف، قابلیت حل مسائل رده بندی و رگرسیون را به همراه تفسیرپذیری زیاد فراهم می کند.



- **استخراج قوانین در قالب اگر - آنگاه**

از دلایل مهمی که باعث شده الگوریتم های درخت تصمیم در حل مسائلی که بدنبال توصیف مفهوم و چرایی یک پدیده هستند موفق و پرکاربرد شود، نمایش الگوهای بدست آمده در قالب اگر – آنگاه می باشد.

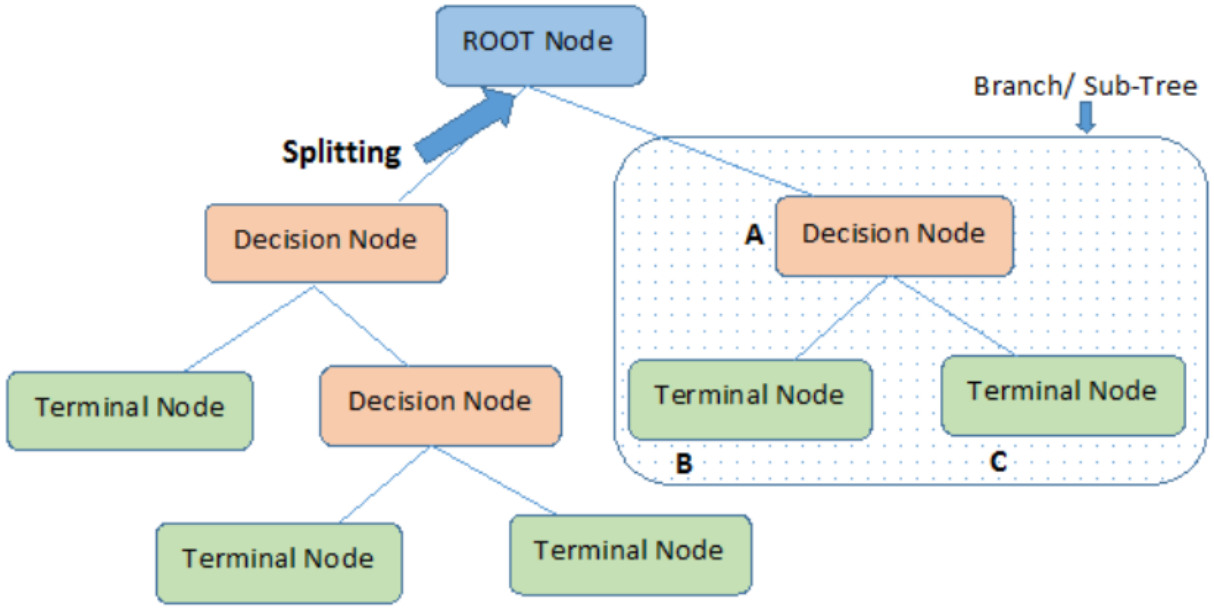
فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم

❑ درخت تصمیم (Decision Tree)

ساختار فلوچارتی درخت تصمیم دارای سه جزء اصلی زیر می باشد:

- گره ریشه (Root Node)
- گره تصمیم (Decision Node)
- گره پایانی – برگ (Terminal Node – Leaf)



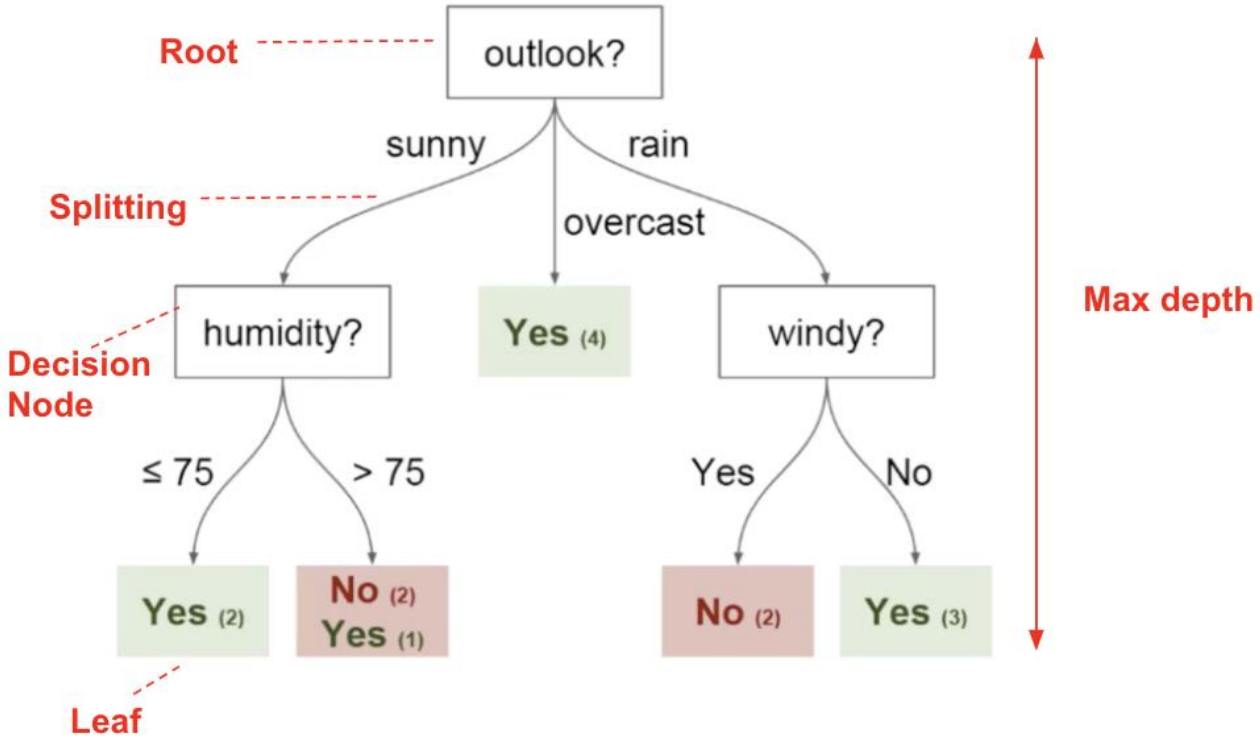
تمام داده ها در گره ریشه قرار دارند و هر رکورد بر اساس پاسخ به سوالات گره تصمیم به مسیر خود ادامه می دهد تا وارد گره پایانی شود. مسیر طی شده از ریشه تا برگ نشان دهنده یک الگو در قالب قانون اگر – آنگاه می باشد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم

❑ درخت تصمیم (Decision Tree)



جهت توسعه درخت تصمیم باید به سوالات زیر پاسخ داد:

- کدام ویژگی برای انشعاب انتخاب شود؟
- حدود آستانه ای برای انشعاب هر ویژگی چه مقادیری باشد؟
- تعداد انشعاب ها تا کجا ادامه پیدا کند؟

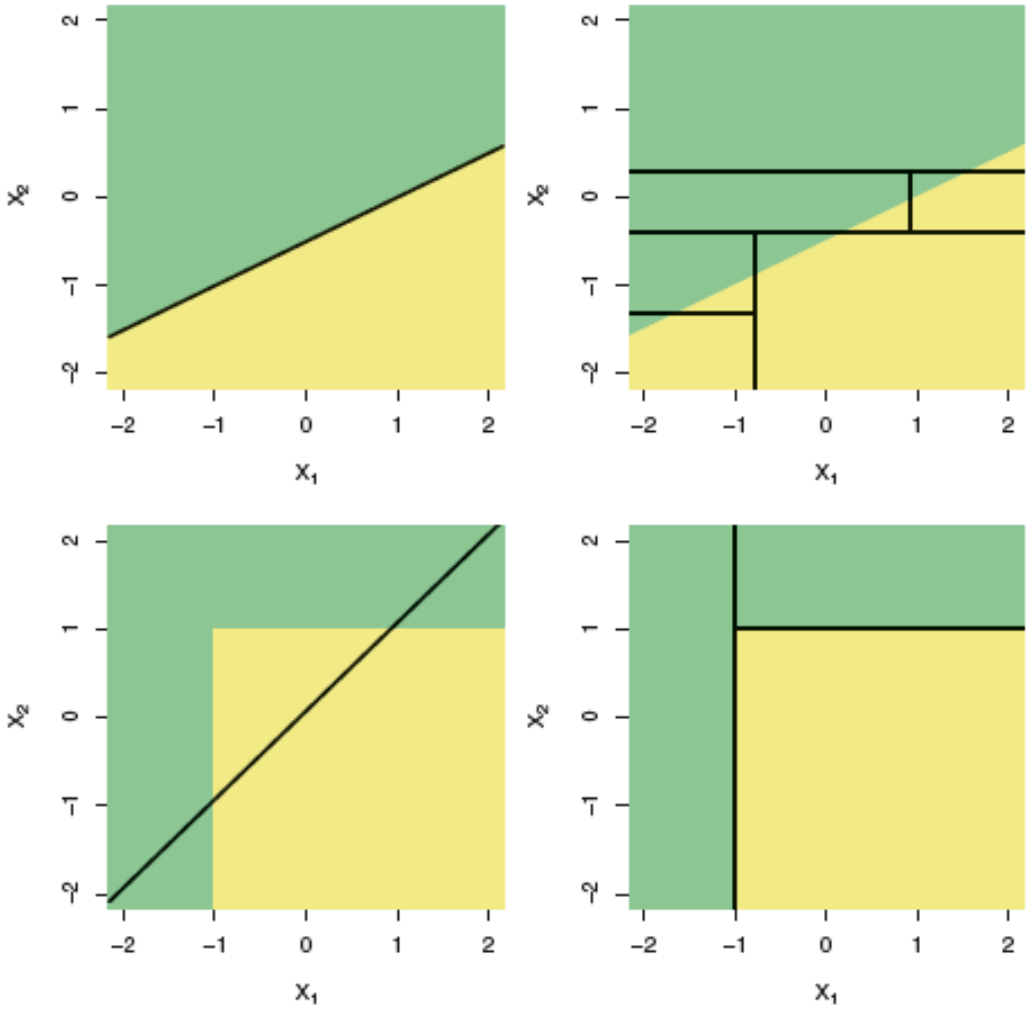
هر الگوریتمی با پاسخ به سوالات بالا می تواند یک درخت تصمیم روی داده های مسئله، آموزش دهد. الگوهای بدست آمده در قالب قوانین درخت نمایش داده می شود. بطور مثال:

$$\textbf{if}\ outlook = sunny\ \&\ humidity \leq 75\ \ \textbf{then}\ \ Target = Yes$$

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم



❑ درخت تصمیم (Decision Tree)

درخت تصمیم با افراز فضای داده ها به زیرفضاهای کوچکتر که در آنها میزان خلوص مقدار فیلد هدف حداکثر شود به شناسایی الگوها می پردازد.

بنابراین درخت تصمیم می تواند به صورت یک مدل غیرخطی روابط پیچیده در داده ها را شناسایی کند.

مثال: فرض کنید در تصاویر روبرو نواحی زرد و سبز مربوط به دو کلاس متفاوت هستند که در دو تصویر بالا، تفکیک خطی نسبت به تفکیک غیرخطی درخت تصمیم عملکرد بهتری دارد ولی در تصاویر پایین تفکیک غیرخطی درخت تصمیم جداسازی بهتری را انجام داده است.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ درخت تصمیم (Decision Tree)

مزایا و نقاط قوت درخت تصمیم

- ماهیت جعبه سفید بودن این روش و استخراج قوانین منجر به درک ساده و سریع از الگوهای بدست آمده می شود.
- نسبت به بسیاری از الگوریتم های دیگر نیاز به مراحل آماده سازی داده کمتری دارد. بطور مثال روش های نرمالسازی و یا تبدیل داده ای کیفی به عددی در توسعه درخت های تصمیم چندان مسئله ساز نیست.
- ماهیت انتخاب ویژگی بصورت درونی در ماهیت این الگوریتم وجود دارد و به همین دلیل نسبت به مشکلات کیفی مانند داده های نامرتبط و افزونگی داده ها مقاوم می باشد.
- ماهیت ناپارامتری این الگوریتم باعث می شود نیاز به فرضیات محدود کننده ای مانند برقراری فرض توزیع نرمال، استقلال ویژگی ها، تثبیت واریانس و ... نداشته باشیم.
- به علت ساختار فلوچارتی، جهت مدلسازی رفتار یا تصمیمات انسانی گزینه مطلوبی می باشد و قابلیت استفاده از آزمونهای آماری برای پایداری نتایج را دارد.

❑ درخت تصمیم (Decision Tree)

معایب و محدودیت های درخت تصمیم

- در مقابل تغییرات در داده های آموزشی الگوریتم مقاومی محسوب نمی شود و با تغییرات کم در داده های ورودی امکان تغییر در خروجی ها وجود دارد.
- به علت ماهیت جستجوی حریصانه (Greedy Search) در انشعاب های انجام شده، تضمینی برای بهینه بودن سراسری الگوها نیست.
- بطور کلی ایجاد سادگی و شفافیت بالا در این الگوریتم، منجر به کاهش صحت مدل (Accuracy) نسبت به برخی الگوریتم های دیگر می شود.

تولید محتوا: زهرا ذوالقدر

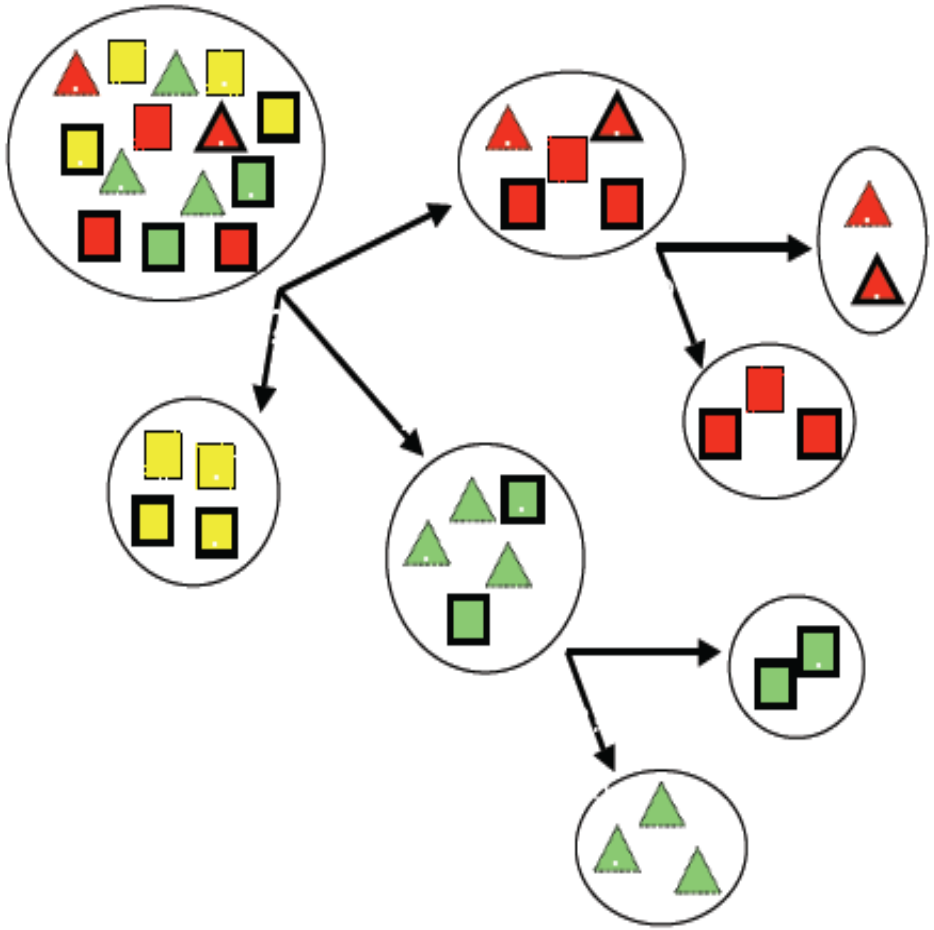
daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم



نحوه آموزش درخت تصمیم

جهت توسعه درخت تصمیم باید به سوالات زیر پاسخ داد:

- کدام ویژگی برای انشعاب انتخاب شود؟
- حدود آستانه ای برای انشعاب هر ویژگی چه مقادیری باشد؟
- تعداد انشعاب ها تا کجا ادامه پیدا کند؟

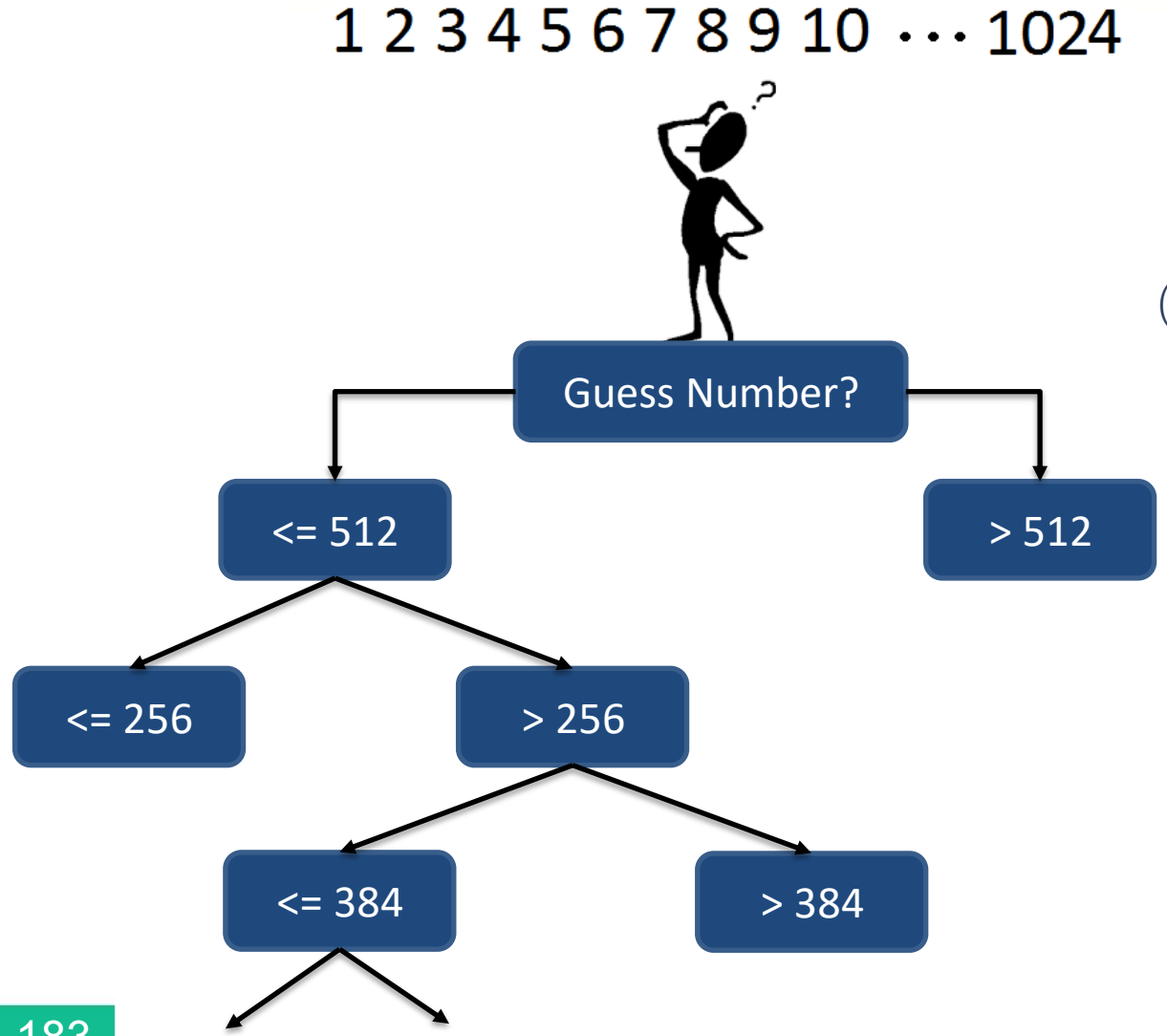
آموزش درخت تصمیم در راستای کاهش ناخالصی در داده ها انجام می شود.

بنابراین در اولین قدم بایستی معیار و شاخصی برای اندازه گیری میزان ناخالصی در نظر گرفت. این معیار در ادبیات آمار و نظریه اطلاع توسط شاخص های متنوعی محاسبه شده و با انشعاب روی سطوح مختلفی از ویژگی های ورودی مدل سعی در کاهش آن می کند تا شرایط توقف الگوریتم احراز گردد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم



❑ نحوه آموزش درخت تصمیم

- **گام اول: انتخاب معیار اندازه گیری ناخالصی**

نحوه انشعاب بایستی به نحوی انجام شود تا با کمترین تعداد سوال (انشعاب) به هدف اصلی یعنی کاهش ناخالصی برسد.

یک بازی فرضی را در نظر بگیرید که شما بایستی یک عدد انتخاب شده در بازه 1 تا 1024 را با کمترین تعداد سوال حدس بزنید. چه نوع سوالاتی می تواند به شما کمک کند تا در این بازی بهترین امتیاز را کسب کنید؟

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ نحوه آموزش درخت تصمیم

اگر به همین ترتیب ادامه دهید با جمعاً 10 سؤال می‌توانید عدد مورد نظر را بیابید. به عبارت دیگر شما به 10 بیت اطلاع نیاز دارید که از فرمول زیر به‌دست می‌آید:

$$\log_2 N = \log_2 1024 = 10$$

- **نظریه اطلاع**

وقتی شما یک مجموعه N عضوی را به دو مجموعه n و p عضوی تقسیم می‌کنید میانگین اطلاعی که همچنان مورد نیاز می باشد برابر است با:

$$\frac{n}{N}\log_2^n + \frac{p}{N}\log_2^p$$

بنابراین میانگین اطلاعی که به‌وسیله یک سؤال فراهم می‌شود برابر است با:

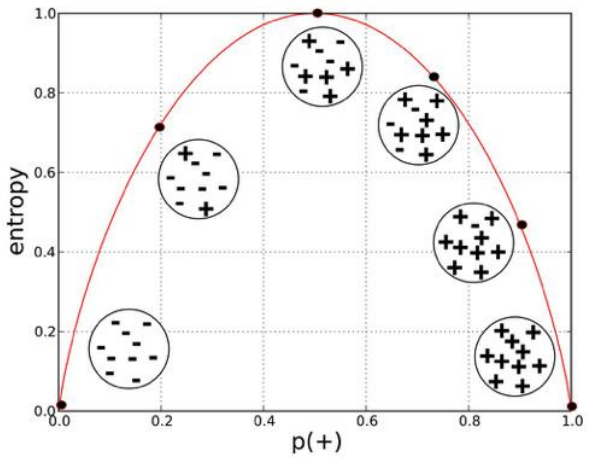
$$\log_2^N - \left(\frac{n}{N}\log_2^n + \frac{p}{N}\log_2^p\right) = -\frac{n}{N}\log_2^{n/N} - \frac{p}{N}\log_2^{p/N}$$

❑ نحوه آموزش درخت تصمیم

- شاخص بی نظمی آنتروپی (Entropy)

بر اساس نظريه اطلاع، اگر رده‌بندي اعضاي يك مجموعه به k رده موردنظر باشد و p_k احتمال تعلق اعضا به رده k ام باشد ، مقدار اطلاع مورد انتظاري كه نياز است تا اعضا به رده درست تعلق گيرند آنتروپي (شاخص بی نظمی) ناميده مي‌شود كه از فرمول زير محاسبه مي‌شود:

$$\text{Entropy}(p_1,\ldots,p_k) = I(p_1,\ldots,p_k) = -\sum_i p_i \log_2^{(p_i)}$$



بیشترین میزان بی نظمی هنگامی می باشد که احتمال تعلق به تمامی رده ها، برابر است. در تصویر، نمودار میزان آنتروپی در توزیع دو جمله ای نشان داده شده است.

تولید محتوا: زهرا ذوالقدر

daychegroup

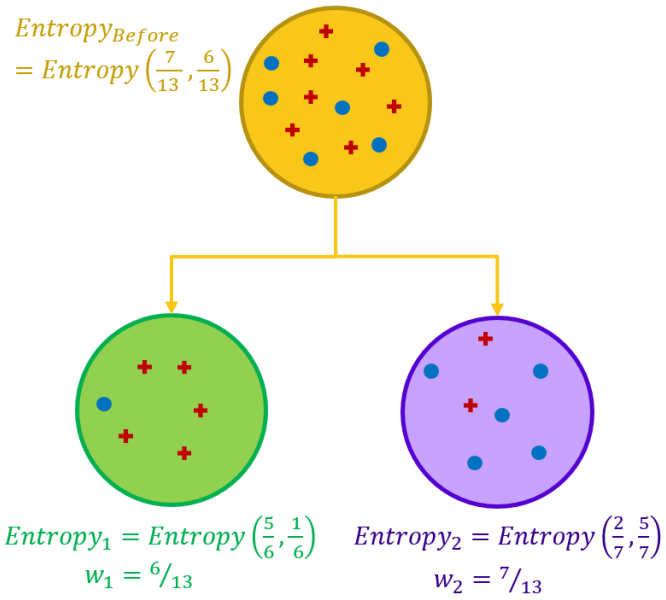
daychegroup

dayche.com | گروه دایکه

❑ نحوه آموزش درخت تصمیم

- شاخص بی نظمی آنتروپی (Entropy)

در صورتی که مجموعه داده بر اساس انشعاب روی سطوح یک ویژگی، به دو یا چند زیرمجموعه تفکیک گردد، میزان آنتروپی حاصل از انشعاب، میانگین وزنی از آنتروپی زیرمجموعه های ایجاد شده می باشد؛ بطوریکه مقدار وزن برابر با نسبت فراوانی هر زیرمجموعه به تعداد کل مجموعه اولیه می باشد.



$$Entropy\ (T, X) = \sum_{i-1}^{k} w_i * Entropy(T_i)\ \ , w_i = \frac{N_{T_i}}{N_T}$$

❑ نحوه آموزش درخت تصمیم

- **شاخص بهره اطلاعاتی (Information Gain)**

به میزان کاهش ناخالصی که بر اساس یک انشعاب ایجاد می شود، میزان بهره اطلاعاتی گفته می شود.

بطور مثال اگر مجموعه داده اولیه T را مجموعه تمام روزهای یک سال در نظر بگیریم که دارای توزیع دو جمله ای از روزهای بازی یا عدم بازی گلف باشد و (T , X) را انشعاب مجموعه T روی ویژگی شرایط آب و هوا (آفتابی / ابری / بارانی) در نظر بگیریم، میزان بهره اطلاعاتی ناشی از انشعاب شرایط آب و هوایی بر اساس رابطه زیر بدست می آید:

$$Gain(T,X) = Entropy(T) - Entropy(T,X)$$

الگوریتم درخت تصمیم ID3 از شاخص بهره اطلاعاتی برای انتخاب بهترین انشعاب استفاده می کنند.

برای این منظور در انشعاب هر گره، میزان بهره اطلاعاتی برای تمام ویژگی ها در تمام حالات امکانپذیر محاسبه می شود و در نهایت ویژگی منتخب و سطوح تفکیک آن بر اساس بیشترین میزان بهره اطلاعاتی بدست می آید.

G(PlayGolf, Outlook) = **E**(PlayGolf) – **E**(PlayGolf, Outlook)

= 0.940 – 0.693 = 0.247

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم

نحوه آموزش درخت تصمیم

- شاخص بهره اطلاعاتی (Information Gain)

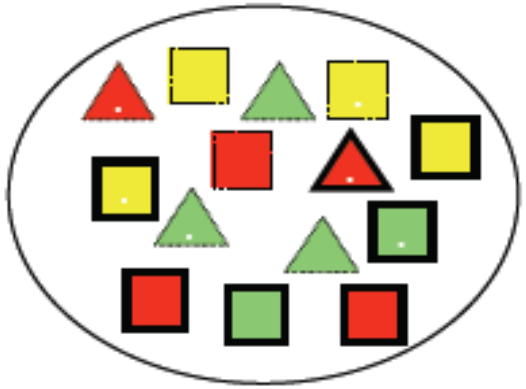
مثال از آموزش درخت تصمیم

فیلد هدف:

- کلاسهای مربع و مثلث

ویژگی های ورودی:

- رنگ: قرمز / سبز / زرد
- نقطه: بله / خیر
- خط: بله / خیر



$$p(\square) = \frac{9}{14}$$

$$p(\triangle) = \frac{5}{14}$$

Entropy :

$$I = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940 \text{ bits}$$

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم

- نحوه آموزش درخت تصمیم
  - شاخص بهره اطلاعاتی (Information Gain)

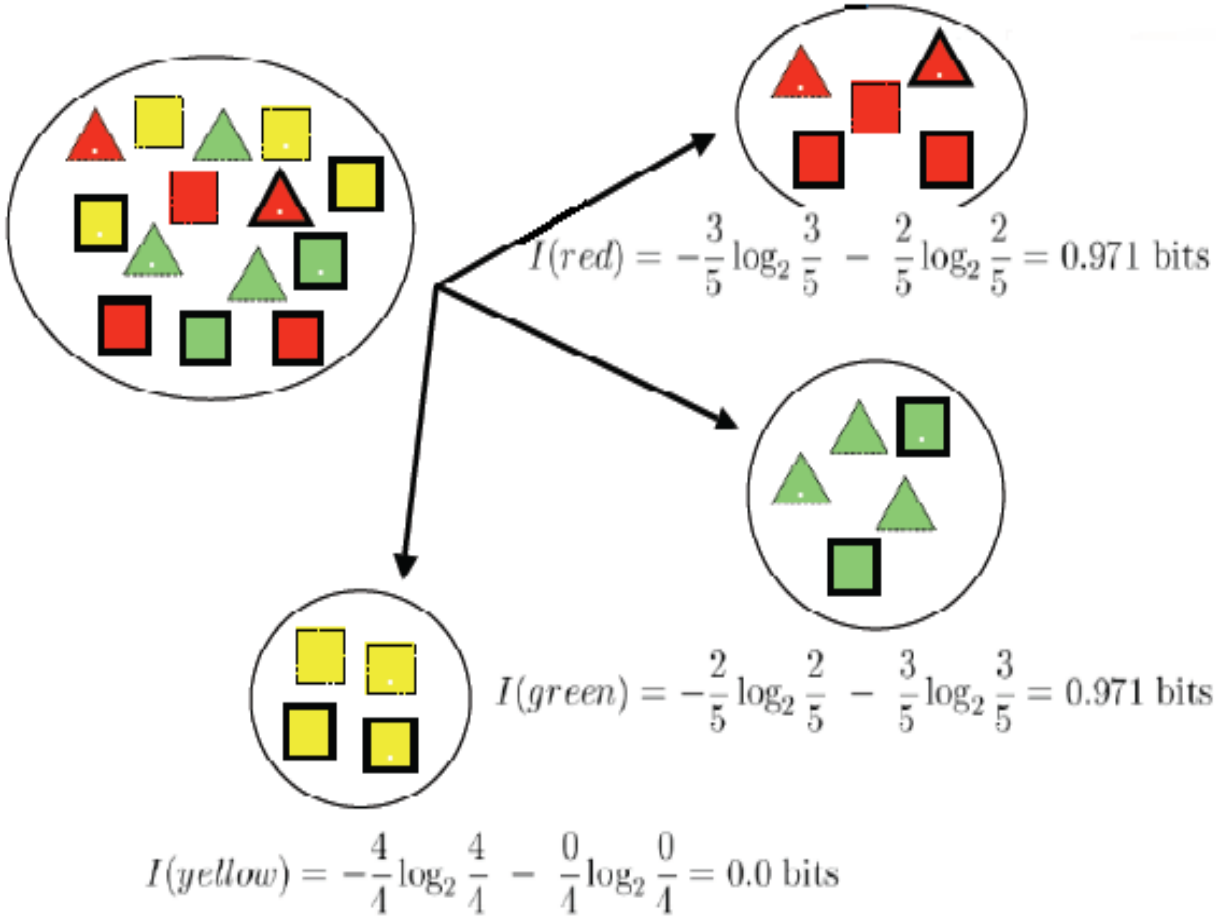
مثال از آموزش درخت تصمیم

فیلد هدف:

- کلاسهای مربع و مثلث

ویژگی های ورودی:

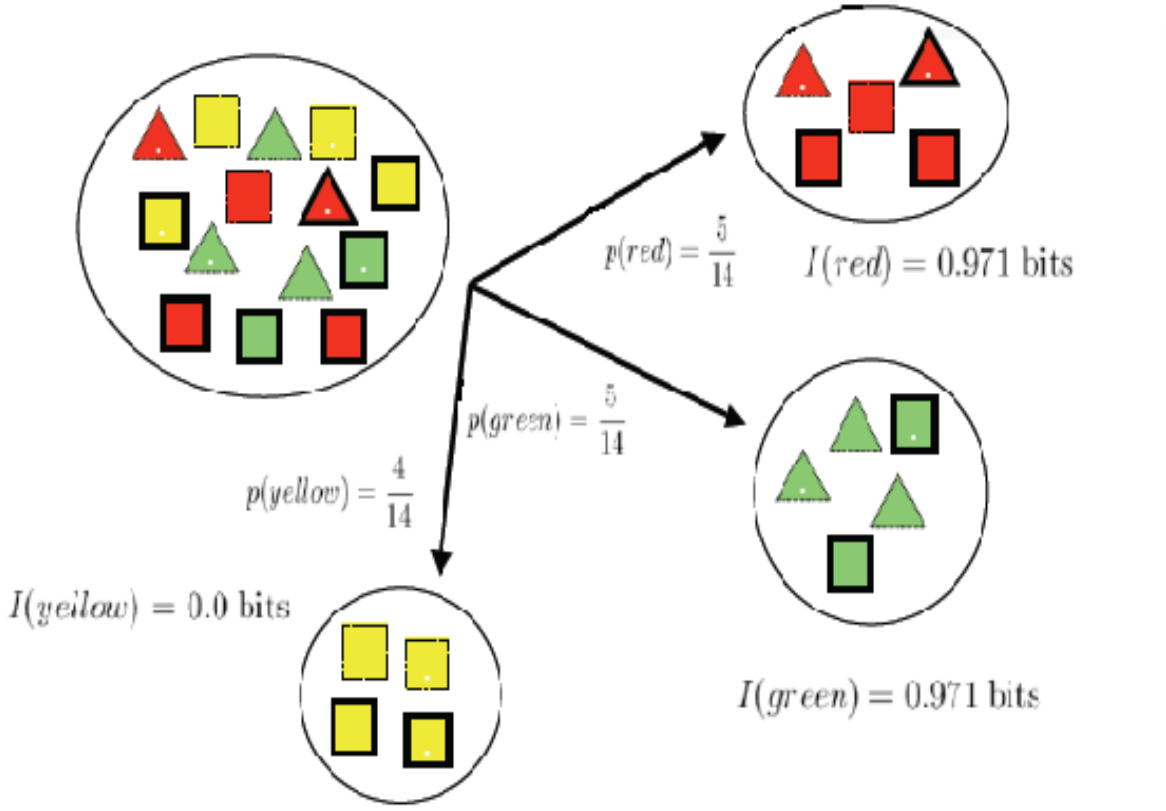
- رنگ: قرمز / سبز / زرد
- نقطه: بله / خیر
- خط: بله / خیر



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم



$$I(Color) = \sum_{v} p(v)I(v) = \frac{5}{14}0.971 + \frac{5}{14}0.971 + \frac{4}{14}0 = 0.694\ bits$$

- نحوه آموزش درخت تصمیم
  - شاخص بهره اطلاعاتی (Information Gain)

مثال از آموزش درخت تصمیم

فیلد هدف:

- کلاسهای مربع و مثلث

ویژگی های ورودی:

- رنگ: قرمز / سبز / زرد
- نقطه: بله / خیر
- خط: بله / خیر

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم

- نحوه آموزش درخت تصمیم
  - شاخص بهره اطلاعاتی (Information Gain)

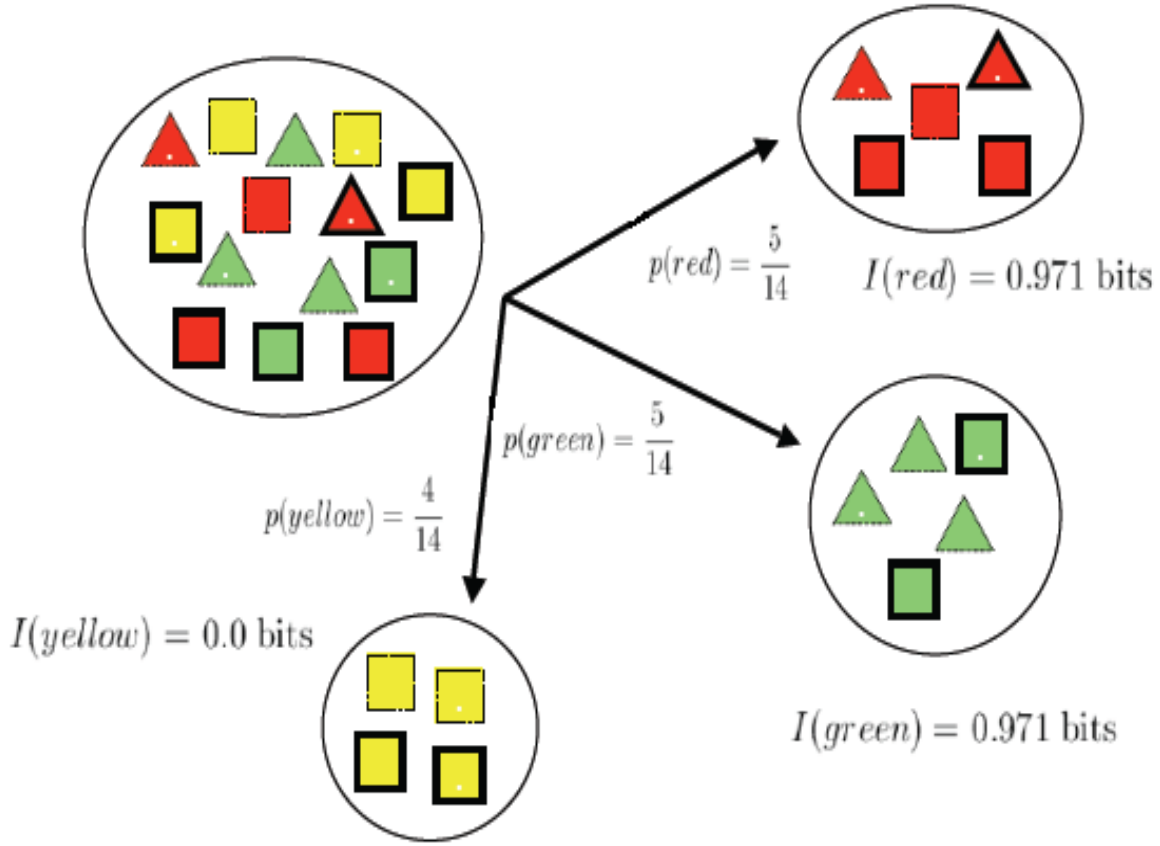
مثال از آموزش درخت تصمیم

فیلد هدف:

- کلاسهای مربع و مثلث

ویژگی های ورودی:

- رنگ: قرمز / سبز / زرد
- نقطه: بله / خیر
- خط: بله / خیر

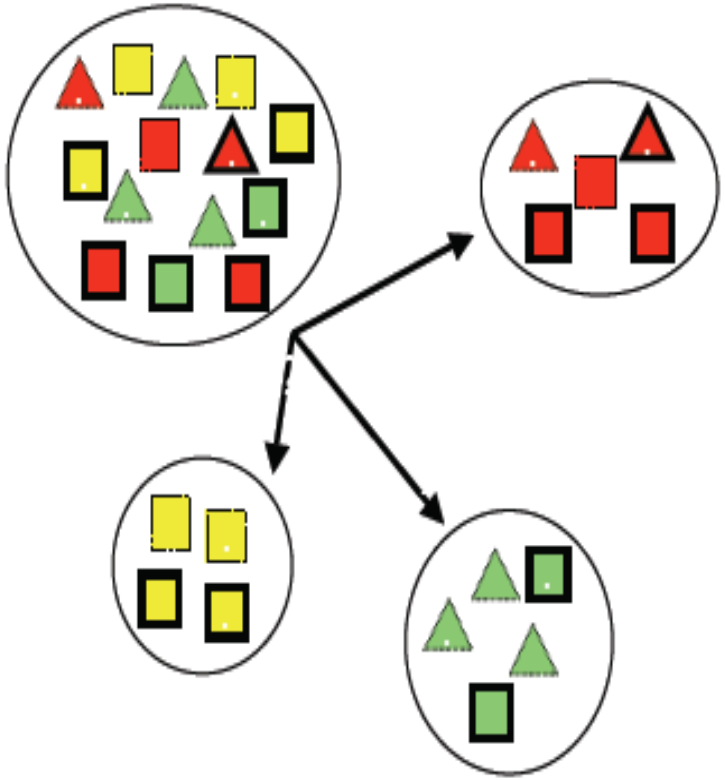


$$IG(Color) = I - I(Color) = I - \sum_{v} p(v)I(v) = 0.940 - 0.694 = 0.246\ bits$$

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم



IG(outline) = 0.971- 0 = 0.971 bits
IG(dot) = 0.971 - 0.951 = 0.020 bits

- نحوه آموزش درخت تصمیم
  - شاخص بهره اطلاعاتی (Information Gain)

مثال از آموزش درخت تصمیم

فیلد هدف:

- کلاسهای مربع و مثلث

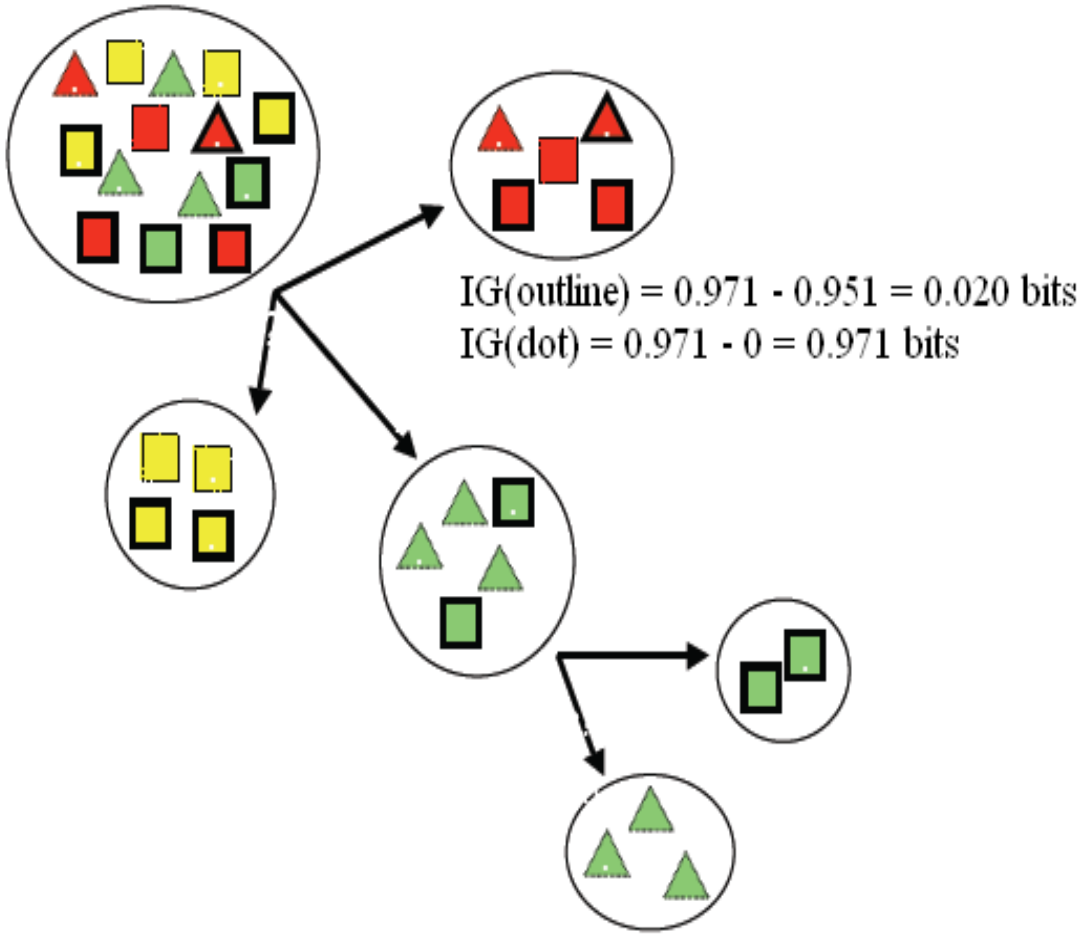
ویژگی های ورودی:

- رنگ: قرمز / سبز / زرد
- نقطه: بله / خیر
- خط: بله / خیر

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم



- نحوه آموزش درخت تصمیم
  - شاخص بهره اطلاعاتی (Information Gain)

مثال از آموزش درخت تصمیم

فیلد هدف:

- کلاسهای مربع و مثلث

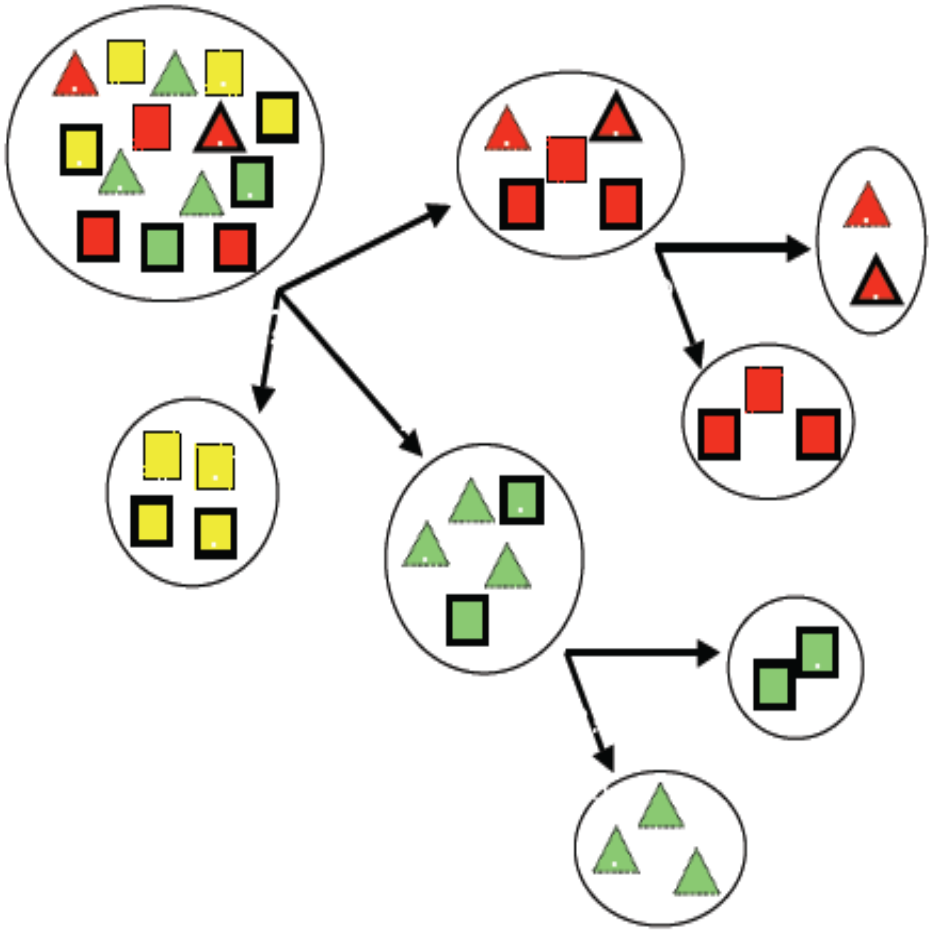
ویژگی های ورودی:

- رنگ: قرمز / سبز / زرد
- نقطه: بله / خیر
- خط: بله / خیر

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم



- نحوه آموزش درخت تصمیم
  - شاخص بهره اطلاعاتی (Information Gain)

مثال از آموزش درخت تصمیم

فیلد هدف:

- کلاسهای مربع و مثلث

ویژگی های ورودی:

- رنگ: قرمز / سبز / زرد
- نقطه: بله / خیر
- خط: بله / خیر

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ نحوه آموزش درخت تصمیم

- شاخص نسبت بهره (Gain Ratio Index)

اندازه بهره اطلاعاتی در انشعاب هایی که تعداد رده های بیشتری دارند بطور کاذب افزایش پیدا می کند و شانس انتخاب ویژگی هایی که دارای تعداد رده بیشتری هستند را نسبت به سایر ویژگی ها بیشتر می کند. برای جلوگیری از این مشکل محاسباتی، کافیست نرمالسازی میزان بهره اطلاعاتی هر ویژگی را با تقسیم بر مقدار آنتروپی توزیع انشعاب محاسبه کنیم. شاخص نسبت بهره در الگوریتم های توسعه یافته بر مبنای ID3 مانند الگوریتم های C4.5 و C5.0 جایگزین شاخص بهره اطلاعاتی شده است.

$$Gain\ Ratio = \frac{Information\ Gain}{SplitInfo} = \frac{Entropy\ (before) - \sum_{j=1}^{K} Entropy(j, after)}{\sum_{j=1}^{K} w_j \log_2 w_j}$$

$$w_j = \frac{\#\ samples\ in\ \ subset\ (j, after)}{\#\ samples\ in\ dataset\ (before)}$$

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

❑ نحوه آموزش درخت تصمیم

- **شاخص ناخالصی جینی (Gini Impurity Index)**

این شاخص در الگوریتم CART برای جداسازی و انشعاب مورد استفاده قرار می گیرد. فرض کنید مجموعه داده در گره T دارای K رده مختلف از کلاس هدف می باشد. شاخص ناخالصی جینی بر اساس رابطه زیر بدست می آید:

$$G(T) = \sum_{i=1}^{k} p_i(1-p_i) = 1 - \sum_{i=1}^{k} p_i^2$$

مقدار بیشینه شاخص ناخالصی جینی در حالتی که احتمال دو رده کلاس هدف برابر باشد، مقدار 0.5 خواهد بود.

محاسبه مقدار شاخص جینی برای هر انشعاب و همچنین میزان بهره حاصل از کاهش میزان شاخص جینی مشابه توضیحات قبلی می باشد.

		Yes	No	Total
Feature 2: Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	3	2	5
	Total	10	4	

Gini (PlayTennis, Outlook=Sunny)
= 1-(⅖)^2 - (⅗)^2 =0.48

Gini (PlayTennis, Outlook=Overcast)
= 1-(4/4)^2 - (0/4)^2=0

Gini (PlayTennis, Outlook=Rainy)
= 1-(⅗)^2 - (⅖)^2 =0.48

The Gigi Index of Outlook (children node)
= 5/14 x 0.48 + 4/14 x 0 + 5/14 x 0.48 = 0.3429

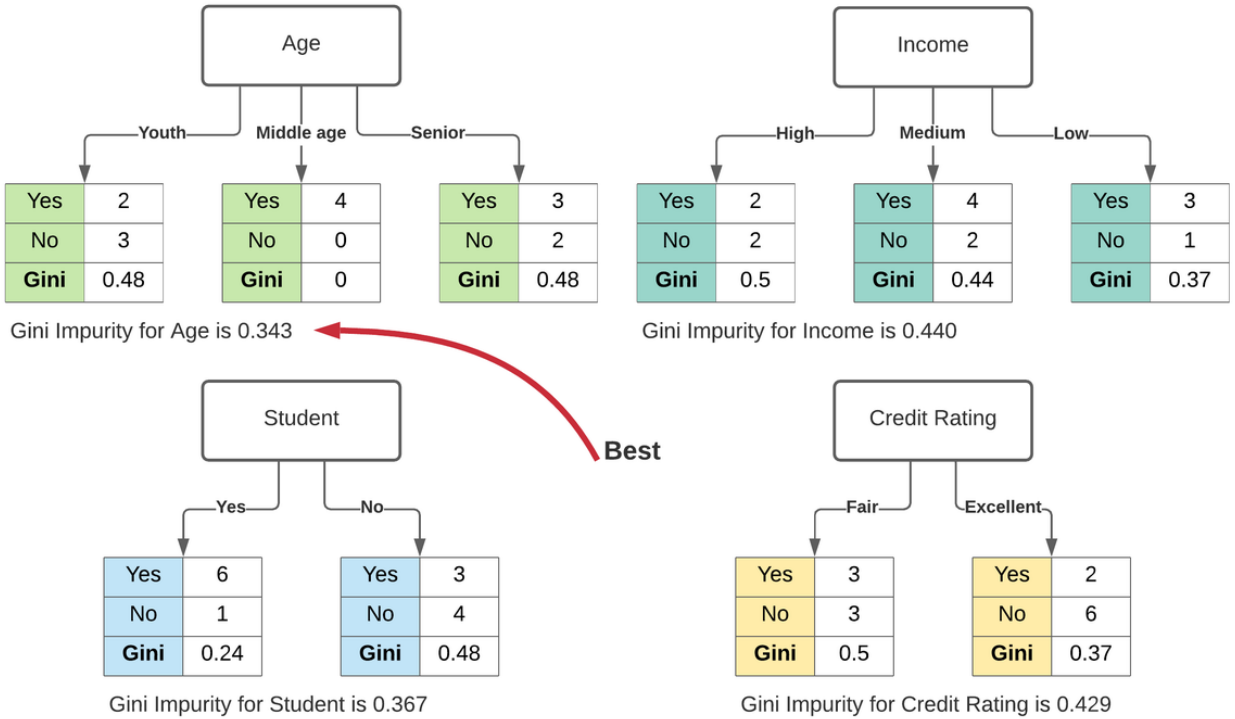
Gini Gain = Gini (parent node) - Gini (children node)

= [1- (10/14)^2 -(4/14)^2] - 0.3429
= 0.4082 - 0.3429
= 0.065

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ نحوه آموزش درخت تصمیم

- شاخص ناخالصی جینی (Gini Impurity Index)



مشابه روال قبل در انشعاب هر گره، میزان شاخص جینی برای تمام ویژگی ها در تمام حالات امکانپذیر محاسبه می شود و در نهایت ویژگی منتخب و سطوح تفکیک آن بر اساس کمترین میزان شاخص جینی بدست می آید.

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم

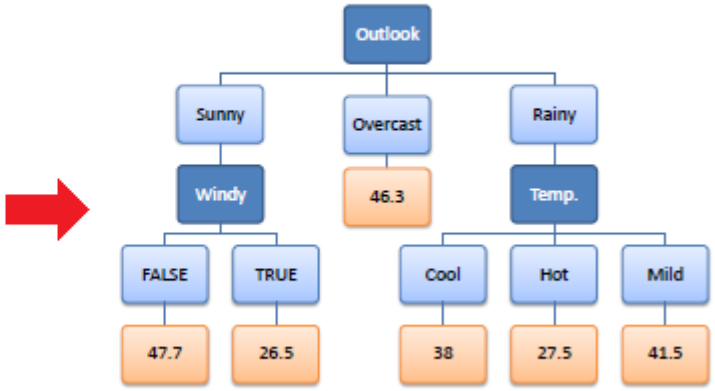
نحوه آموزش درخت تصمیم

- شاخص کاهش انحراف معیار (Standard Deviation Reduction)

در الگوریتم هایی مانند CART، کاهش انحراف معیار در مسائلی که فیلد هدف دارای توزیع کمی باشد مورد استفاده قرار می گیرد. در حالت Regression Tree مقدار میانگین یا میانه در هر گره محاسبه شده و بر اساس رابطه شاخص پراکندگی (واریانس / انحراف معیار / ...) نسبت به معیار مرکزی آن گره میزان ناخالصی آن گره و انشعاب های آن محاسبه می شود.

$$SDR = sd(T) - \Sigma_{i=1}^{n} \frac{|T_i|}{|T|} * sd(T_i)$$

Predictors				Target
Outlook	Temp.	Humidity	Windy	Hours Played
Rainy	Hot	High	False	25
Rainy	Hot	High	True	30
Overcast	Hot	High	False	48
Sunny	Mild	High	False	45
Sunny	Cool	Normal	False	52
Sunny	Cool	Normal	True	23
Overcast	Cool	Normal	True	43
Rainy	Mild	High	False	35
Rainy	Cool	Normal	False	38
Sunny	Mild	Normal	False	48
Rainy	Mild	Normal	True	48
Overcast	Mild	High	True	52
Overcast	Hot	Normal	False	44
Sunny	Mild	High	True	30



محاسبه مقدار شاخص پراکندگی برای هر انشعاب و همچنین میزان بهره حاصل از کاهش میزان شاخص پراکندگی مشابه توضیحات قبلی می باشد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
گروه دایکه | dayche.com

❑ نحوه آموزش درخت تصمیم

- شاخص اختلاف کای-دو (Chi-Square Index)

در الگوریتم درخت تصمیم CHAID از شاخص آماری کای –دو برای محاسبه شاخص ناخالصی استفاده شده است. با توجه به رابطه کای – دو ایده اصلی در بررسی اختلاف بین توزیع واقعی و توزیع قابل انتظار می باشد. در این الگوریتم توزیع قابل انتظار، شرایط کاملا برابر کلاس های فیلد هدف در هر گره می باشد؛ به همین دلیل هر چه مقدار کای – دو بزرگتر باشد نشان دهنده خلوص بیشتر در گره می باشد.

$$x^2 = \sum \frac{(Observed\ value - expected\ value)^2}{(Expected\ value)}$$

الگوریتم CHAID در حالت درخت رگرسیونی برای پیش بینی مقادیر کمی، با محاسبه آماره F و تعیین بزرگترین مقدار آن به عنوان بهترین انشعاب، آموزش درخت تصمیم را انجام می دهد.

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

❑ نحوه آموزش درخت تصمیم

- **گام دوم: انتخاب بهترین انشعاب برای هر ویژگی**

پس از انتخاب معیار مناسب برای اندازه گیری میزان ناخالصی، برای ایجاد انشعاب روی هر ویژگی بسته به توزیع کمی یا کیفی ویژگی مربوطه از روش های زیر استفاده می شود:

- **ویژگی های کیفی**
- در صورتی که ویژگی باینری باشد، انشعاب دوتایی ایجاد شده و معیار ناخالصی برای آن محاسبه و ثبت می گردد.
- در صورتی که ویژگی اسمی دارای k مقدار باشد، تمام حالت های ممکن برای انشعاب دوتایی، 3تایی، ... و k-تایی ایجاد و معیار ناخالصی برای همه آنها محاسبه شده و انشعاب دارای کمترین مقدار ناخالصی برای آن ویژگی انتخاب می شود.
- در صورتی که ویژگی ترتیبی و دارای k مقدار باشد، تمام حالت ها ممکن برای انشعاب دوتایی، 3تایی، ... و k-تایی با شرط حفظ همسایگی مقادیر ویژگی ایجاد و معیار ناخالصی برای همه آنها محاسبه شده و انشعاب دارای کمترین مقدار ناخالصی برآن آن ویژگی انتخاب می شود.

❑ نحوه آموزش درخت تصمیم

- **گام دوم: انتخاب بهترین انشعاب برای هر ویژگی**

پس از انتخاب معیار مناسب برای اندازه گیری میزان ناخالصی، برای ایجاد انشعاب روی هر ویژگی بسته به توزیع کمی یا کیفی ویژگی مربوطه از روش های زیر استفاده می شود:

- **ویژگی های کمی**

جهت تعیین تعداد انشعاب ویژگی های کمی و مقادیر آستانه ای آن، مراحل زیر انجام می شود:

- کلیه مقادیر یکتای ویژگی کمی مورد نظر، از کوچک به بزرگ مرتب می شوند.
- حد وسط مقادیر به عنوان نقاط آستانه ای اولیه انتخاب شده و گسسته سازی بر اساس بازه های ایجاد شده انجام می شود.
- مشابه داده های ترتیبی، تمامی انشعاب های ممکن انجام و معیار ناخالصی برای آنها محاسبه شده و انشعاب دارای کمترین میزان ناخالصی به عنوان بهترین انشعاب برای آن ویژگی انتخاب می شود.

❑ نحوه آموزش درخت تصمیم

- گام سوم: انتخاب بهترین ویژگی برای هر گره

در هریک از گره های تصمیم گیری، پس از بررسی تمامی حالات ممکن انشعاب برای هر یک از ویژگی ها و انتخاب بهترین حالت برای آنها، الگوریتم درخت تصمیم به مقایسه میزان ناخالصی ایجاد شده بین تمام ویژگی ها پرداخته و بهترین ویژگی برای ایجاد کمترین ناخالصی را برای انشعاب در نظر می گیرد.

بعد از ایجاد انشعاب، هر یک از گره های ایجاد شده جدید مستقلا به عنوان یک گره تصمیم در نظر گرفته شده و کلیه مراحل گفته شده برای هریک از آنها تکرار می شود تا بر اساس قوانین توقف الگوریتم از رشد بیشتر درخت جلوگیری شود.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم

❑ نحوه آموزش درخت تصمیم

- **گام چهارم: هرس کردن درخت تصمیم**

مرحله مهم دیگر در ساخت درخت تصمیم استفاده از روش های هرس کردن درخت به منظور جلوگیری از بیش برازش شدن مدل می باشد.

در نتیجه با حذف قوانین ضعیف و ساده کردن مدل درخت تصمیم، در کنار افزایش قدرت تعمیم پذیری، میزان تفسیرپذیری مدل نیز افزایش می یابد.

تکنیک های هرس درخت با دو رویکرد انجام می پذیرد:

پیش هرس
Pre-Pruning

با تعریف قوانین توقف در هنگام توسعه درخت تصمیم از رشد بیش از حد آن جلوگیری می کند.

پس هرس
Post-Pruning

پس از توسعه کامل درخت تصمیم، بر اساس محاسبه هزینه اقدام به هرس مدل می کند.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ نحوه آموزش درخت تصمیم

- **روش پیش هرس کردن درخت تصمیم (Pre-Pruning)**

بطور معمول از طریق تنظیم قوانین توقف زیر می توان از توسعه درخت تصمیم جلوگیری کرد:

- **تعیین عمق درخت تصمیم**

  با تعیین عمق درخت تصمیم می توان تعداد انشعاب ها را کنترل کرد. بدین ترتیب در صورتی که انشعاب یک گره، از حداکثر عمق مشخص شده بیشتر گردد، اجازه انشعاب وجود نداشته و آن گره به عنوان گره پایانی یا برگ در نظر گرفته می شود

- **تعیین حداقل رکورد برای انشعاب**

  با تعیین تعداد حداقل رکورد برای انشعاب، در صورتی که تعداد رکوردهای یک گره از میزان مشخص شده کمتر باشد، اجازه انشعاب صادر نشده و آن گره به عنوان گره پایانی در نظر گرفته می شود.

- **تعیین حداقل رکورد برای گره پایانی**

  با تعیین تعداد حداق رکورد برای گره پایانی، در صورتی که انشعاب گره والد، منجر به ایجاد گره هایی با کمتر از تعداد رکورد تعیین شده شود، اجازه انشعاب صادر نشده و گره والد به عنوان گره پایانی یا برگ در نظر گرفته می شود.

تولید محتوا: زهرا ذوالقدر

daychegroup

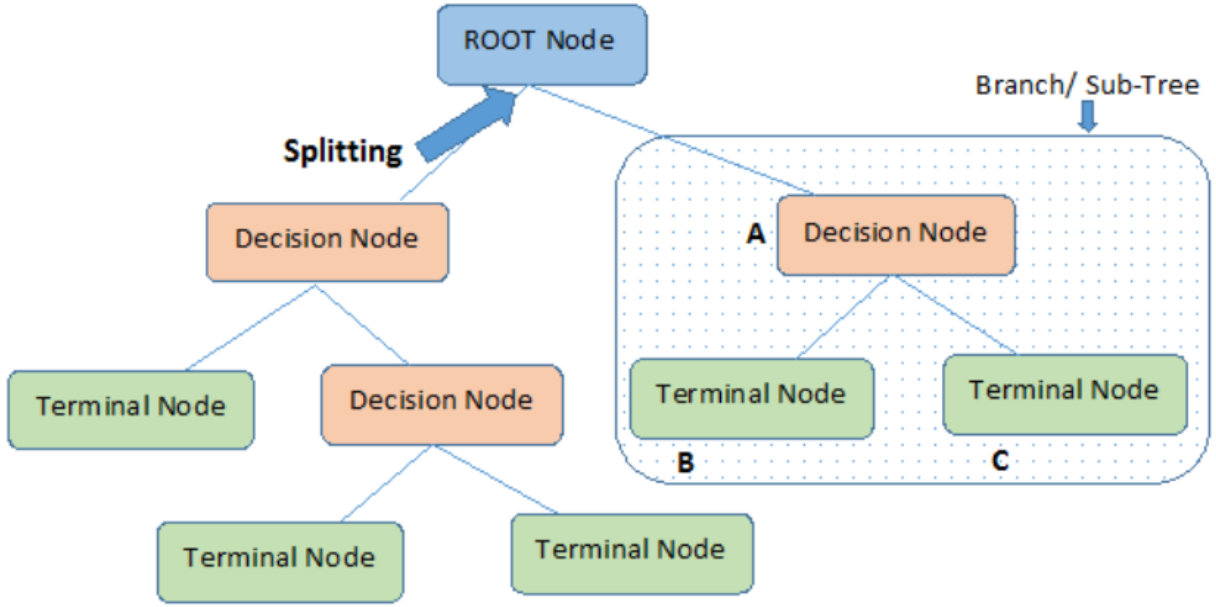
daychegroup

dayche.com | گروه دایکه

❑ نحوه آموزش درخت تصمیم

- روش پس هرس کردن درخت تصمیم (Post-Pruning)

در این رویکرد، درخت تصمیم اولیه بدون محدودیت ساخته می شود. سپس بر اساس اندازه گیری میزان هزینه توسعه آن، در خصوص حذف و هرس کردن بخش هایی از درخت تصمیم گیری می شود.



$$R_\alpha(T) = \mathrm{R(T)} + \alpha|T|$$

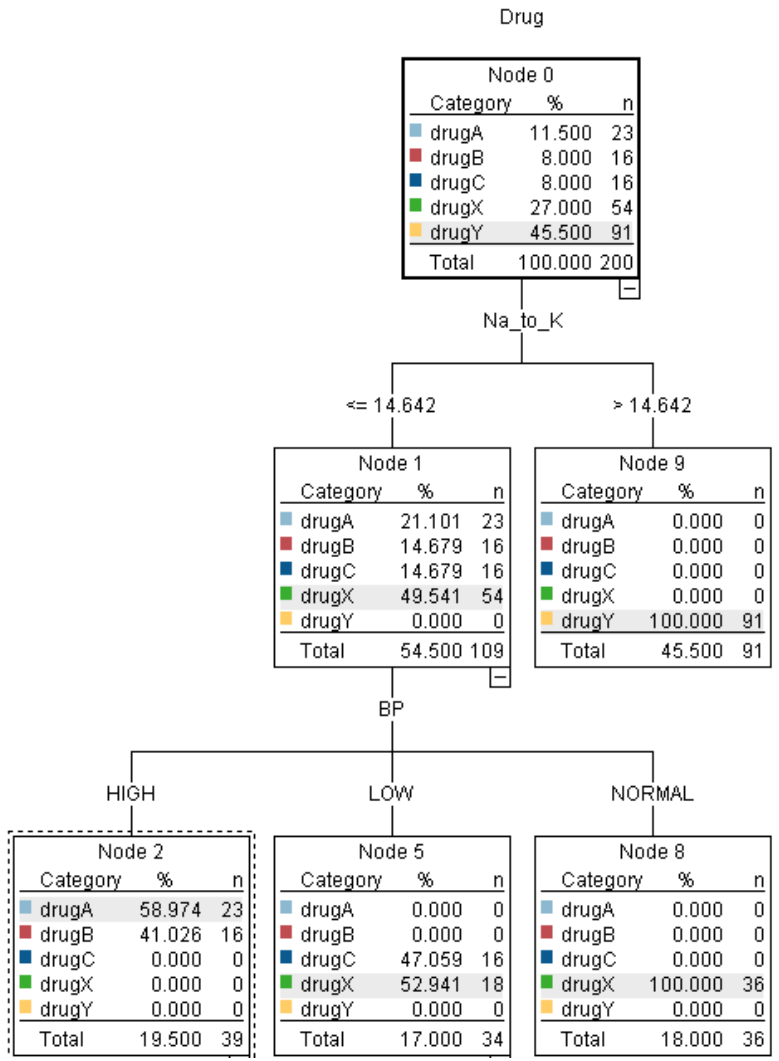
R(T): مجموع خطای آموزش در گره های پایانی (برگ ها)

|T|: تعداد کل گره های پایانی (برگ ها)

α: میزان شدت هرس

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم



❑ تفسیر قوانین درخت تصمیم

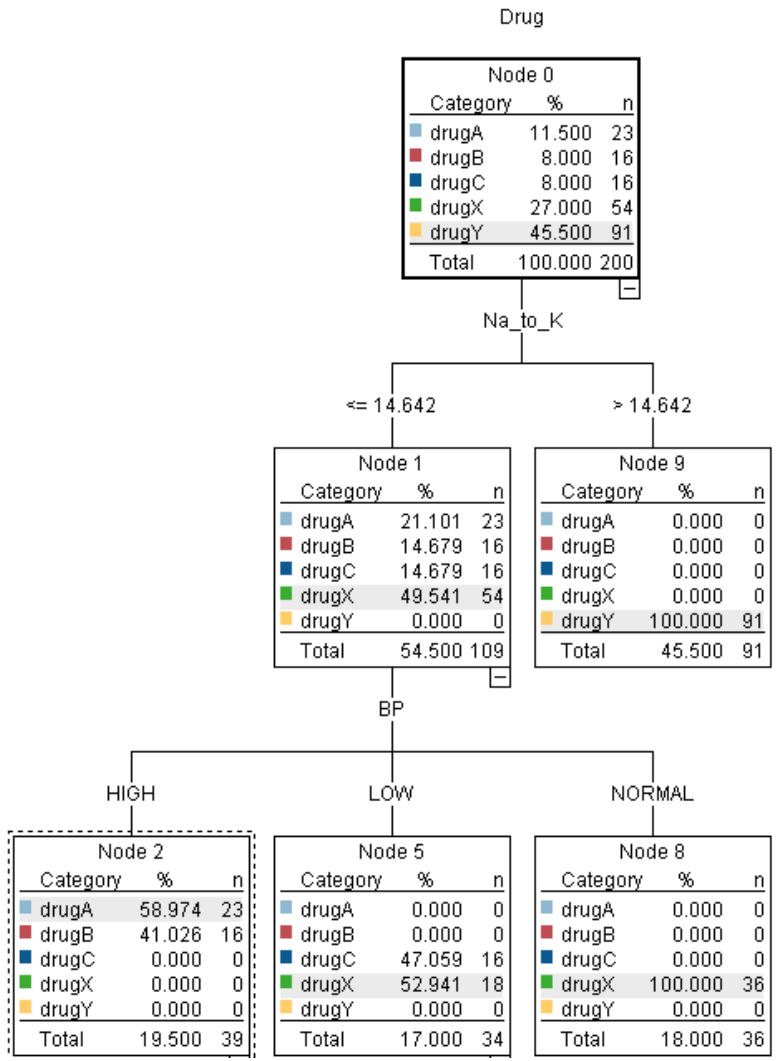
طبق ساختار درخت تصمیم، مسیر شروط و تصمیم هایی که از گره ریشه به گره پایانی (برگ) می رسد، یک قانون در نظر گرفته می شود. بنابراین بر اساس نحوه رشد و توسعه درخت تصمیم، کل فضای داده های اولیه به مجموعه قوانین بدست آمده افراز می شود.

در نتیجه هر قانون، یک زیر مجموعه از کل داده های آموزشی می باشد و افزایش تعداد برگ ها در درخت تصمیم، به معنی تعداد قوانین بیشتر و در نتیجه ایجاد زیرمجموعه های کوچکتر (با تعداد رکورد کمتر) می باشد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم



❑ تفسیر قوانین درخت تصمیم

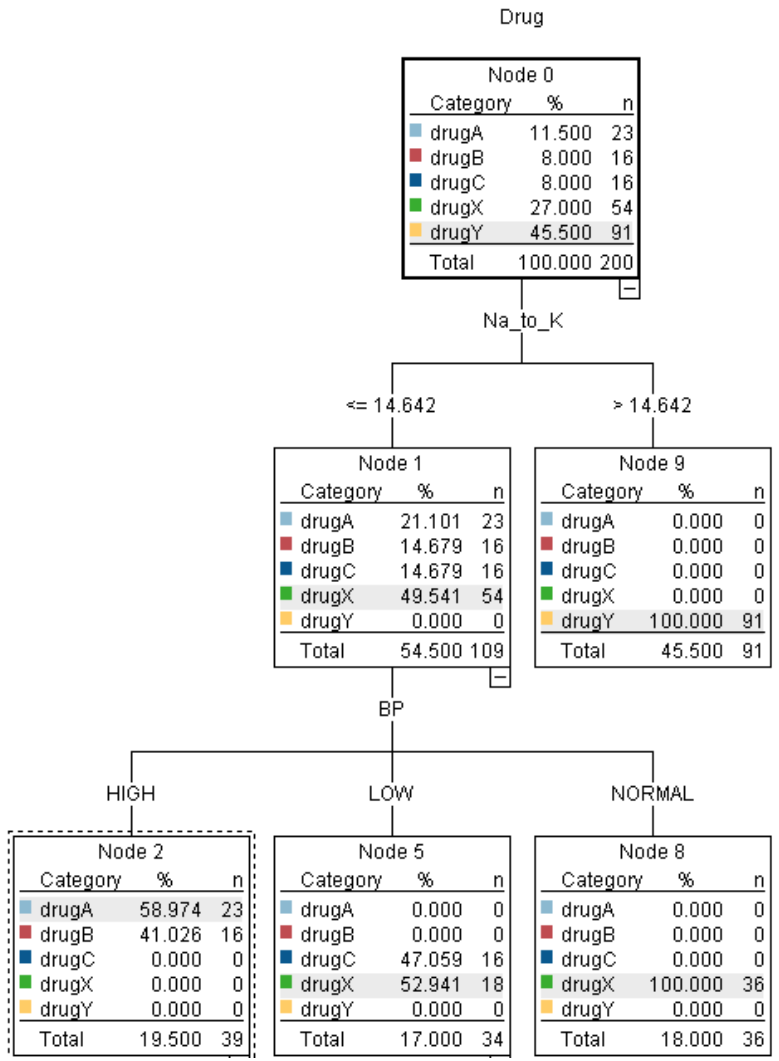
همچنین بر اساس معیارهای انتخاب ویژگی در ساختار درخت، هر انشعاب به سمت خلوص بیشتر در توزیع مقادیر فیلد هدف حرکت می کند. بدین ترتیب با ایجاد انشعاب بیشتر، پیش بینی مقدار هدف با دقت بالاتری انجام می شود.

در نتیجه افزایش تعداد انشعاب، منجر به افزایش تعداد برگ ها شده که تعداد قوانین بدست آمده از درخت را بیشتر می کند و این موضوع بصورت مستقیم منجر به افزایش دقت قوانین و مدل درخت تصمیم می شود.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
گروه دایکه | dayche.com

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم



❑ تفسیر قوانین درخت تصمیم

- شاخص های ارزیابی کیفیت قوانین
- **معیار پشتیبانی (Support Index)**

این معیار به نسبت تعداد رکوردهای یک قانون به کل داده های آموزشی/آزمایشی اشاره دارد و نشان دهنده قدرت تعمیم پذیری آن قانون می باشد.

$$Support\ (A \rightarrow B) = \frac{frq(A)}{N}$$

بدیهی هست افزایش میزان پشتیبانی قوانین در جهت کاهش پیچیدگی مدل و کوچکتر کردن درخت تصمیم است. در این حالت میزان تعمیم پذیری مدل افزایش می یابد و شانس بیش برازشی مدل کاهش می یابد.

تولید محتوا: زهرا ذوالقدر

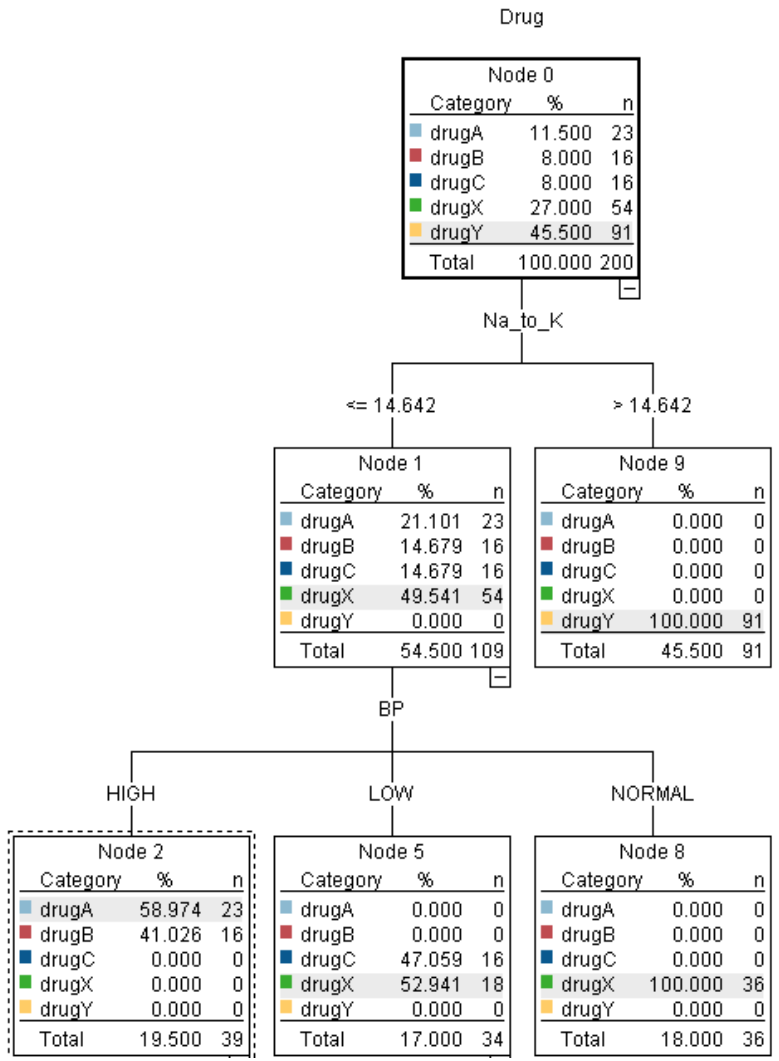
daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم



❑ تفسیر قوانین درخت تصمیم

- شاخص های ارزیابی کیفیت قوانین
  - **معیار اطمینان** (Confidence Index)

این معیار به نسبت تعداد رکوردهای با برچسب های واقعی کلاس هدف و تعداد کل رکوردهای گره پایانی (برگ) گفته می شود و نشان دهنده اطمینان قانون می باشد.

$$Confidence\ (A \rightarrow B) = P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{frq(A,B)}{frq(A)}$$

بدیهی هست افزایش میزان اطمینان قوانین در جهت افزایش پیچیدگی مدل و بزرگتر کردن درخت تصمیم است. در این حالت میزان صحت مدل افزایش می یابد ولی باید توجه نمود شانس بیش برازشی مدل نیز افزایش می یابد.

تولید محتوا: زهرا ذوالقدر

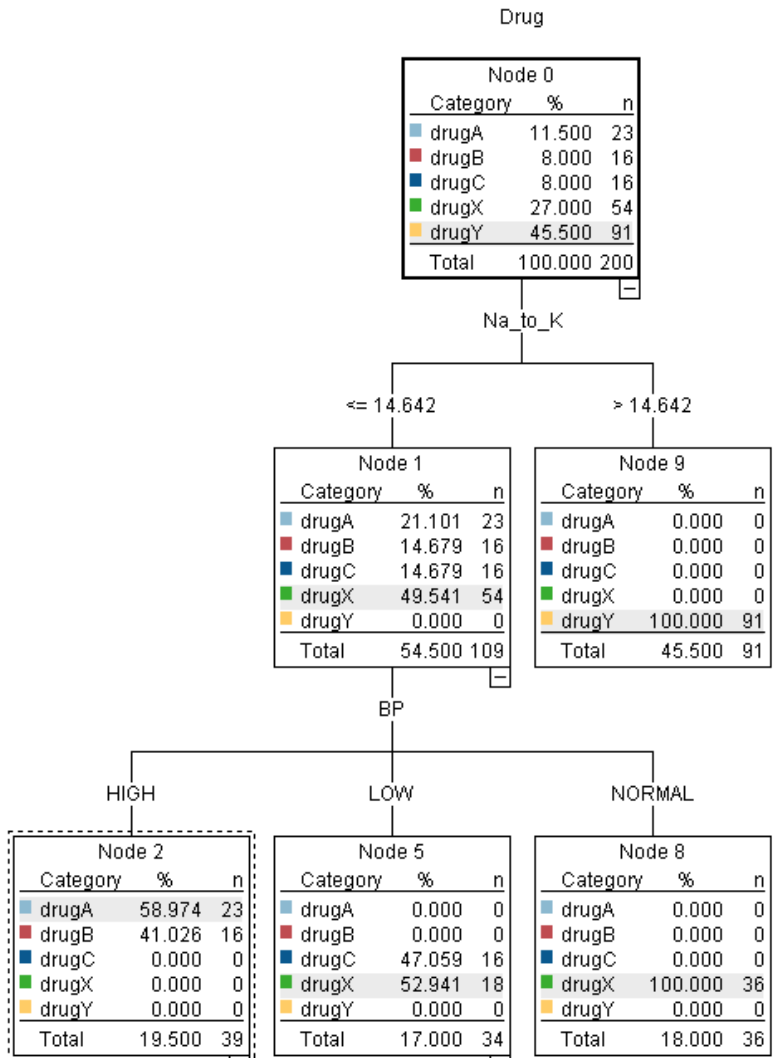
daychegroup

daychegroup

گروه دایکه | dayche.com

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم



❑ تفسیر قوانین درخت تصمیم

- شاخص های ارزیابی کیفیت قوانین
  - معیار ارتقا (Lift Index)

این معیار به نسبت اطمینان حاصل از یک قانون به احتمال وقوع اولیه مقدار پیش بینی شده توسط آن قانون گفته می شود و نشان دهنده میزان بدیع بودن یک قانون می باشد.

$$Lift\ (A \rightarrow B) = \frac{Confidence(A \rightarrow B)}{P(B)}$$

بدیهی هست مقدار Lift برابر یا نزدیک به یک به این معنی می باشد که قانون حاصل شده تاثیری بر شانس وقوع B نداشته است. هرچقدر عدد Lift از مقدار یک فاصله بگیرد قانون بدست آمده جذابیت بیشتری خواهد داشت.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – درخت های تصمیم

❑ تفسیر قوانین درخت تصمیم

- مقایسه شاخص های ارزیابی و رتبه بندی قوانین

طبق تعاریفی که از شاخص های ارزیابی کیفیت قوانین گفته شده میزان شاخص های پشتیبانی و اطمینان در تقابل با یکدیگر هستند و افزایش یکی از آنها می تواند منجر به کاهش دیگری گردد.

- در انتخاب و رتبه بندی قوانین معمولا مقدار حداقل شاخص پشتیبانی بر اساس شرایط مسئله و کسب و کار تعیین شده و سپس قوانین منتخب بر مبنای میزان شاخص اطمینان و یا شاخص ارتقا رتبه بندی می شوند.

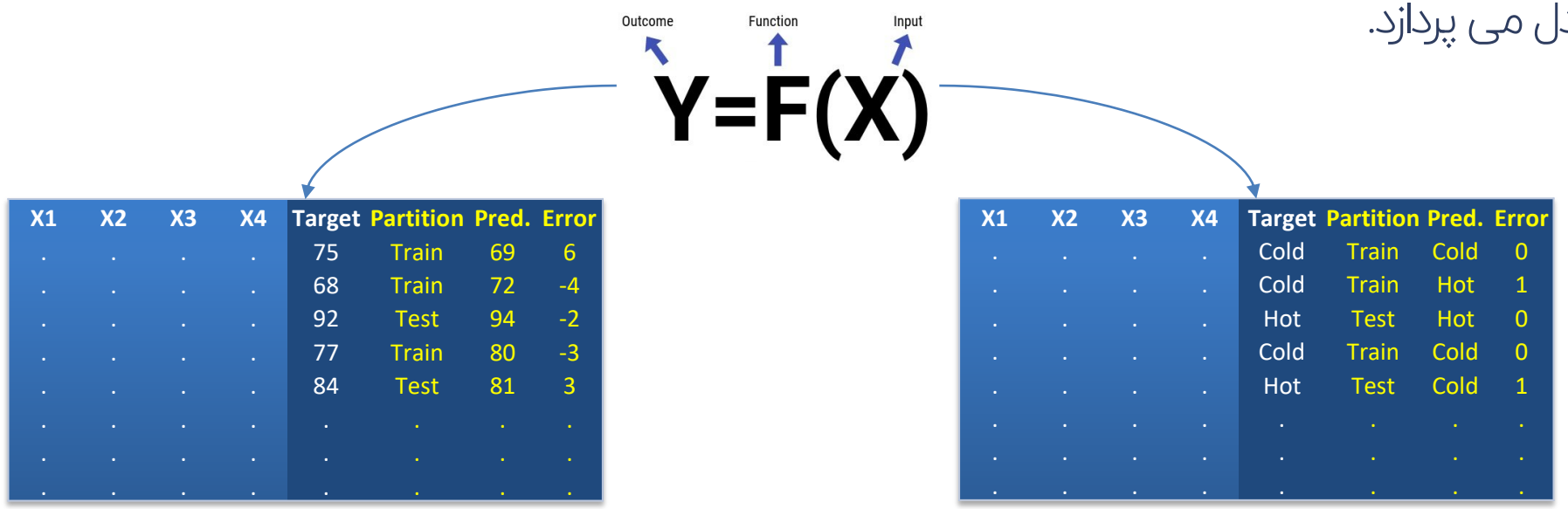
- در مسائل رده بندی با کلاس های نامتوازن، معمولا میزان شاخص اطمینان برای کلاس مینور عدد کوچکتری می باشد و شاخص ارتقا درک بهتری از کیفیت قوانین بدست آمده می دهد.

فرآیند داده کاوی

مدل های پیش بینانه – ارزیابی مدل ها

❑ ارزیابی مدل ها

ارزیابی مدل ها بر مبنای طرح آزمون استفاده شده در فاز مدلسازی، به مقایسه مقادیر واقعی و مقادیر پیش بینی شده توسط مدل می پردازد.



X1	X2	X3	X4	Target	Partition	Pred.	Error
.	.	.	.	75	Train	69	6
.	.	.	.	68	Train	72	-4
.	.	.	.	92	Test	94	-2
.	.	.	.	77	Train	80	-3
.	.	.	.	84	Test	81	3
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

ارزیابی مدل های رگرسیون

X1	X2	X3	X4	Target	Partition	Pred.	Error
.	.	.	.	Cold	Train	Cold	0
.	.	.	.	Cold	Train	Hot	1
.	.	.	.	Hot	Test	Hot	0
.	.	.	.	Cold	Train	Cold	0
.	.	.	.	Hot	Test	Cold	1
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

ارزیابی مدل های رده بندی

فرآیند داده کاوی

مدل های پیش بینانه – ارزیابی مدل ها

- ارزیابی مدل های رده بندی
  - شاخص صحت (Accuracy)

شاخص ساده و پرکاربرد در ارزیابی مدل های رده بندی است که نسبت رکوردهای با پیش بینی صحیح به تمام رکوردها را محاسبه می کند.

X1	X2	X3	X4	Target	Partition	Pred.	Error
.	.	.	.	Cold	Train	Cold	0
.	.	.	.	Cold	Train	Hot	1
.	.	.	.	Hot	Test	Hot	0
.	.	.	.	Cold	Train	Cold	0
.	.	.	.	Hot	Test	Cold	1
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

$$Accuracy = 1 - Classification\ Error$$

سوال: در مسئله تشخیص یک بیماری نادر که شانس وقوع آن 2% می باشد،

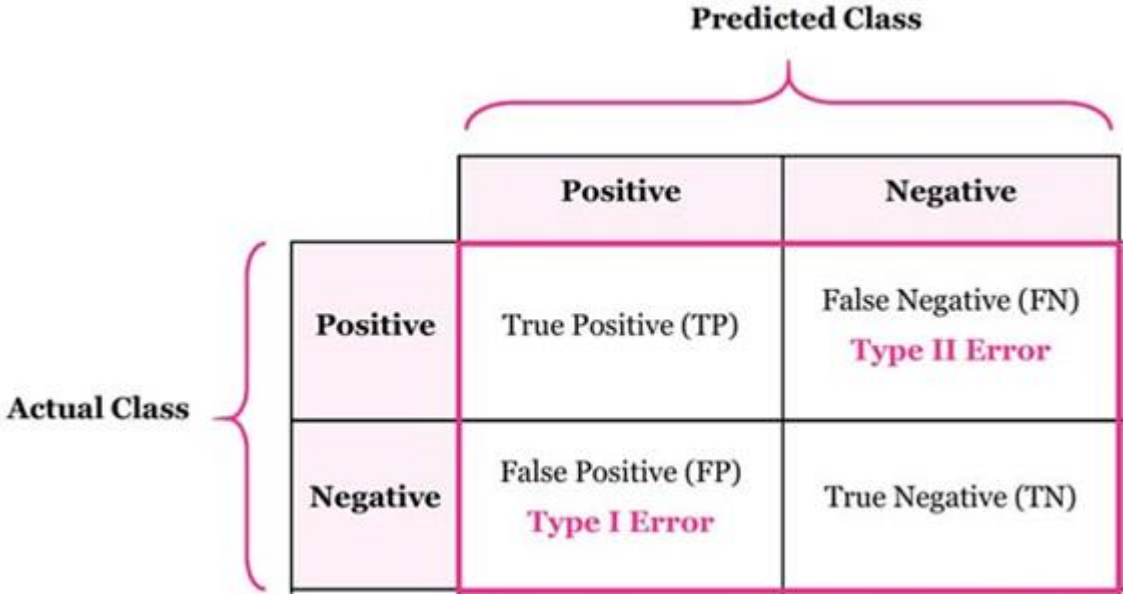
آیا شاخص صحت مدل (Accuracy) متر مناسبی برای ارزیابی مدل می باشد؟

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

ارزیابی مدل های رده بندی

- ماتریس درهم ریختگی (Confusion Matrix)

در ماتریس درهم ریختگی مقایسه مقادیر واقعی و مقادیر پیش بینی مدل، به تفکیک هر یک از کلاس های فیلد هدف توزیع می شود.



TP: مجموعه رکوردهایی که توسط مدل به درستی، کلاس مثبت پیش بینی شد.

TN: مجموعه رکوردهایی که توسط مدل به درستی، کلاس منفی پیش بینی شد.

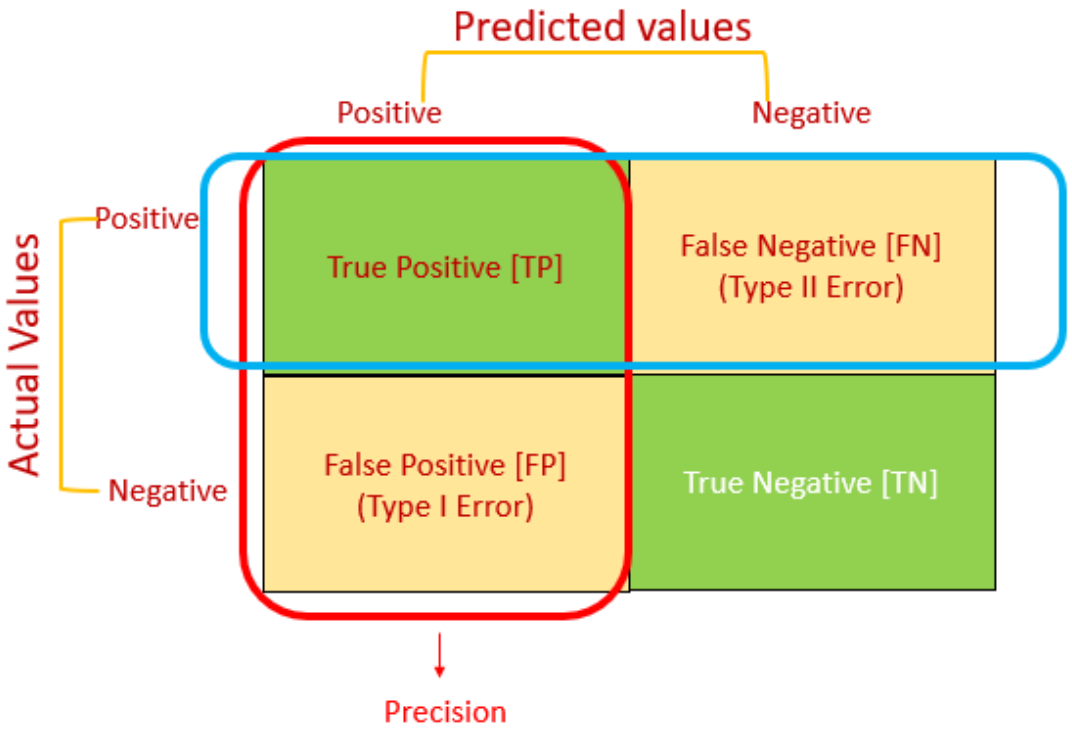
FP: مجموعه رکوردهایی که توسط مدل به اشتباه، کلاس مثبت پیش بینی شد.

FN: مجموعه رکوردهایی که توسط مدل به اشتباه، کلاس منفی پیش بینی شد.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

❑ ارزیابی مدل های رده بندی

- ماتریس درهم ریختگی (Confusion Matrix)



شاخص Recall (بازیابی) یا Sensitivity (حساسیت) نشان دهنده اینست که چه نسبتی از مقادیر واقعی کلاس مثبت به درستی توسط مدل شناسایی و پیش بینی شده است.

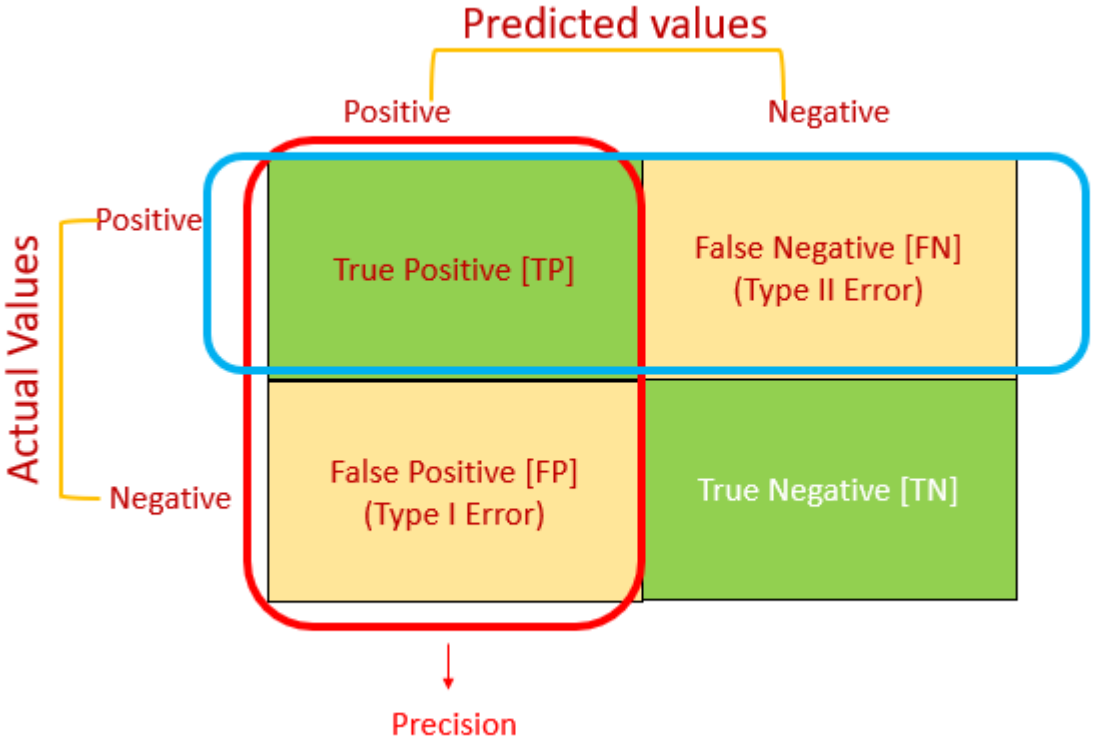
$$Recall = Sensitivity = True\ Positive\ Rate(TPR) = \frac{TP}{TP + FN}$$

شاخص Recall برای کلاس منفی را به عنوان Specificity (ویژگی) می شناسند.

$$Specificity = \frac{TN}{TN + FP}$$

❑ ارزیابی مدل های رده بندی

- ماتریس درهم ریختگی (Confusion Matrix)



نرخ مثبت کاذب، نشان دهنده اینست که چه درصدی از کلاس های منفی به اشتباه کلاس مثبت در نظر گرفته می شوند. این شاخص به عنوان **نرخ هشدار کاذب** (False Alarm Rate) نیز شناخته می شود.

$$False\ Positive\ Rate\ (FPR) = 1 - Specificity = \frac{FP}{TN + FP}$$

شاخص Precision، نشان دهنده اینست که چه نسبتی از مقادیر پیش بینی شده کلاس مثبت صحیح بوده است. (دقت مدل در کلاس مثبت)

$$Precision = \frac{TP}{TP + FP}$$

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

❑ ارزیابی مدل های رده بندی

- ماتریس درهم ریختگی (Confusion Matrix)

برای انتخاب مدل کدام یک از شاخص های Recall یا Precision اهمیت دارد؟

پاسخ به این سوال وابسته به اهمیت هریک از خطاهای نوع اول یا دوم می باشد.

	Predicted class POSITIVE (spam ✉)	Predicted class NEGATIVE (normal 🕵)	
Actual class POSITIVE (spam ✉)	TRUE POSITIVE (TP) ✉ ✉ 320	FALSE NEGATIVE (FN) ✉ 🕵 43	$Recall = \frac{TP}{TP+FN} = \frac{320}{320+43} = 0.882$
Actual class NEGATIVE (normal ✉)	FALSE POSITIVE (FP) ✉ ✉ 20	TRUE NEGATIVE (TN) ✉ 🕵 538	
	$Precision = \frac{TP}{TP+FP} = \frac{320}{320+20} = 0.941$		

مثال1: فرض کنید کلاس مثبت تشخیص یک غده سرطانی بدخیم است؛ در این صورت انتظار داریم مقدار FN برابر با صفر باشد (یعنی مقدار Recall برابر با 100). بنابراین مدلی را انتخاب می کنیم که با دارا بودن این شرط، مقدار Precision را ماکسیمم کند.

مثال2: فرض کنید کلاس مثبت وقوع زلزله شدید باشد؛ در این صورت ترجیح می دهیم برای جلوگیری از وقوع آلارم های اشتباه، مقدار FP از یک مقدار تعیین شده بیشتر نباشد (بطور مثال حداقل 90% Precision) و مدلی را انتخاب می کنیم که با دارا بودن این شرط مقدار Recall را ماکسیمم کند.

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

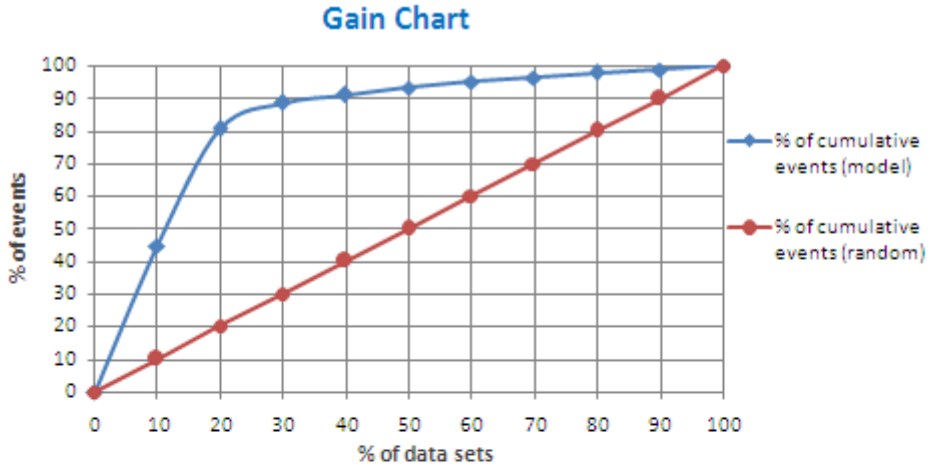
- ارزیابی مدل های رده بندی
  - ماتریس درهم ریختگی (Confusion Matrix)

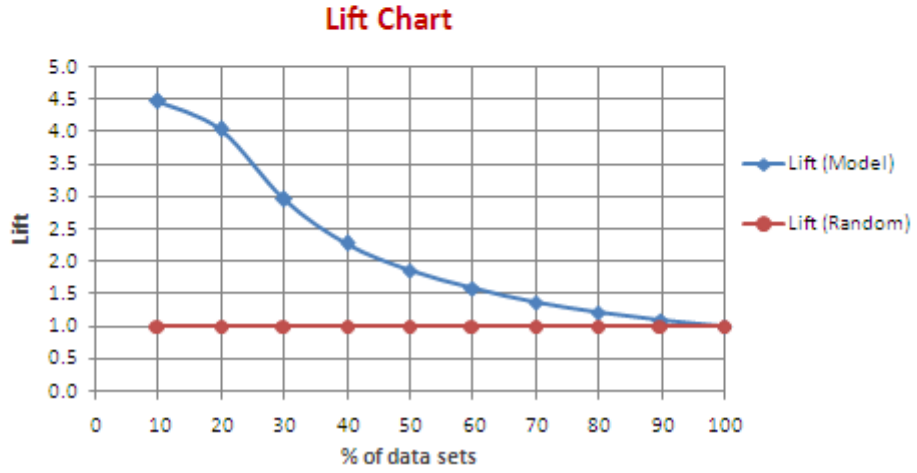
در مواردی که چندین مدل ساخته شده، شرایط حداقلی میزان Recall و Precision را دارا می باشند، می توان از شاخص هیبریدی که بر اساس دو شاخص فوق بدست می آید، به عنوان معیار ارزیابی استفاده نمود:

$$\mathrm{F} - Score = F - Measure = \frac{2 * Recall * Precision}{Recall + Precision}$$

فرآیند داده کاوی

مدل های پیش بینانه – ارزیابی مدل ها





- ارزیابی مدل های رده بندی
  - نمودار بهره و ارتقا (Gain & Lift Chart)

ابزار دیگری که برای ارزیابی مدل های رده بندی استفاده می شود استفاده از روش های بصری Gain Chart و Lift Chart می باشد. بر خلاف ماتریس درهم ریختگی که شاخص های ارزیابی روی کل مجموعه داده ها محاسبه می شدند، این ابزارها روی نسبت های مختلفی از داده ها محاسبه می شوند.

هر دو ابزار Gain Chart و Lift Chart با محاسبه نسبت بین نتایج بدست آمده با مدل و بدون مدل (تصادفی)، اثربخشی مدل را اندازه گیری می کنند.
به عبارتی این ابزارها نشان می دهد آیا استفاده از مدل اثربخشی دارد یا خیر؟

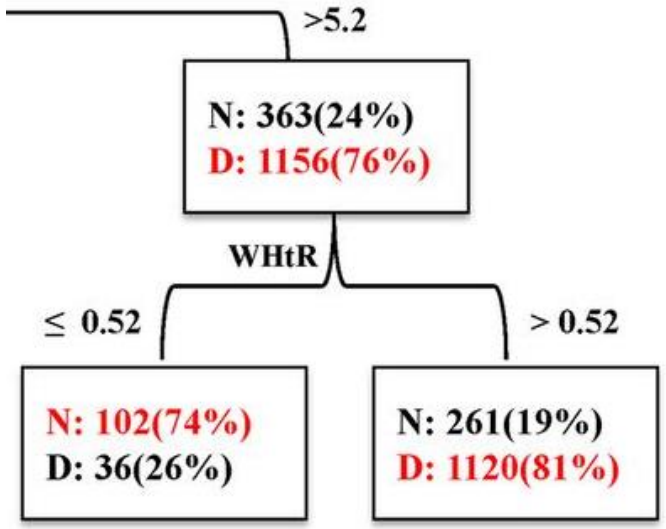
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

- ارزیابی مدل های رده بندی
  - محاسبه و رسم نمودار بهره و ارتقا (Gain & Lift Chart)

گام اول: انتخاب کلاس هدف (hit) برای رسم چارت (بطور مثال کلاس مثبت)

گام دوم: مرتب سازی بزرگ به کوچک کلیه رکوردها بر اساس احتمال دارا بودن کلاس هدف

(بر اساس میزان Confidence هر پیش بینی توسط مدل ساخته شده)



$$Propensity = P(+|A) = \begin{cases} Confidence & Predict = Positive\ Class \\ 1 - Confidence & Predict = Negative\ Class \end{cases}$$

- ارزیابی مدل های رده بندی
  - محاسبه و رسم نمودار بهره و ارتقا (Gain & Lift Chart)

گام سوم: تقسیم بندی کلیه رکوردهای مرتب شده در n دسته برابر (بطور مثال 10 دسته) با حفظ ترتیب آنها

(گام دوم و سوم، در واقع عملیات گسسته سازی رکوردها، روی میزان اطمینان کلاس هدف با روش چندک ها می باشد.)

گام چهارم: محاسبه شاخص Gain و Lift برای هر دسته و محاسبه میزان تجمعی آنها

$$Gain = \frac{Number\ of\ hits\ in\ quantile}{Total\ number\ of\ hits} * 100 \qquad Lift = \frac{Gain}{percent(\%)\ of\ data\ in\ quantiles}$$

گام پنجم: رسم چارت.

فرآیند داده کاوی

مدل های پیش بینانه – ارزیابی مدل ها

- ارزیابی مدل های رده بندی
  - محاسبه و رسم نمودار بهره و ارتقا (Gain & Lift Chart)

مثال:

مجموع رکوردها: 25000

مجموع رکوردهای کلاس هدف: 4874

گسسته سازی: در 10 دهک

	Input Values					
Decile	Number of Cases	Number of Responses	Cumulative Responses	% of events	Gain	Cumulative Lift
1	2500	2179	2179	44.71	44.71	4.47
2	2500	1753	3932	35.97	80.67	4.03
3	2500	396	4328	8.12	88.80	2.96
4	2500	111	4439	2.28	91.08	2.28
5	2500	110	4549	2.26	93.33	1.87
6	2500	85	4634	1.74	95.08	1.58
7	2500	67	4701	1.37	96.45	1.38
8	2500	69	4770	1.42	97.87	1.22
9	2500	49	4819	1.01	98.87	1.10
10	2500	55	4874	1.13	100.00	1.00
	25000	4874				

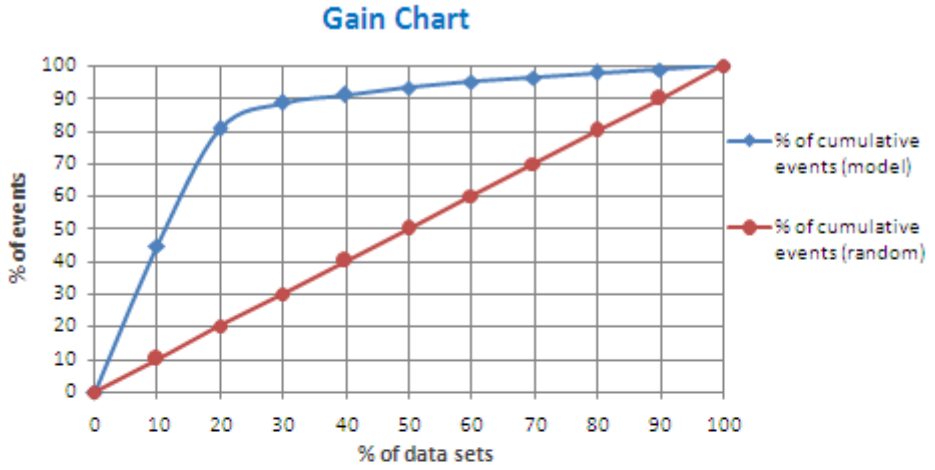
$$Gain = \frac{Number\ of\ hits\ in\ quantile}{Total\ number\ of\ hits} * 100$$

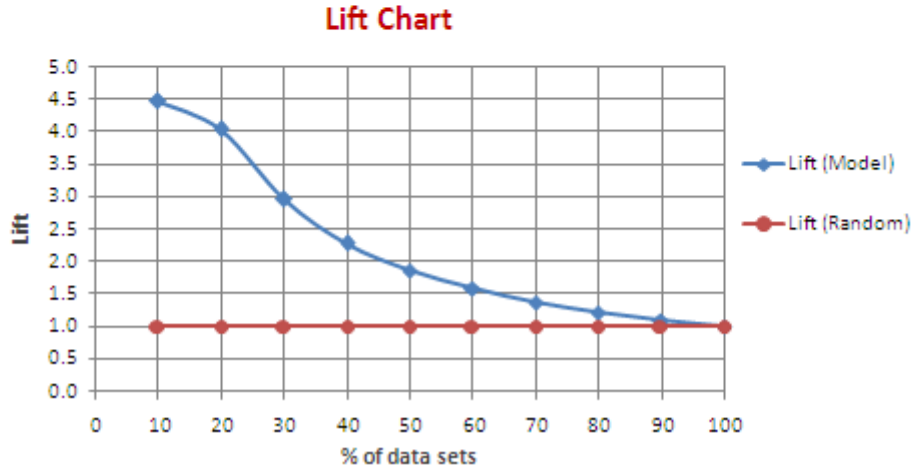
$$Lift = \frac{Gain}{percent(\%)\ of\ data\ in\ quantiles}$$

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – ارزیابی مدل ها





- ارزیابی مدل های رده بندی
  - محاسبه و رسم نمودار بهره و ارتقا (Gain & Lift Chart)

مثال:

مجموع رکوردها: 25000

مجموع رکوردهای کلاس هدف: 4874

گسسته سازی: در 10 دهک

$$Gain = \frac{Number\ of\ hits\ in\ quantile}{Total\ number\ of\ hits} * 100$$

$$Lift = \frac{Gain}{percent(\%)\ of\ data\ in\ quantiles}$$

تولید محتوا: زهرا ذوالقدر

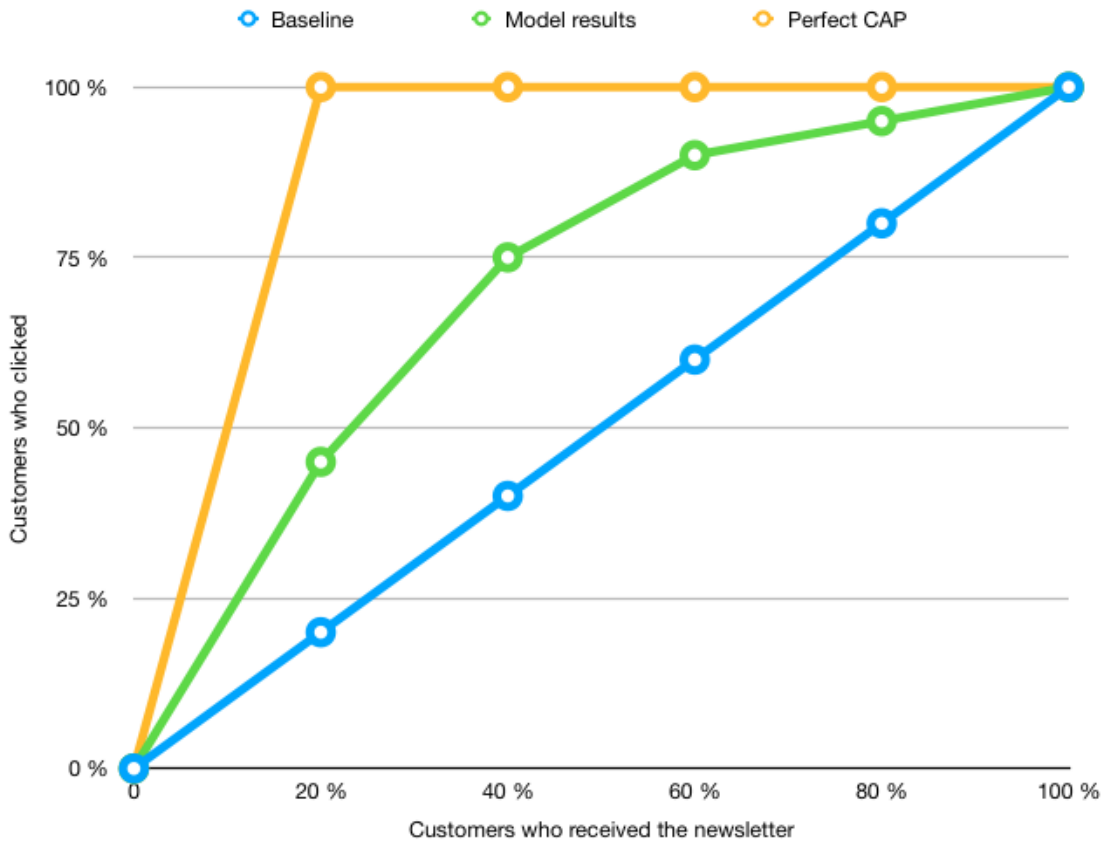
daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – ارزیابی مدل ها



- ارزیابی مدل های رده بندی
  - محاسبه و رسم نمودار بهره و ارتقا (Gain & Lift Chart)
    - محاسبه Baseline

در این حالت با فرض برچسب گذاری تصادفی، میزان احتمال وقوع کلاس هدف در مجموعه داده ها محاسبات مربوط به شاخص های Gain و Lift انجام می شود. **بطور مثال** در صورتی که 20% داده ها کلاس مثبت باشند، در هر دسته به میزان 20% از داده ها برچسب مثبت داده می شود.

    - محاسبه Best Line (Perfect CAP)

در این حالت مدل با خطای صفر در نظر گرفته می شود و در هر دسته تمامی رکوردهای ممکن دارای برچسب کلاس هدف در نظر گرفته می شود تا جایی که تعداد رکوردهای کلاس هدف به اتمام برسد.

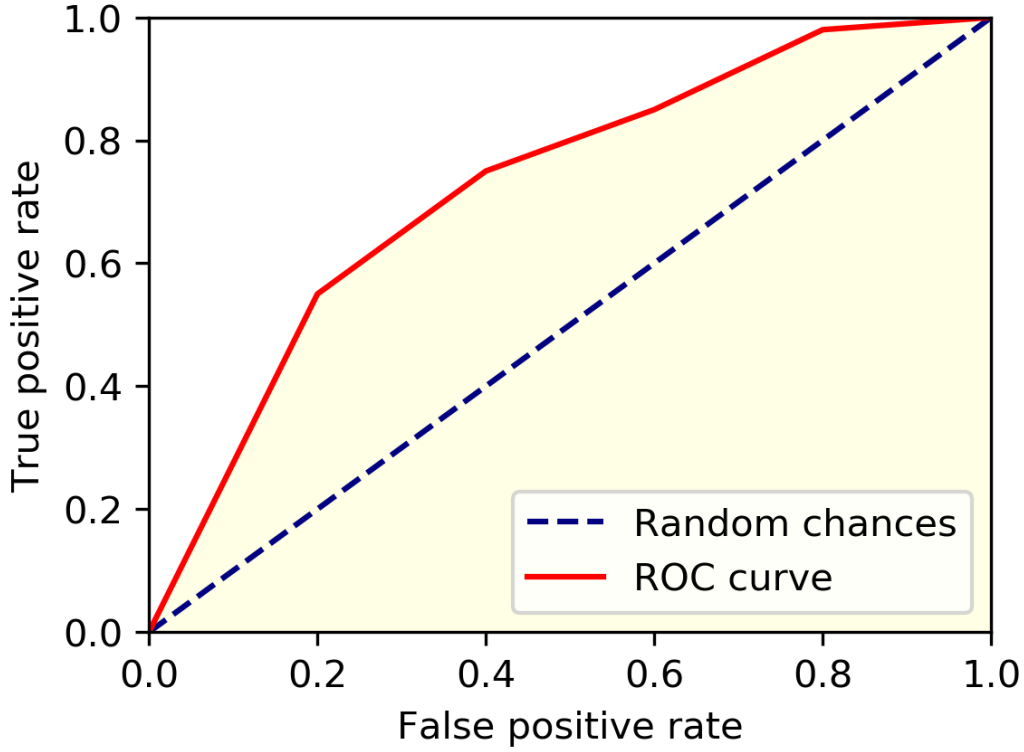
فرآیند داده کاوی

مدل های پیش بینانه – ارزیابی مدل ها

- ارزیابی مدل های رده بندی
  - منحنی ROC (Receiver Operating Characteristics curve)

یکی از رایج ترین ابزارهای بصری جهت ارزیابی کارایی مدل های رده بندی استفاده از منحنی ROC می باشد.

در این ابزار، مقادیر TPR (نرخ مثبت هایی که به درستی توسط مدل شناسایی شدند - شاخص Recall/Sensitivity) به عنوان نقطه قوت مدل (میزان سود) در مقابل مقادیر FPR (نرخ مثبت کاذب – 1-Specificity) به عنوان نقطه ضعف مدل (میزان هزینه) قرار می گیرد و برای ارزیابی مدل به دنبال مصالحه بین سود و هزینه می باشد.



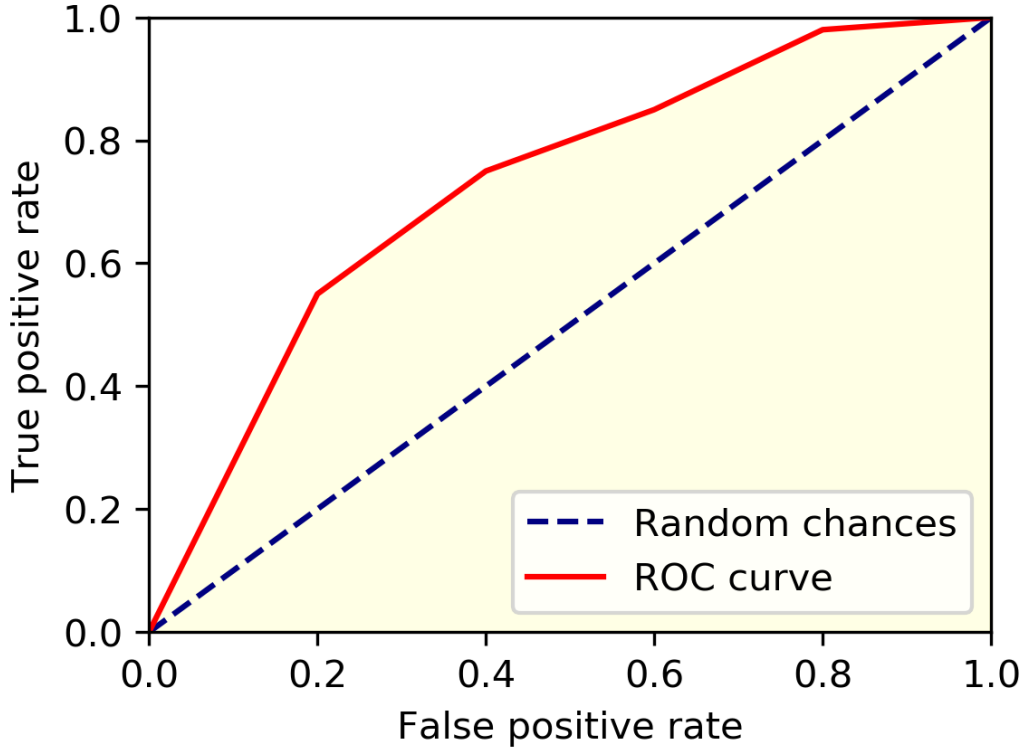
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

ارزیابی مدل های رده بندی

- منحنی ROC (Receiver Operating Characteristics curve)

یکی از رایج ترین ابزارهای بصری جهت ارزیابی کارایی مدل های رده بندی استفاده از منحنی ROC می باشد.

در این نمودار، خط نیمساز به عنوان مدل انتخاب تصادفی است و هرچه منحنی ROC یک مدل، فاصله بیشتری از آن بگیرد نشان دهنده کارایی بیشتر آن مدل است.



تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – ارزیابی مدل ها

- ارزیابی مدل های رده بندی
  - محاسبه و رسم منحنی ROC

برای محاسبه نقاط منحنی ROC، بایستی مقادیر مختلفی از مقادیر TPR و FPR به ازای هر مدل محاسبه شود.

مثال: فرض کنید برای 10 رکورد جدول فوق، با استفاده از مدل رده بند A، احتمال تعلق به کلاس مثبت برای تمام رکوردها محاسبه شده است:

Instance	P(+|A)	True Class
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

گام اول: مرتب سازی رکوردها بر اساس احتمال تعلق به کلاس مثبت

Source: Tan, Steinbach, Kumar

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – ارزیابی مدل ها

- ارزیابی مدل های رده بندی
  - محاسبه و رسم منحنی ROC

گام دوم: با در نظر گرفتن هر یک از مقادیر احتمال تعلق به کلاس مثبت، به عنوان حد آستانه ای (برش) برای تفکیک کلاسهای مثبت و منفی، ماتریس درهم ریختگی مربوط به آن ایجاد و بر اساس آن مقادیر شاخص های TPR و FPR محاسبه می شود.

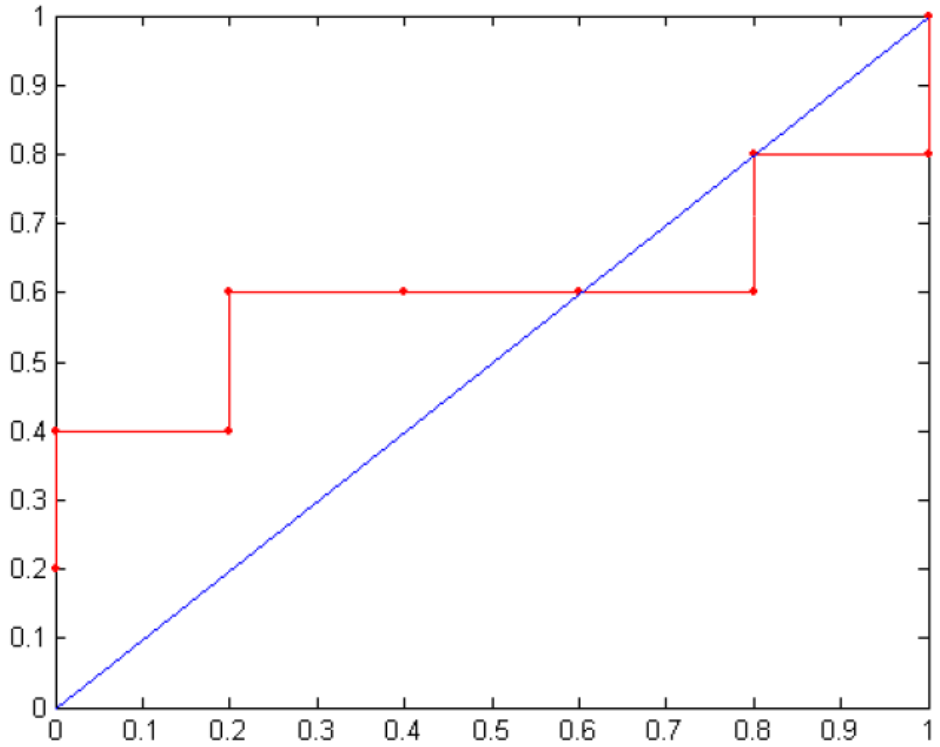
Class	+	-	+	-	-	-	+	-	+	+	
	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
TP	5	4	4	3	3	3	3	2	2	1	0
FP	5	5	4	4	3	2	1	1	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5
TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – ارزیابی مدل ها

- ارزیابی مدل های رده بندی
  - محاسبه و رسم منحنی ROC



گام سوم: رسم منحنی ROC بر اساس نقاط TPR و FPR به تفکیک هر حد آستانه ای (برش) و اتصال آنها

بدین ترتیب استفاده از نمودار ROC در تعیین بهترین نقطه آستانه ای برای تفکیک کلاس ها نیز مفید هست.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

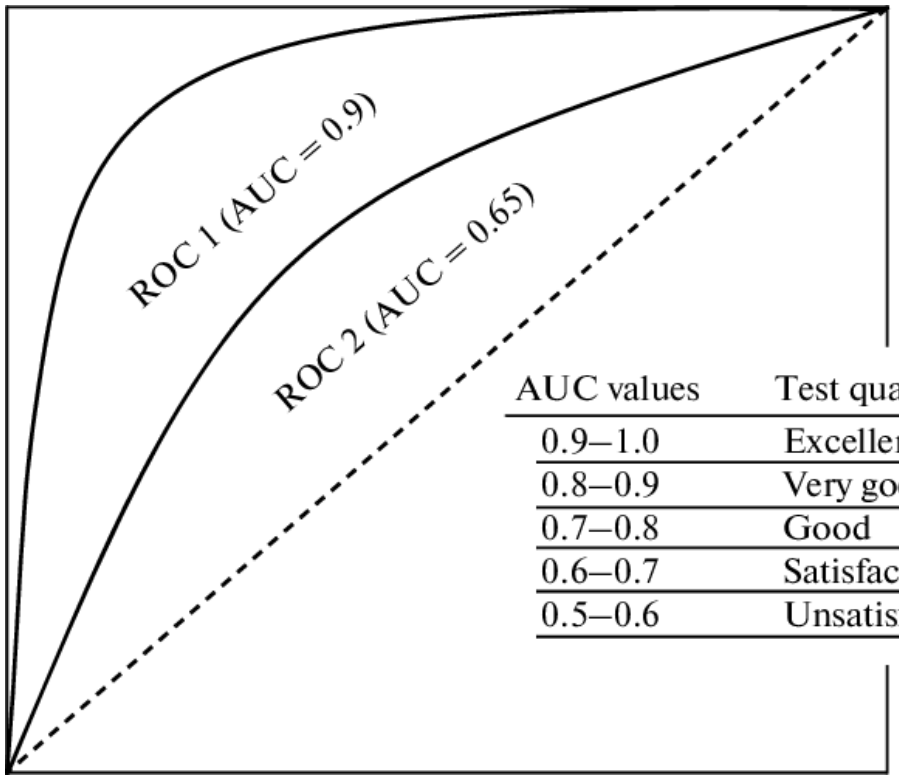
فرآیند داده کاوی

مدل های پیش بینانه – ارزیابی مدل ها

- ارزیابی مدل های رده بندی
  - شاخص Area Under Curve (AUC)

شاخص AUC منحنی ROC را در قالب یک مقدار کمی اندازه گیری کرده و قابلیت مقایسه بین مدل ها را فراهم می کند.

با توجه به اینکه مساحت زیر خط نیمساز برابر با عدد 0.5 می باشد، بنابراین مدل مطلوب بایستی مقداری بیش از 0.5 داشته باشد. بدیهی هست در بهترین حالت که مقدار FPR برابر با صفر و مقدار TPR برابر با یک باشد، مقدار بیشینه مساحت زیر منحنی برابر با یک خواهد بود و عدد AUC در اغلب مدلها در بازه (0,1) قرار می گیرند.

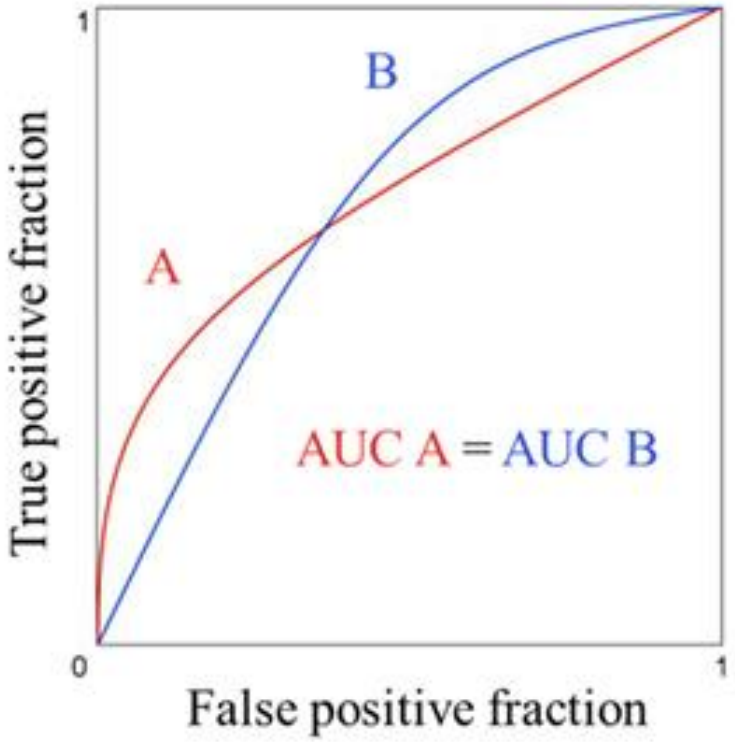


AUC values	Test quality
0.9–1.0	Excellent
0.8–0.9	Very good
0.7–0.8	Good
0.6–0.7	Satisfactory
0.5–0.6	Unsatisfactory

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ ارزیابی مدل های رده بندی

- شاخص Area Under Curve (AUC)



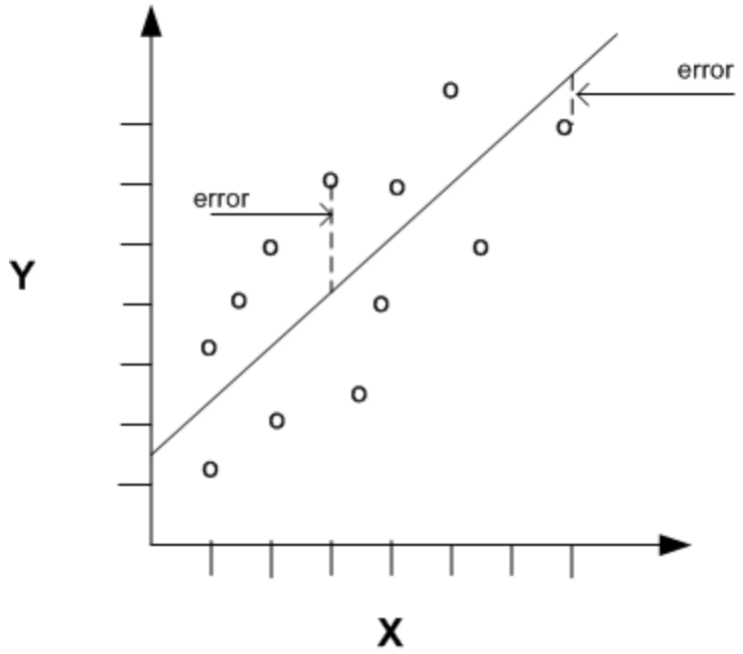
استفاده از شاخص AUC برای مسائلی که توزیع فیلد هدف نامتوازن می باشد، می تواند گمراه کننده باشد. در این نوع مسائل استفاده از شاخص های Recall و Precision می تواند گزینه بهتری برای ارزیابی مدل ها باشد.

مثال: میزان AUC در مدل های A و B برابر است؛ اما به وضوح می توان دید در مسائلی که نرخ هشدار اشتباه دارای هزینه زیادی هست مدل A دارای نتیجه بهتری نسبت به مدل B می باشد. همچنین مدل B نیز در مسائلی که دسترسی به کلاس مثبت اهمیت بالایی دارد نسبت به مدل A دارای ارجحیت است.

فرآیند داده کاوی

مدل های پیش بینانه – ارزیابی مدل ها

X1	X2	X3	X4	Target	Partition	Pred.	Error
.	.	.	.	75	Train	69	6
.	.	.	.	68	Train	72	-4
.	.	.	.	92	Test	94	-2
.	.	.	.	77	Train	80	-3
.	.	.	.	84	Test	81	3
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.



ارزیابی مدل های رگرسیونی

بر خلاف مدل های رده بندی، در مسائل پیش بینی داده های کمی، شاخص هایی مانند صحت مدل (با مفهومی که در مسائل رده بندی آشنا شدیم) وجود ندارد. بلکه به دنبال معیارهایی برای اندازه گیری میزان نزدیکی مقدار پیش بینی شده با مقدار واقعی هستیم.

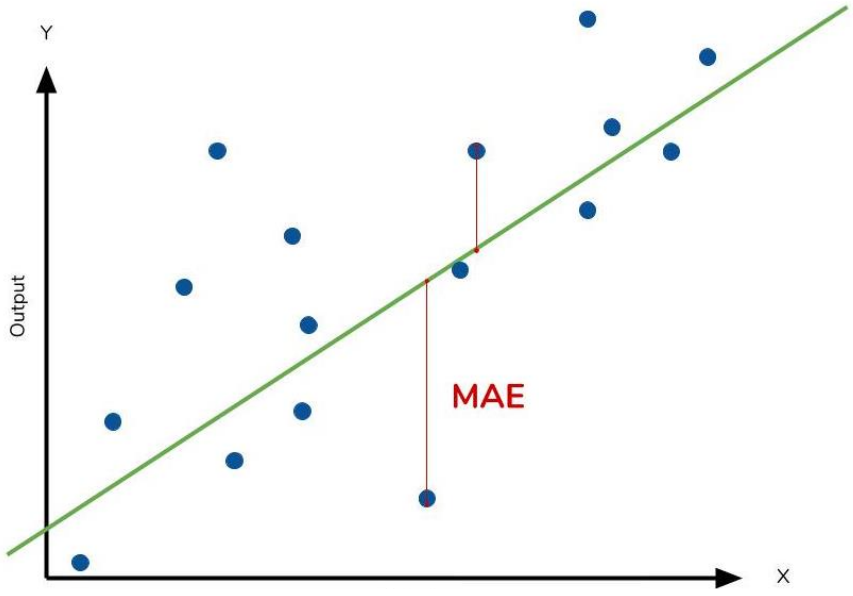
به همین دلیل پایه اغلب شاخص های ارزیابی مدل های رگرسیونی، بر مبنای محاسبه اندازه خطای پیش بینی می باشد.

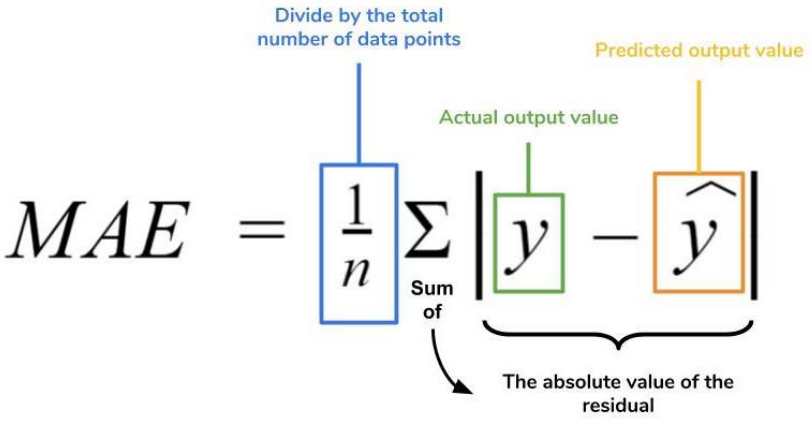
Error = Real Value – Predicted Value

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – ارزیابی مدل ها





ارزیابی مدل های رگرسیونی

- شاخص Mean Absolute Error (MAE)

شاخص MAE یکی از ساده ترین معیارهای ارزیابی مدلهای رگرسیونی می باشد. این شاخص متوسط اندازه خطای پیش بینی را محاسبه می کند و تفسیر خوبی از میزان اثربخشی مدل فراهم می کند.

نکته: با حذف قدر مطلق و محاسبه میانگین خطا (Mean Error)، با توجه به گذر خط رگرسیونی از مرکز ثقل داده ها، بایستی عدد صفر بدست آید. در صورت محاسبه مقدار کوچکتر یا بزرگتر از صفر، می توان گفت مدل دارای سوگیری (Bias) به سمت بیش برآوردی یا کم برآوردی می باشد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

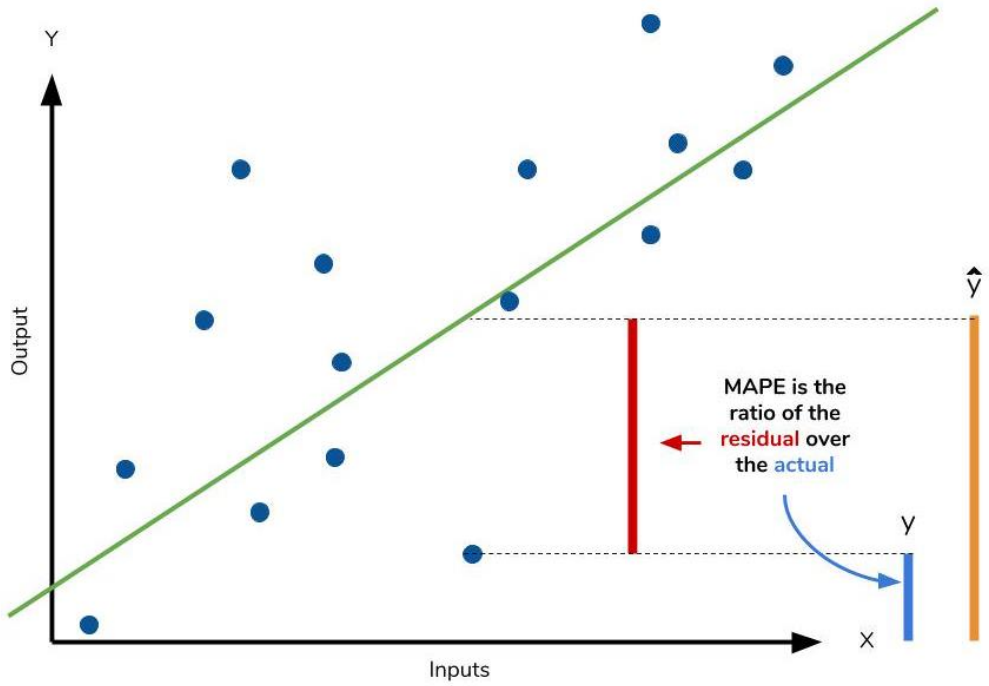
مدل های پیش بینانه – ارزیابی مدل ها

ارزیابی مدل های رگرسیونی

- شاخص Mean Absolute Percentage Error (MAPE)

شاخص MAPE شباهت زیادی به MAE دارد، با این تفاوت که اندازه خطای پیش بینی در این شاخص بصورت درصدی از مقدار واقعی بیان می شود. به همین علت تفسیر راحت تری از ارزیابی مدل ایجاد می کند.

در صورت برداشتن قدر مطلق از رابطه، می توان جهت مثبت و منفی خطا را در مقدار شاخص Mean Percentage Error (MPE) مشاهده کرد و مشابه شاخص Mean Error در خصوص امکان سوگیری مدل قضاوت نمود.



$$MAPE = \frac{100\%}{n} \sum \left| \frac{y - \hat{y}}{y} \right|$$

Multiplying by 100% converts to percentage

The residual

Each residual is scaled against the actual value

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ ارزیابی مدل های رگرسیونی

- شاخص Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{100\%}{n} \sum \left| \frac{y - \hat{y}}{y} \right|$$

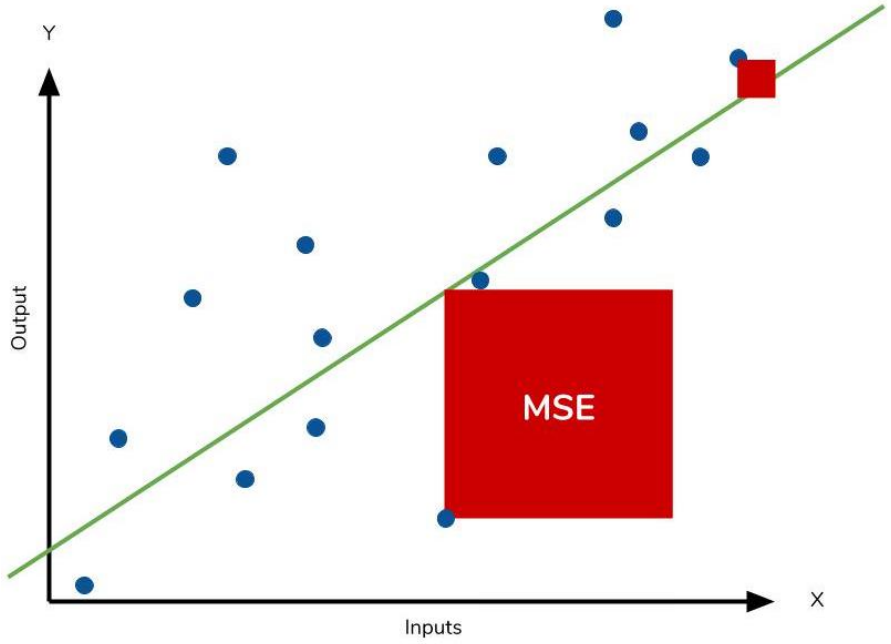
در استفاده از شاخص MAPE به دو مورد بایستی توجه نمود:

- در صورت صفر بودن مقدار واقعی در یک رکورد، مقدار MAPE قابل محاسبه نمی باشد.
- درصد خطای پیش بینی محاسبه شده در این شاخص در ارتباط با اندازه مقدار واقعی می تواند تفاوت زیادی ایجاد نماید.

$\hat{y}$ is smaller than the actual value	$\hat{y}$ is greater than the actual value
n = 1 $\hat{y}$ = 10 y = 20	n = 1 $\hat{y}$ = 20 y = 10
MAPE = 50%	MAPE = 100%

فرآیند داده کاوی

مدل های پیش بینانه – ارزیابی مدل ها



- ارزیابی مدل های رگرسیونی
  - شاخص Mean Square Error (MSE)

در محاسبه شاخص MSE از میانگین توان دوم خطای پیش بینی استفاده می شود. به همین دلیل بر خلاف شاخص MAE که در محاسبه آن تمام خطاها هم وزن بوده اند، در این شاخص هر چه میزان خطای پیش بینی بزرگتر باشد، وزن بیشتری در میزان MSE خواهد داشت.

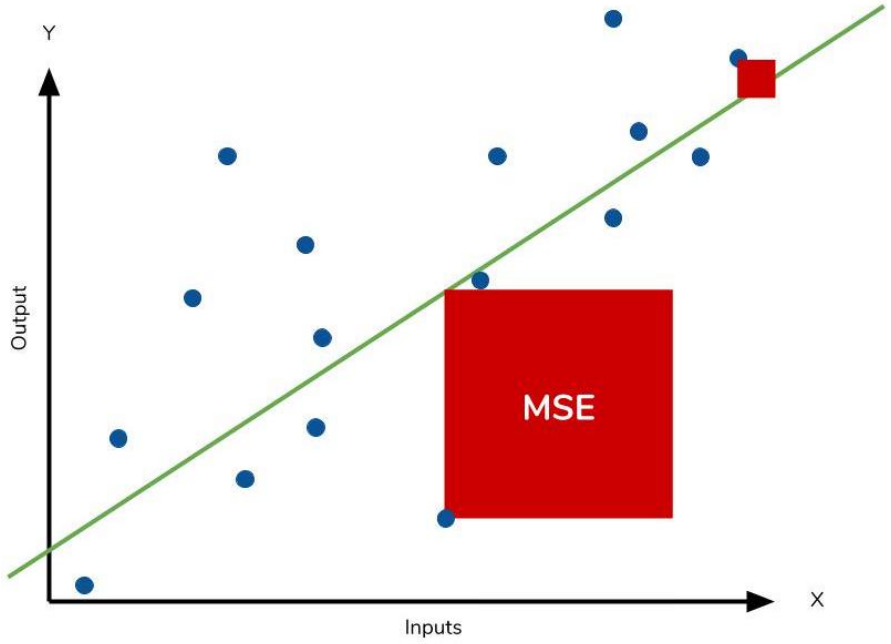
در واقع این شاخص حساسیت بیشتری به مقادیر پرت نشان می دهد.

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – ارزیابی مدل ها



- ارزیابی مدل های رگرسیونی
  - شاخص Mean Square Error (MSE)

انتخاب بین شاخص های MAE و MSE کاملا به اهداف مسئله وابسته هست. استفاده از شاخص MSE این اطمینان را به ما می دهد که مدل حساسیت لازم در ارزیابی مقادیر پرت را داراست.

با توجه به اینکه توان دوم خطای پیش بینی، منجر به ایجاد داده های بزرگ شده و مقیاس داده ها را تغییر می دهد، معمولا برای تفسیر بهتر نتایج ارزیابی از جذر این شاخص به عنوان Root Mean Square Error (RMSE) استفاده می شود.

$$MSE = \frac{1}{n}\sum\left(\underbrace{y - \hat{y}}\right)^2$$

The square of the difference between actual and predicted

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

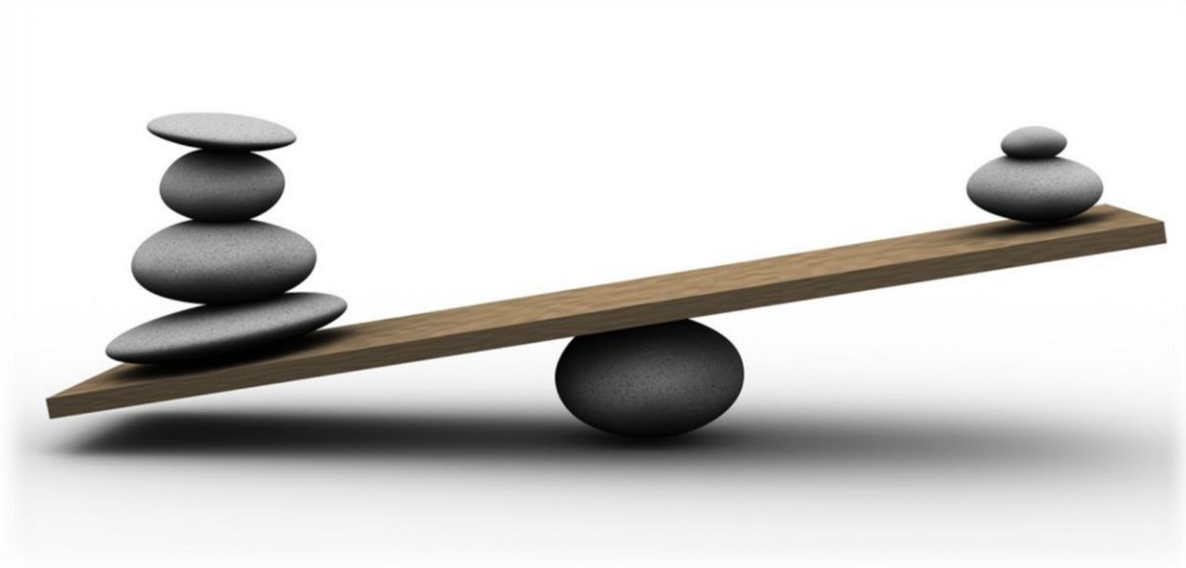
فرآیند داده کاوی

مدل های پیش بینانه – داده های نامتوازن

❑ چالش رده بندی داده های نامتوازن (Imbalanced Dataset)

یکی از مسائل رایج در رده بندی، توزیع نامتوازن در کلاس های فیلد هدف می باشد. در بسیاری از داده ها، از جمله داده های تخلفات مالی یا بیمه ای، بیماری های خاص، پدیده های نادر و ... نمی توان انتظار داشت فراوانی همه کلاس ها به یک اندازه باشد.

این موضوع باعث می شود بسیاری از الگوریتم ها، برچسب کلاس های نادر را به عنوان رکوردهای پرت و یا حتی خطای پیش بینی نادیده بگیرند، در صورتیکه معمولا، این کلاس های نادر اهمیت بالایی در مسئله دارند.



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – داده های نامتوازن

❑ چالش رده بندی داده های نامتوازن (Imbalanced Dataset)

بطور کلی شدت نامتوازن بودن داده ها پیچیدگی مسئله را افزایش می دهد و اغلب الگوریتم ها به علت اینکه به دنبال افزایش صحت کلی مدل (Accuracy) یا کاهش خطا هستند عموما در داده های نامتوازن عملکرد خوبی را نخواهند داشت.

نسبت کلاس اقلیّت	درجه نامتعادلی
20-40%	Mild
1-20%	Moderate
<1%	Extreme

بطور مثال فرض کنید برای حل مسئله تشخیص تخلف در یک شرکت بیمه، مجموعه داده ای در اختیار شما گذاشته شده است و پس از یک آماده سازی اولیه بر روی داده، الگوریتم درخت تصمیم را روی داده ها آموزش دادید و از نتیجه مدل شگفت زده شدید: **99.5% صحت مدل در پیش بینی!!**

ولی با کمی دقت خواهید دید که مدل شما به همه رکوردها برچسب سالم داده، و 99.5% از داده های شما برچسب سالم داشته اند! در واقع **مدلی با ارزش صفر.**

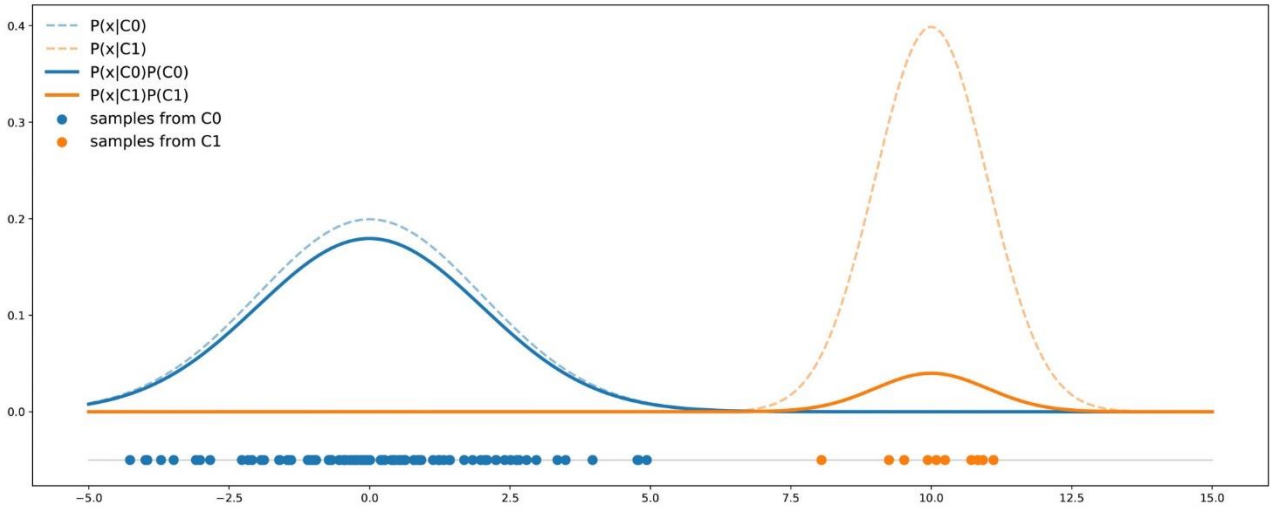
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – داده های نامتوازن

❑ چالش رده بندی داده های نامتوازن (Imbalanced Dataset)

برای فائق آمدن بر چالش داده های نامتوازن، روش های متنوعی وجود دارد که در اجرای پروژه های داده کاوی سعی می شود از ساده ترین روشها برای حل آن استفاده شود و در صورت نیاز به سمت روش های پیچیده تر رفت.



- **مرحله اول: هیچ کار اضافه ای نکنید!**

شاید خوش شانس باشیم و فارغ از شدت نامتوازن بودن داده ها، مدل های بدست آمده از ویژگی های در دسترس، به خوبی کلاس های نامتوازن را تفکیک کنند.

- **مرحله دوم: ویژگی موثر دیگری به داده ها اضافه کنید!**

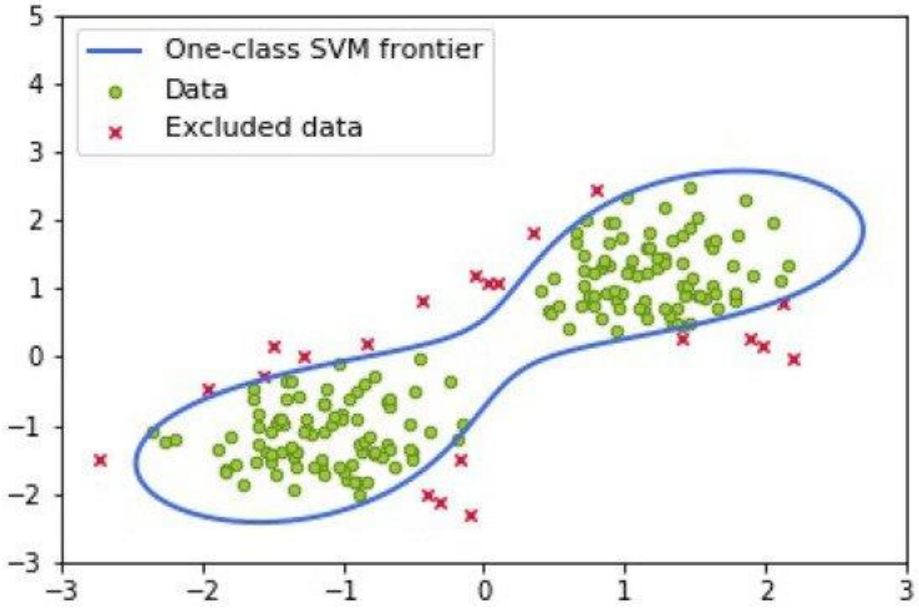
اگر به اندازه مرحله اول خوش شانس نبودیم، بررسی کنید آیا امکان اضافه کردن ویژگی موثری که تفکیک پذیری کلاس ها را ارتقا دهد وجود دارد.

تولید محتوا: زهرا ذوالقدر

- ## چالش رده بندی داده های نامتوازن (Imbalanced Dataset)

  - ### مرحله سوم: به داده های کلاس حداقلی اضافه کنید!

شاید امکان دسترسی به داده های جدید با کلاس اقلیت وجود داشته باشد. در اینصورت بهتر هست ادامه حل مسئله پس از تکمیل داده ها و متعادل کردن آنها انجام پذیرد.



  - ### مرحله چهارم: رویکرد حل مسئله را می توان تغییر داد!

گاهی وقت ها باید واقعیت را پذیرفت و شرایط مسئله را از زاویه دیگری نگاه کرد.

بطور مثال، تبدیل مسئله رده بندی به مسئله شناسایی انحرافات

(یا Anomaly Detection) می تواند یکی از راهکارهای موجود برای مدلسازی

و حل مسئله در نظر گرفته شود.

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – داده های نامتوازن

- چالش رده بندی داده های نامتوازن (Imbalanced Dataset)
  - مرحله پنجم: استفاده از رویکردهای مواجهه با داده های نامتوازن

پس از بررسی چهار مرحله قبل و در صورت تداوم مشکل، سه رویکرد عمده در حل مسئله رده بندی داده های نامتوازن وجود دارد:

رویکرد مبتنی بر
نمونه گیری

رویکرد مبتنی بر
تابع هزینه

رویکرد مبتنی بر
معماری الگوریتم

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

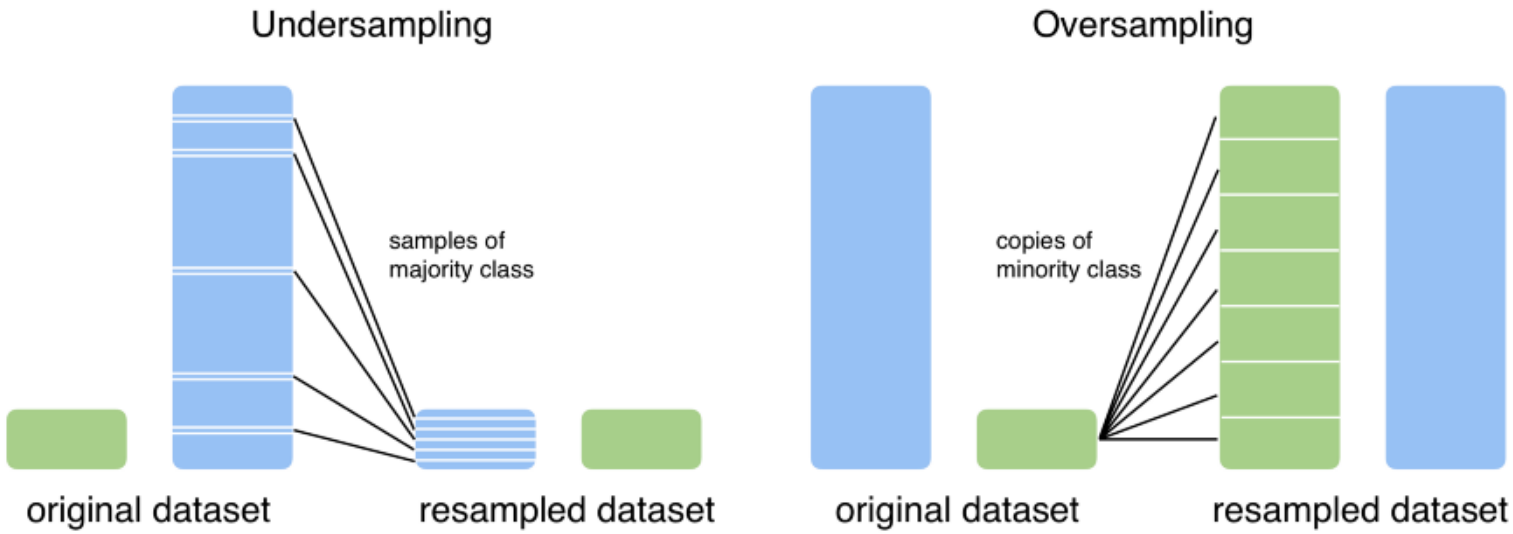
فرآیند داده کاوی

مدل های پیش بینانه – داده های نامتوازن

- رویکردهای مواجهه با داده های نامتوازن
  - رویکرد مبتنی بر نمونه گیری

هدف از تکنیک های مورد استفاده در این رویکرد، متعادل سازی توزیع کلاس های فیلد هدف می باشد.

- Under-Sampling
- Over-Sampling
- Hybrid Approach



نکته مهم: تغییر توزیع داده ها فقط و فقط در مجموعه داده های آموزشی برای ساخت مدل انجام می پذیرد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

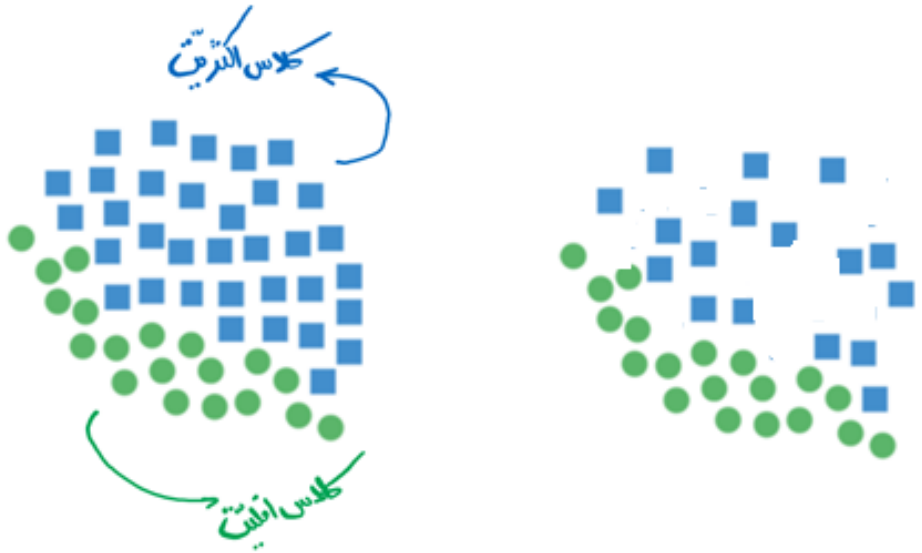
مدل های پیش بینانه – داده های نامتوازن

- رویکردهای مواجهه با داده های نامتوازن
  - رویکرد مبتنی بر نمونه گیری

کاهش رکوردهای کلاس اکثریت به روش های مختلفی قابل انجام هست. سه روش رایج به شرح زیر است:

روش حذف تصادفی (Random Under-Sampling)

در این روش، از طریق نمونه گیری تصادفی ساده، تعداد رکوردهای کلاس اکثریت به میزان (یا نزدیک به) تعداد رکوردهای کلاس اقلیت می رسد.



Under-Sampling

Over-Sampling

Hybrid Approach

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – داده های نامتوازن

- Under-Sampling
- Over-Sampling
- Hybrid Approach

رویکردهای مواجهه با داده های نامتوازن

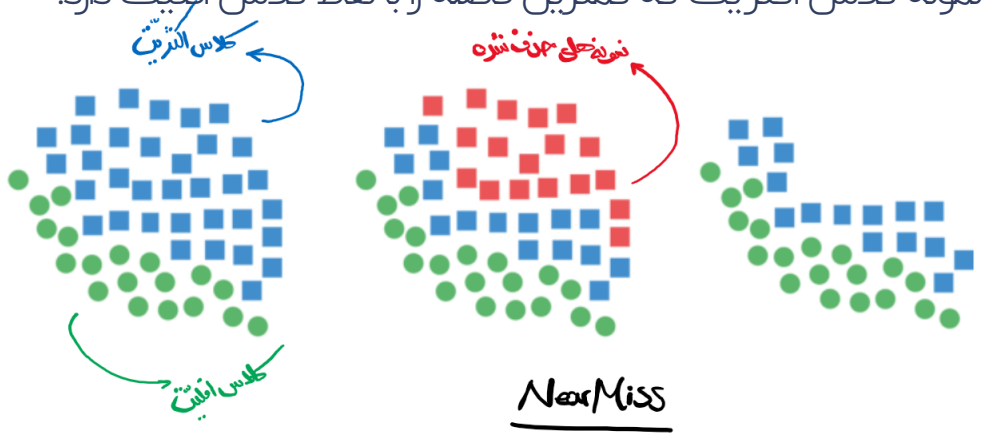
- رویکرد مبتنی بر نمونه گیری

کاهش رکوردهای کلاس اکثریت به روش های مختلفی قابل انجام هست. سه روش رایج به شرح زیر است:

روش حذف نزدیک ترین همسایه ها (Near Miss Under-Sampling)

در این روش، بصورت غیر تصادفی و با الگوی زیر رکوردهای کلاس اکثریت کاهش می یابد:

- محاسبه فواصل بین تمام نمونه‌های کلاس اکثریّت و کلاس اقلیّت
- شناسایی و نگهداشت k نمونه کلاس اکثریّت که کمترین فاصله را با نقاط کلاس اقلیّت دارد.



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – داده های نامتوازن

- Under-Sampling
- Over-Sampling
- Hybrid Approach

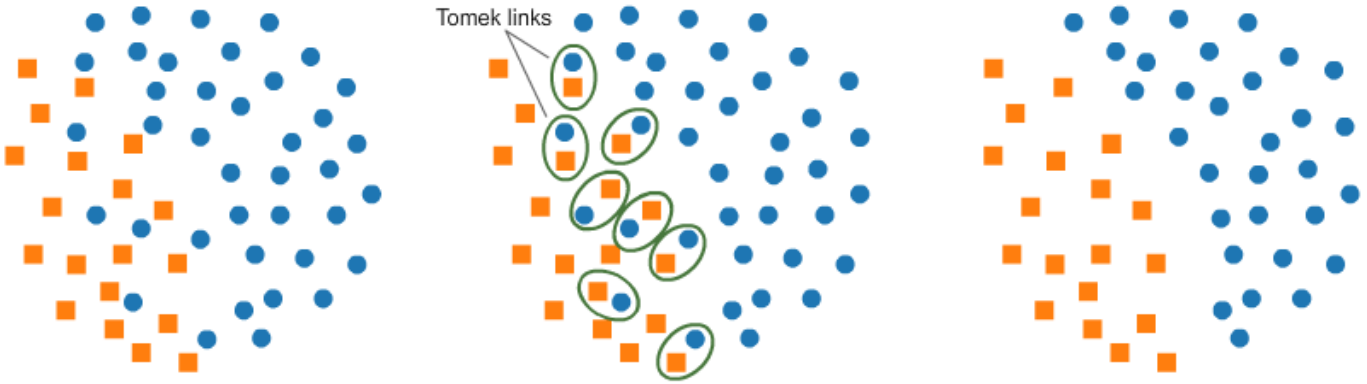
رویکردهای مواجهه با داده های نامتوازن

- رویکرد مبتنی بر نمونه گیری

کاهش رکوردهای کلاس اکثریت به روش های مختلفی قابل انجام هست. سه روش رایج به شرح زیر است:

روش حذف TomeKLinks

- شناسایی جفت نمونه‌هایی در داده‌ها که هر کدام به کلاس متفاوتی تعلق دارند.
  (این جفت نمونه‌ها در اصل در نزدیکی مرز بین دو کلاس قرار دارند)
- حذف نمونه‌های کلاس اکثریت در این جفت‌ها.
  (علاوه بر متعادل شدن تعداد نمونه‌ها، مرز بین دو کلاس هم افزایش می‌یابد)



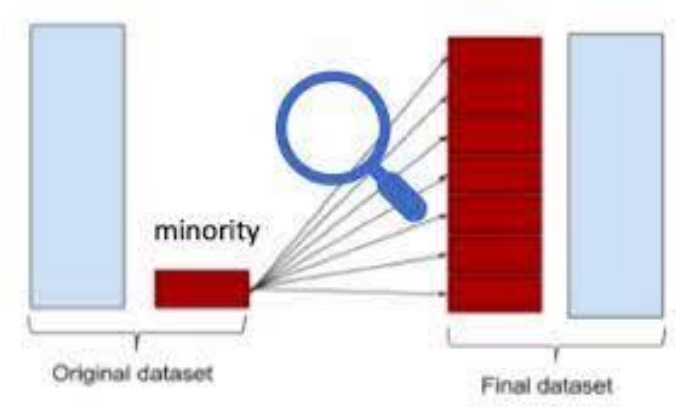
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

- رویکردهای مواجهه با داده های نامتوازن
  - رویکرد مبتنی بر نمونه گیری

افزایش رکوردهای کلاس اقلیت به روش های مختلفی قابل انجام هست. سه روش رایج به شرح زیر است:

روش افزایش تصادفی (Random Over-Sampling)

در این روش، از طریق نمونه گیری تصادفی با جایگذاری (بوت استرپ Bootstrap)، با تکرار رکوردهای کلاس اقلیت، تعداد آنها به میزان (یا نزدیک به) تعداد رکوردهای کلاس اکثریت می رسد.



Under-Sampling

Over-Sampling

Hybrid Approach

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

- Under-Sampling
- Over-Sampling
- Hybrid Approach

رویکردهای مواجهه با داده های نامتوازن

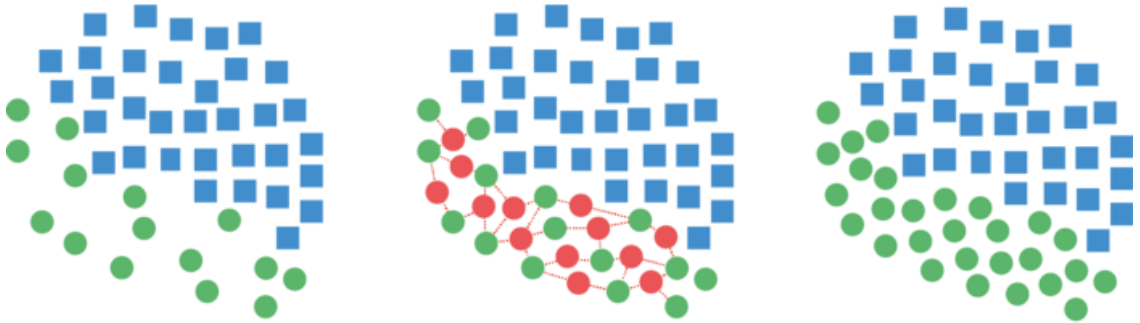
رویکرد مبتنی بر نمونه گیری

افزایش رکوردهای کلاس اقلیت به روش های مختلفی قابل انجام هست. سه روش رایج به شرح زیر است:

روش نمونه سازی Synthetic Minority Oversampling Technique (SMOTE)

در این روش، از طریق ساخت نمونه های شبیه به کلاس اقلیت، تعداد آن افزایش می یابد تا توازن کلاس ها برقرار گردد.

- ابتدا k-نزدیک‌ترین همسایه نمونه‌های کلاس اقلیّت برای هر نمونه از کلاس اقلیّت مشخص می‌شوند
- برای هر نمونه کلاس اقلیت به صورت تصادفی یکی از همسایه‌ها انتخاب می‌شود
- با استفاده از درون‌یابی (Interpolation) یک نمونه جدید بین دو نمونه مذکور ایجاد می‌کنیم



- Under-Sampling
- Over-Sampling
- Hybrid Approach

رویکردهای مواجهه با داده های نامتوازن

- رویکرد مبتنی بر نمونه گیری

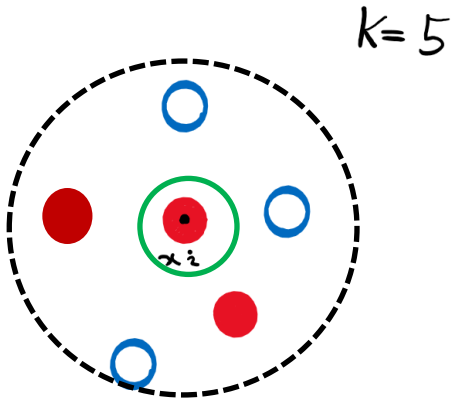
افزایش رکوردهای کلاس اقلیت به روش های مختلفی قابل انجام هست. سه روش رایج به شرح زیر است:

روش نمونه سازی ADASYN

در این روش، با استفاده از محاسبه توزیع چگالی داده های اقلیت، عملیات نمونه سازی را برای نمونه هایی از کلاس اقلیت که برای یادگیری مدل سخت تر هستند، انجام می دهد.

محاسبه چگالی:

(در همسایگی یک رکورد اقلیت)



$$r_i = \frac{\#\text{کلاس اکثریت}}{K} = \frac{3}{5} = 0.6$$

فرآیند داده کاوی

مدل های پیش بینانه – داده های نامتوازن

- Under-Sampling
- Over-Sampling
- Hybrid Approach

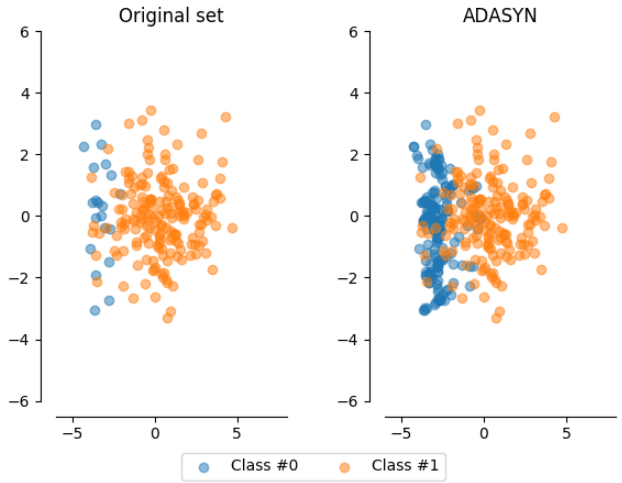
رویکردهای مواجهه با داده های نامتوازن

- رویکرد مبتنی بر نمونه گیری

افزایش رکوردهای کلاس اقلیت به روش های مختلفی قابل انجام هست. سه روش رایج به شرح زیر است:

روش نمونه سازی ADASYN

- در این روش k-نزدیک‌ترین همسایه از کل داده‌ها برای هر نمونه از کلاس اقلیّت مشخص می‌شود.
- سپس برای هر نمونه از کلاس اقلیّت چگالی کلاس اکثریت در همسایگی آن (r_i) محاسبه می‌شود.
- تنها روی نمونه‌های کلاس اقلیّتی که یادگیری آن‌ها برای مدل سخت‌تر است (نمونه‌های مرزی)، نمونه جدید ساخته می‌شود.



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – داده های نامتوازن

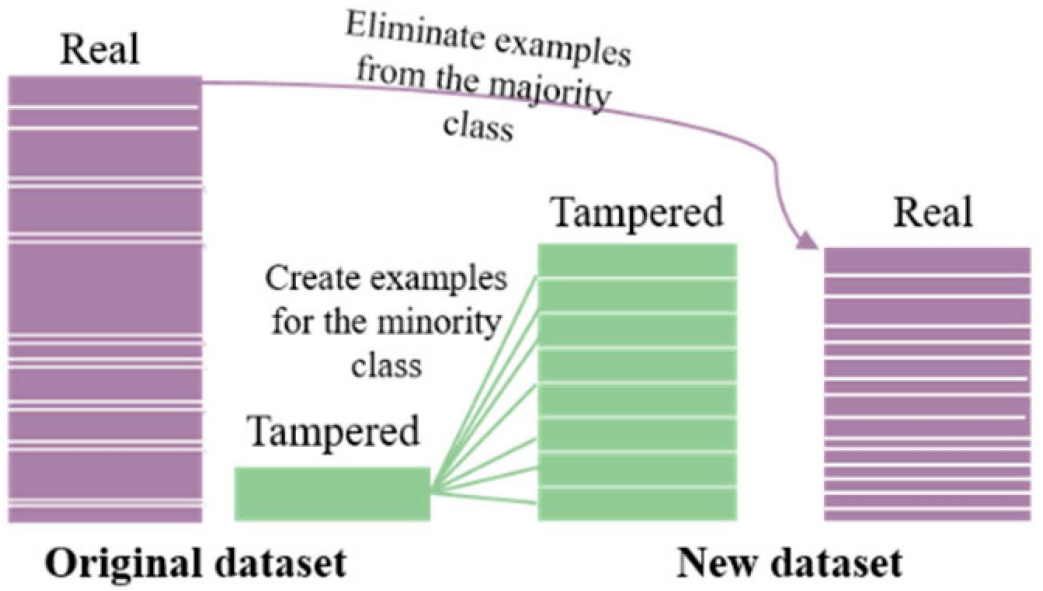
- رویکردهای مواجهه با داده های نامتوازن
  - رویکرد مبتنی بر نمونه گیری

در بسیاری از مسائلی که شدت نامتوازن بودن داده ها زیاد باشد، استفاده از روش های ترکیبی از رویکردهای کاهش داده های اکثریت و افزایش داده های اقلیت مورد استفاده قرار می گیرد.

Under-Sampling

Over-Sampling

Hybrid Approach



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
گروه دایکه | dayche.com

❑ رویکردهای مواجهه با داده های نامتوازن

○ رویکرد مبتنی بر تابع هزینه

در این رویکرد با تغییر تابع هزینه خطاهای پیش بینی، روند یادگیری مدل به سمت پیش بینی کلاس اقلیت سوق داده می شود.

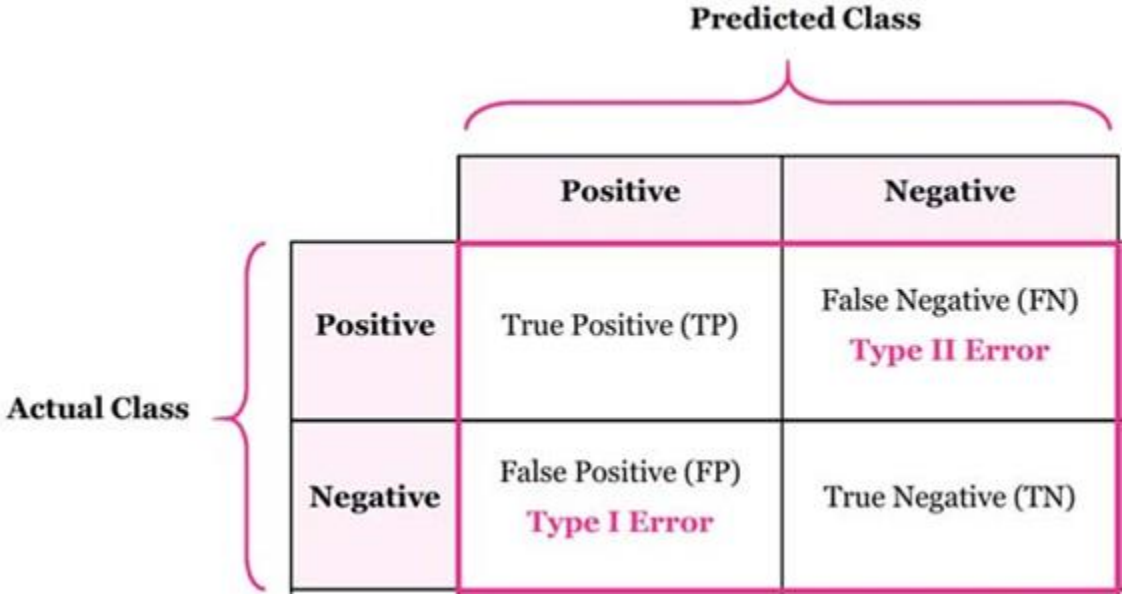
- در اکثر رده بند ها فرض بر این است که هزینه خطای رده بندی برای کلاس‌های متفاوت یکسان است. (آیا در دنیای واقعی این‌گونه است؟)
- اصل اولیّه در این رویکرد، نابرابری هزینه خطاهای رده بندی است.
- در این روش به‌جای محاسبه ساده خطا برای هر نمونه، هزینه رده بندی اشتباه برای هر کلاس، متفاوت در نظر گرفته می‌شود.
- مدل به جای تلاش برای بیشینه‌سازی صحت (Accuracy)، سعی در کمینه‌سازی کل هزینه‌های رده بندی اشتباه را دارد.

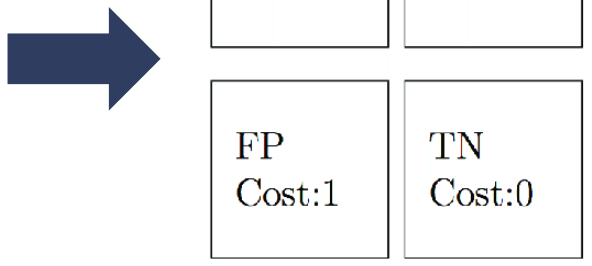
فرآیند داده کاوی

مدل های پیش بینانه – داده های نامتوازن

- رویکردهای مواجهه با داده های نامتوازن
  - رویکرد مبتنی بر تابع هزینه

با تعریف ماتریس هزینه، آموزش مدل بر اساس یادگیری حساس به هزینه (Cost-Sensitive Learning) خواهد بود.





انتخاب مقدار پارامتر هزینه:

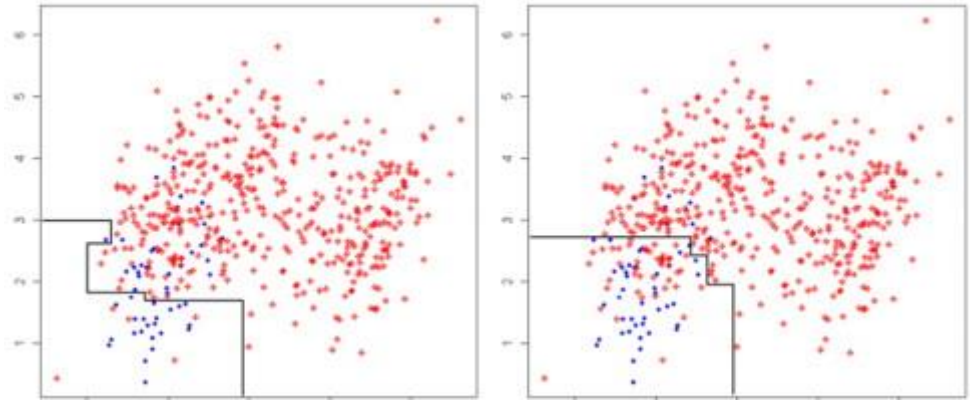
- بر اساس تحلیل اقتصادی و محاسبه هزینه اقتصادی خطا در پیش بینی
- بر اساس نسبت عدم توازن داده ها
- به روش آزمون و خطا

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

- ❑ رویکردهای مواجهه با داده های نامتوازن
  - رویکرد مبتنی بر معماری الگوریتم

الگوریتم های مختلف بر اساس ویژگی های ریاضیاتی، آماری و هندسی خود، می توانند نتایج متفاوتی در برخورد با داده های نامتوازن داشته باشند. بنابراین شناخت لازم و آگاهی از جزئیات الگوریتم ها می تواند در انتخاب الگوریتم های مناسب برای حل یک مسئله تعیین کننده باشد.

- بطور مثال الگوریتم هایی مانند درخت تصمیم به علت ماهیت جستجو و افزاری که دارند، عموما در مقابل چالش نامتوازن بودن داده ها مقاومت بیشتری خواهند داشت.



تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

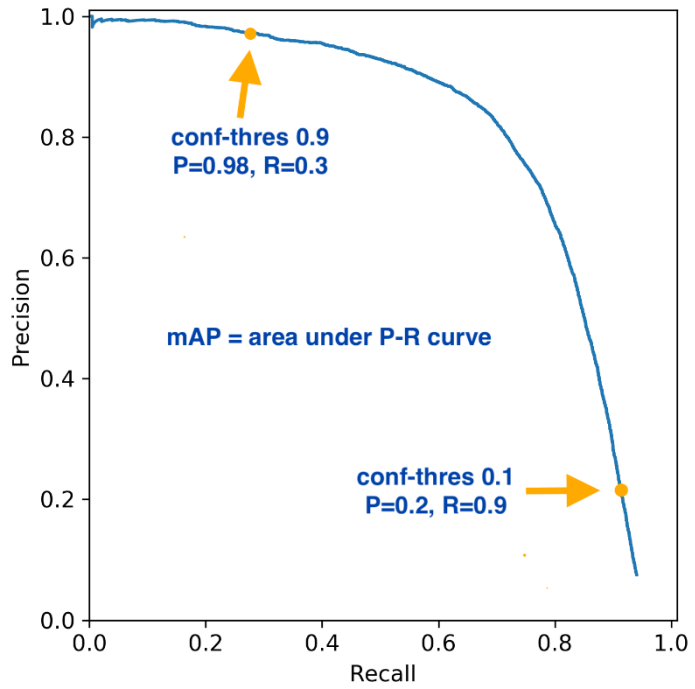
dayche.com | گروه دایکه

❑ رویکردهای مواجهه با داده های نامتوازن

○ رویکرد مبتنی بر معماری الگوریتم

تعیین حد آستانه مدل رده بندی (Classification Threshold) برای برچسب گذاری کلاس های پیش بینی شده، یکی از روش های مناسب در مواجهه با داده های نامتوازن است.

اغلب الگوریتم ها بطور پیش فرض در مسائل رده بندی (باینری) مقدار 0.5 را به عنوان حد آستانه ای در نظر می گیرند. در صورتیکه مقدار احتمال کلاس مثبت در مدل برازش داده شده A یعنی $P(+|A)$ بالای این حد باشد، برچسب کلاس مثبت و در غیر اینصورت برچسب کلاس منفی تخصیص داده می شود.



استفاده از نمودار Recall-Precision برای حدود آستانه ای متفاوت، ابزار رایجی در تعیین بهترین حد آستانه ای داده های نامتوازن می باشد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

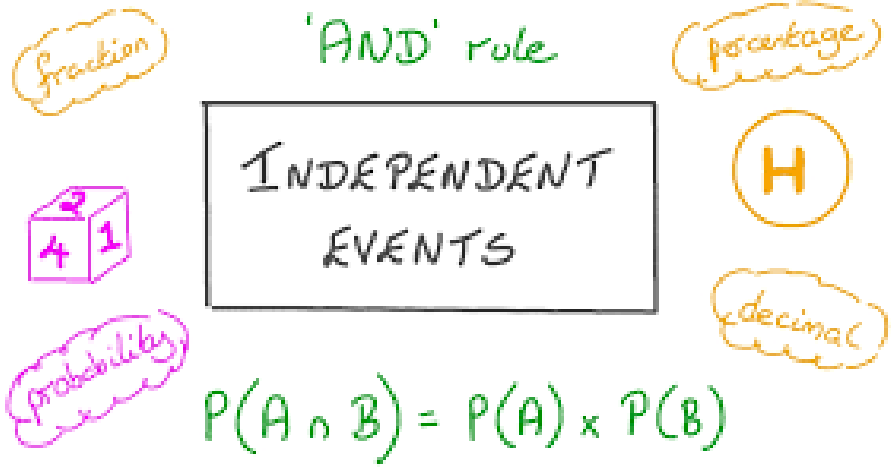
- رویکردهای مواجهه با داده های نامتوازن
  - رویکرد مبتنی بر معماری الگوریتم

یکی از روش‌های مناسب جهت مواجهه با داده‌های نامتعادل استفاده از قدرت چندین مدل رده بند به جای استفاده از یک مدل است. در این روش حل مسئله، به جای تمرکز بر ساخت یک مدل بسیار خوب به سمت ساخت یک سیستم خرد جمعی مطمئن می رویم.

این رویکرد تحت عنوان مدل های تجمیعی (Ensemble Models) شناخته می شوند.

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم بیز ساده



Conditional Probability Formula

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Probability of A given B

Probability of A and B

Probability of B

❑ مروری بر مفاهیم تئوری احتمال

○ پیشامد های مستقل و قانون ضرب احتمال

وقتی وقوع یک پیشامد، مستقل از وقوع پیشامد دیگری باشد، احتمال وقوع توام (همزمان) آنها برابر با حاصل ضرب احتمال هریک از پیشامدها می باشد.

○ احتمال شرطی

در بسیاری از مواقعی که بین وقوع دو پیشامد استقلال وجود نداشته باشد، می توان با آگاهی از نتیجه پیشامد اول، احتمال وقوع پیشامد دوم را در فضای نمونه کوچکتری جستجو کرد و مقدار احتمال را محاسبه نمود.

تولید محتوا: زهرا ذوالقدر

daychegroup

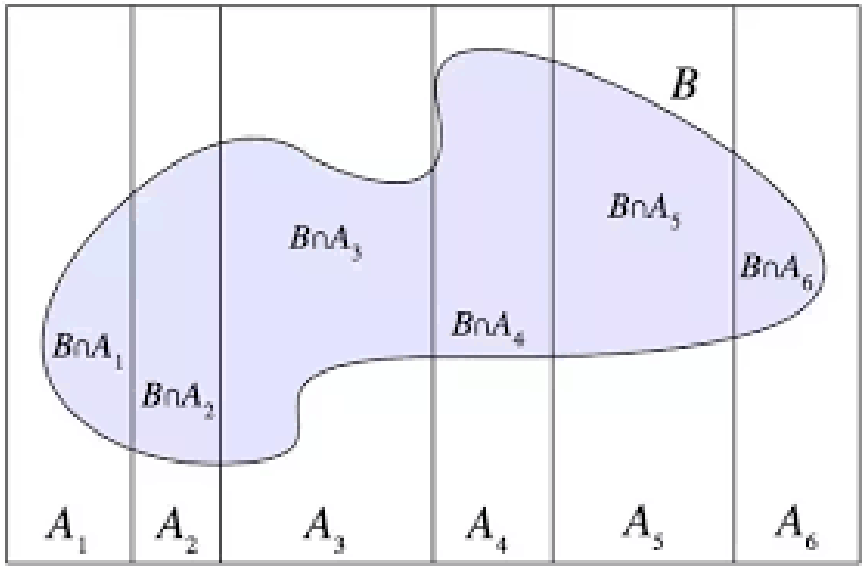
daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم بیز ساده

مروری بر مفاهیم تئوری احتمال



- **قانون احتمال کل**

احتمال یک پیشامد را می توان، به صورت حاصل جمع احتمال های شرطی آن در افراز پیشامد دیگری بازنویسی کرد.

$$P(B)=P(B \cap A_1)+ P(B \cap A_2)+ P(B \cap A_3)+ P(B \cap A_4)+ P(B \cap A_5) + P(B \cap A_6)$$

$$= P(B|A_1)\,P(A_1) + P(B|A_2)\,P(A_2)+ P(B|A_3)\,P(A_3)+ P(B|A_4)\,P(A_4)+ P(B|A_5)\,P(A_5)+ P(B|A_6)\,P(A_6)$$

$$=\sum_i P(B|Ai)\,P(Ai)$$

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم بیز ساده

- مروری بر مفاهیم تئوری احتمال
  - قانون بیز

محاسبه احتمال شرطی بدون نیاز به محاسبه احتمال توام. (محاسبه احتمال توام دو پیشامد در بسیاری از موارد به سادگی قابل محاسبه نیست)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

  - بسط قانون بیز بر اساس قانون احتمال کل

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)}.$$

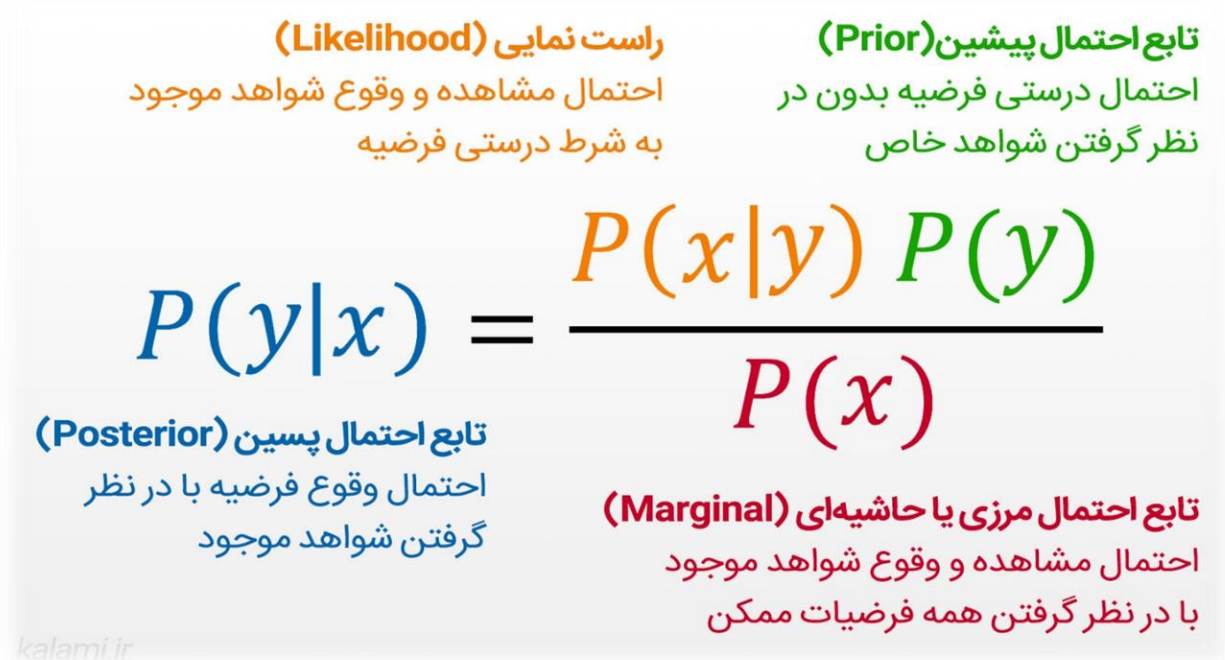
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم بیز ساده

❑ الگوریتم بیز ساده (Naïve Bayes)

این الگوریتم یک مدل رده بند احتمالاتی تولید کرده و بر اساس مشاهدات مربوط به داده های ورودی، مقداری از فیلد هدف را که دارای بیشترین احتمال وقوع باشد را بر می گرداند.



الگوریتم بیز ساده، یکی از انواع الگوریتم های یادگیری با نظارت از نوع رده بندی برای فیلد هدف کیفی می باشد.

فرض کنید بر اساس ویژگی ورودی x به دنبال محاسبه احتمال وقوع فیلد هدف y هستیم. بنابراین بر اساس قانون بیز رابطه روبرو را خواهیم داشت:

Posterior = Likelihood * Prior / Evidence

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم بیز ساده

❑ الگوریتم بیز ساده (Naïve Bayes)

استفاده از الگوریتم رده بندی بیز ساده در شرایط واقعی، شامل مشاهدات چند بعدی $x_i = x_1, x_2, \dots x_n$ می باشد. این الگوریتم، با در نظر گرفتن فرض قوی استقلال بین بردارهای ورودی و استفاده از قانون ضرب احتمال و قانون احتمال کل، محاسبه قانون بیز را به صورت زیر انجام می دهد:

$$\boldsymbol{P(Y|X)} = P(Y|x_1, x_2, \dots, x_n) = P(Y|x_1) * P(Y|x_2) * \cdots * P(Y|x_n) = \prod_{i=1}^{n} P(Y|x_i)$$

$$\boldsymbol{P(Y|X)} = \prod_{i=1}^{n} P(Y|x_i) = \prod_{i=1}^{n} \frac{P(x_i|Y)P(Y)}{P(x_i)} = \prod_{i=1}^{n} \frac{\boldsymbol{P(x_i|Y)P(Y)}}{\boldsymbol{P(x_i|Y)P(Y)P(Y) + P(x_i|\sim Y)P(\sim Y)}}$$

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ الگوریتم بیز ساده (Naïve Bayes)

- **محاسبه بیشینه احتمال پسین:** Maximum a Posterior (MAP)

پیش بینی برچسب فیلد هدف با محاسبه احتمال وقوع هر یک از کلاس ها، به شرط مقادیر ویژگی های ورودی و تعیین کلاسی که مقدار احتمال را بیشینه کند (MAP) انجام می پذیرد.

فرض کنید فیلد هدف به تعداد K کلاس دارد؛ در اینصورت پیش بینی برچسب کلاس فیلد هدف $\hat{y}$ بر اساس رابطه زیر بدست می آید:

$$\hat{y} = \arg\ max\ \prod_{i=1}^{n} \frac{P(x_i|Y=c_k)P(Y=c_k)}{P(x_i)}$$

$$k \in \{1,2,\dots,K\}$$

مقداری از کلاس فیلد هدف c_k به عنوان رده پیش بینی شده انتخاب می گردد، که مقدار تابع رده بند بیز را بیشینه کند.

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم بیز ساده

ویژگی های الگوریتم بیز ساده (Naïve Bayes)

به علت فرض استقلال بین ویژگی های ورودی، محاسبه رابطه بیز با جداسازی توزیع احتمال شرطی برای هر ویژگی بصورت مستقل از سایر ویژگی ها انجام می شود و این موضوع دو نتیجه مهم در پی خواهد داشت:

- نسبت به بسیاری از الگوریتم های دیگر رده بندی، نیاز به داده های کمتری برای محاسبات خود دارد و از این رو یکی از کاندیداهای مناسب برای مواقعی که با داده های کم مواجه هستیم می باشد.
- هرچند با ساده سازی روابط بین ویژگی های ورودی، مقدار احتمال پسین بدست آمده اغلب دارای خطا و انحراف از مقدار واقعی می باشد، اما با توجه به نحوه تخصیص برچسب کلاس هدف بر اساس MAP، فارغ از اینکه برآورد احتمال کم یا زیاد دارای خطا باشد، اما برچسب پیش بینی شده به درستی گزینه محتمل تری نسبت به سایر مقادیر است.

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم بیز ساده

ویژگی های الگوریتم بیز ساده (Naïve Bayes)

به علت فرض استقلال بین ویژگی های ورودی، محاسبه رابطه بیز با جداسازی توزیع احتمال شرطی برای هر ویژگی بصورت مستقل از سایر ویژگی ها انجام می شود و این موضوع دو نتیجه مهم در پی خواهد داشت:

- به علت محاسبات بسیار سریع، در مسائلی که نیاز به پیش بینی بلادرنگ (Real Time) و یا در شرایط وجود داده های با حجم بسیار زیاد، گزینه مطلوبی هست.

به علت پشتیبانی کامل از فیلد هدف چند کلاسه و محاسبه احتمال پسین برای هر کلاس (روش MAP)، در اغلب مسائل رده بندی مانند رده بندی متون (Text Classification)، تشخیص اسپم (Spam Filtering)، تحلیل احساسات در شبکه های اجتماعی (Sentiment Analysis) و سیستم های پیشنهاد دهنده (Recommendation System) بسیار مورد توجه می باشد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم بیز ساده

spam: "send us your password"
ham: "send us your review"
ham: "password review"
spam: "review us "
spam: "send your password"
spam: "send us your account"

new email "review us now"

- مثال عددی از الگوریتم بیز ساده (Naïve Bayes)
  - استفاده از الگوریتم بیز ساده در تشخیص اسپم.

فرض کنید داده های آموزشی در دسترس شامل 6 متن ایمیل مطابق جدول روبرو است که با برچسب های "ham" و "spam" کلاس آنها مشخص شده است.

می خواهیم با استفاده از الگوریتم بیز ساده، برچسب ایمیل جدید با متن "review us now" را تشخیص دهیم.

Send	Us	Your	Password	Review	Account	Class
1	1	1	1	0	0	Spam
1	1	1	0	1	0	Ham
0	0	0	1	1	0	Ham
0	1	0	0	1	0	Spam
1	0	1	1	0	0	Spam
1	1	1	0	0	1	Spam

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
گروه دایکه | dayche.com

- مثال عددی از الگوریتم بیز ساده (Naïve Bayes)
  - استفاده از الگوریتم بیز ساده در تشخیص اسپم.

بر اساس روابط بیز، احتمال پیشین کلاس های هدف و همچنین احتمال پسین به شرط وقوع یک ویژگی با کمک داده های آموزشی بدست می آید:

spam: "send us your password"
ham: "send us your review"
ham: "password review"
spam: "review us "
spam: "send your password"
spam: "send us your account"

new email "review us now"

Prior probabilities are:

$$\Pr(\text{spam}) = \frac{4}{6} \qquad \Pr(\text{ham}) = \frac{2}{6}$$

The posterior probability that an email containing the word "review" is a spam is:

$$\Pr(\text{spam} \mid \text{review}) = \frac{\Pr(\text{review}|\text{spam})\Pr(\text{spam})}{\Pr(\text{review}|\text{spam})\Pr(\text{spam}) + \Pr(\text{review}|\text{ham})\Pr(\text{ham})} = \frac{\frac{1}{4}\cdot\frac{4}{6}}{\frac{1}{4}\cdot\frac{4}{6} + \frac{2}{2}\cdot\frac{2}{6}} = \frac{1}{3}$$

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم بیز ساده

spam: "send us your password"
ham: "send us your review"
ham: "password review"
spam: "review us "
spam: "send your password"
spam: "send us your account"

new email "review us now"

- مثال عددی از الگوریتم بیز ساده (Naïve Bayes)
  - استفاده از الگوریتم بیز ساده در تشخیص اسپم.

برای استفاده از رده بند بیز ساده، بایستی محاسبات قانون بیز بصورت جداگانه برای هر ویژگی و همچنین به تفکیک کلاس های فیلد هدف انجام شود. در جدول زیر محاسبه مقدار درستنمایی تمام ویژگی های مورد بررسی به تفکیک دو کلاس نشان داده شده است:

	$\Pr(\cdot \mid \text{spam})$	$\Pr(\cdot \mid \text{ham})$
review	1/4	2/2
send	3/4	1/2
us	3/4	1/2
your	3/4	1/2
password	2/4	1/2
account	1/4	0/2

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم بیز ساده

- مثال عددی از الگوریتم بیز ساده (Naïve Bayes)
  - استفاده از الگوریتم بیز ساده در تشخیص اسپم.

spam: "send us your password"
ham: "send us your review"
ham: "password review"
spam: "review us "
spam: "send your password"
spam: "send us your account"

new email "review us now"

	Pr(· | spam)	Pr(· | ham)
review	1/4	2/2
send	3/4	1/2
us	3/4	1/2
your	3/4	1/2
password	2/4	1/2
account	1/4	0/2

محاسبه مقدار درستنمایی برای عبارت ایمیل جدید به تفکیک کلاس های هدف انجام می شود:

$$\Pr(\text{review us now} \mid \text{spam}) = \Pr(\{1,0,1,0,0,0\} \mid \text{spam}) = \frac{1}{4}\left(1-\frac{3}{4}\right)\frac{3}{4}\left(1-\frac{3}{4}\right)\left(1-\frac{2}{4}\right)\left(1-\frac{1}{4}\right) = 0.0044$$

$$\Pr(\text{review us now} \mid \text{ham}) = \Pr(\{1,0,1,0,0,0\} \mid \text{ham}) = \frac{2}{2}\left(1-\frac{1}{2}\right)\frac{1}{2}\left(1-\frac{1}{2}\right)\left(1-\frac{1}{2}\right)\left(1-\frac{0}{2}\right) = 0.0625$$

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

spam: "send us your password"
ham: "send us your review"
ham: "password review"
spam: "review us "
spam: "send your password"
spam: "send us your account"

new email "review us now"

- مثال عددی از الگوریتم بیز ساده (Naïve Bayes)
  - استفاده از الگوریتم بیز ساده در تشخیص اسپم.

در نهایت محاسبه احتمال پسین برای عبارت ایمیل جدید محاسبه می شود و با توجه به اینکه مقدار آن برای کلاس spam کوچکتر می باشد، در نتیجه با احتمال 0.877 ایمیل جدید متعلق به کلاس ham می باشد.

Then, the posterior probability that the new email "review us now" is a spam is:

$$\Pr(\text{spam} \mid \text{review us now}) = \Pr(\text{spam} \mid \{1,0,1,0,0,0\})$$

$$= \frac{\Pr(\{1,0,1,0,0,0\}|\text{spam})\Pr(\text{spam})}{\Pr(\{1,0,1,0,0,0\}|\text{spam})\Pr(\text{spam}) + \Pr(\{1,0,1,0,0,0\}|\text{ham})\Pr(\text{ham})}$$

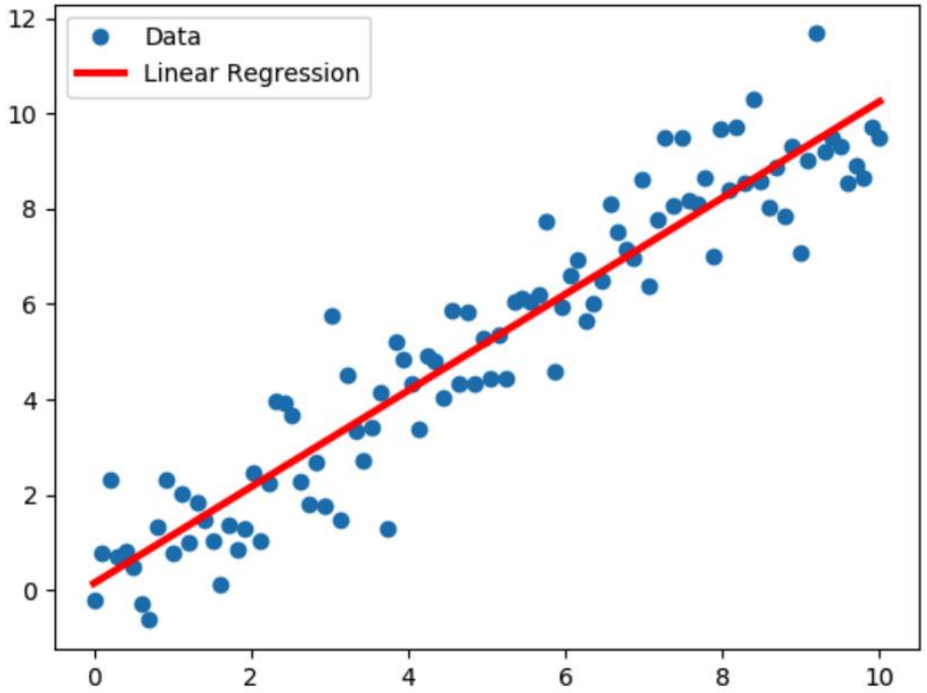
$$= \frac{0.0044 \cdot \frac{4}{6}}{0.0044 \cdot \frac{4}{6} + 0.0625 \cdot \frac{2}{6}} = 0.123$$

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی

❑ مقدمه ای بر رگرسیون خطی

یکی از پرکاربرد ترین الگوریتم های یادگیری آماری با هدف بررسی و مدلسازی ارتباط خطی بین یک یا چند ویژگی ورودی مستقل از هم با متغیر وابسته (پاسخ) می باشد. رگرسیون خطی از نوع یادگیری با نظارت با هدف پیش بینی مقادیر کمی است.



فقط یک ویژگی ورودی

رگرسیون خطی ساده

Simple Linear Regression

بیش از یک ویژگی ورودی

رگرسیون خطی چندگانه

Multiple Linear Regression

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

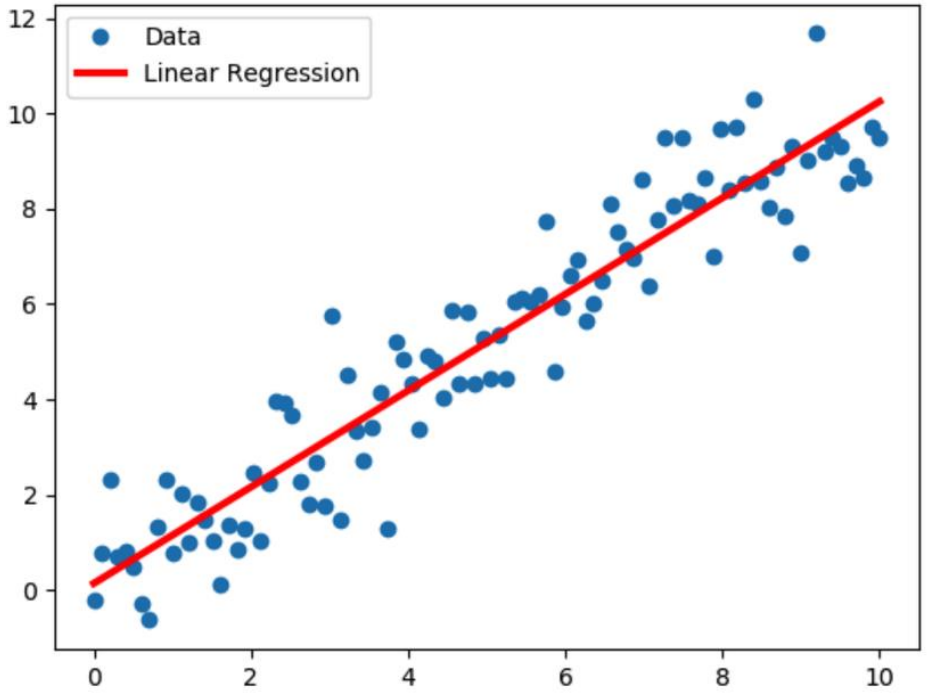
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی

❑ مقدمه ای بر رگرسیون خطی

یکی از پرکاربرد ترین الگوریتم های یادگیری آماری با هدف بررسی و مدلسازی ارتباط خطی بین یک یا چند ویژگی ورودی مستقل از هم با متغیر وابسته (پاسخ) می باشد. رگرسیون خطی از نوع یادگیری با نظارت با هدف پیش بینی مقادیر کمی است.



فقط یک ویژگی ورودی

رگرسیون خطی ساده

Simple Linear Regression

Regression Analysis Formula

$$Y = mx + b$$

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی

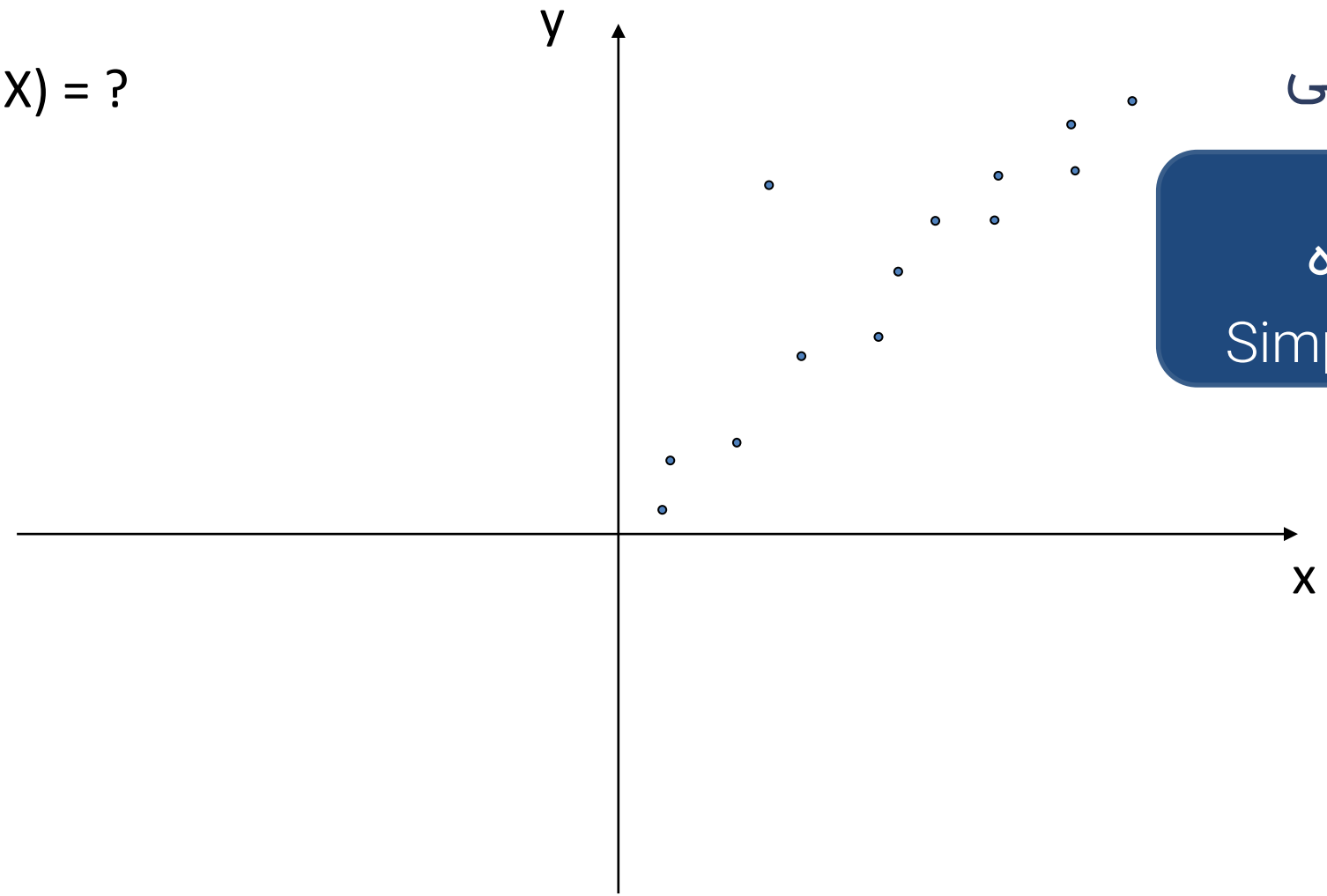
- مقدمه ای بر رگرسیون خطی

فقط یک ویژگی ورودی

رگرسیون خطی ساده

Simple Linear Regression

$E(Y|X) = ?$



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی

- مقدمه ای بر رگرسیون خطی

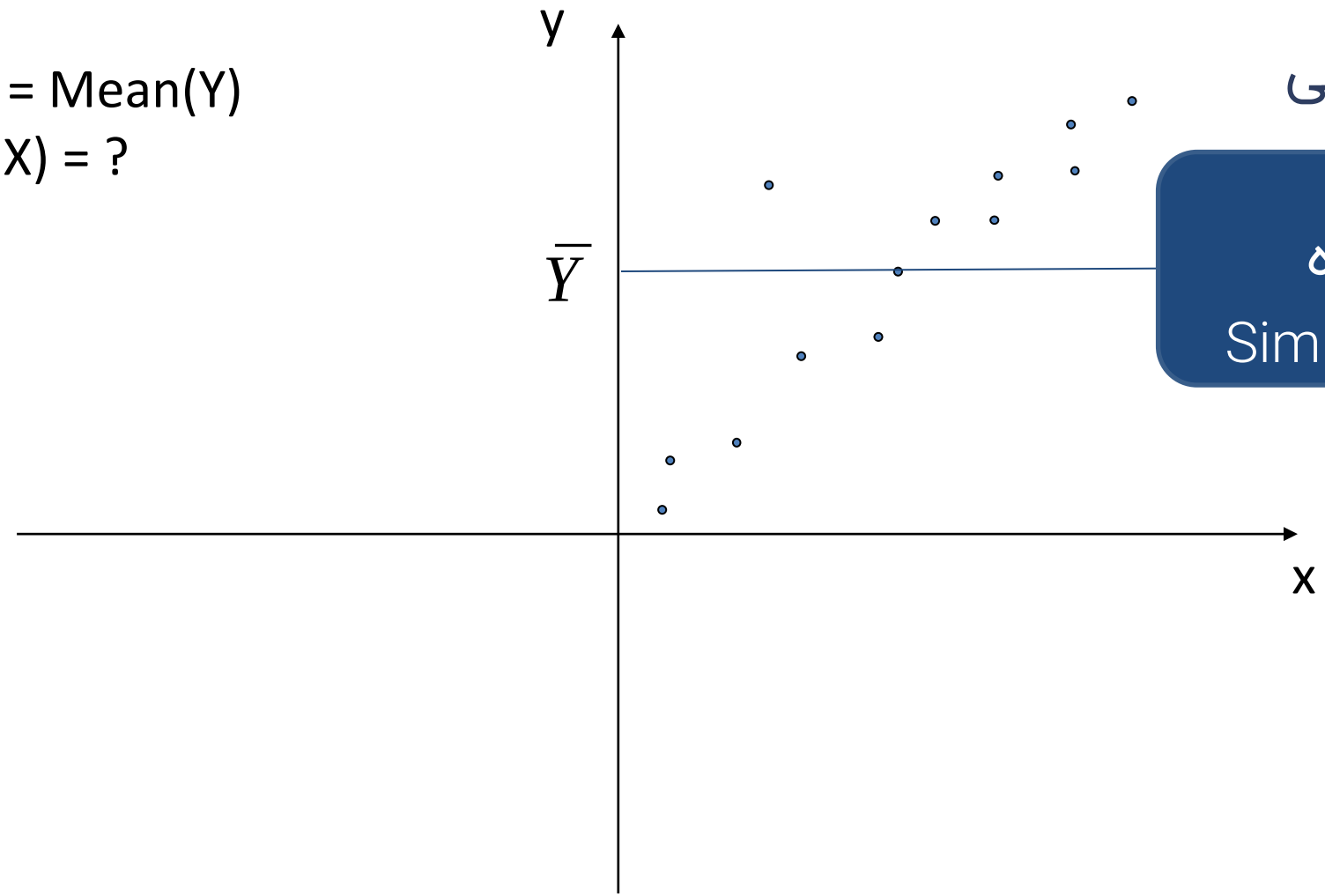
فقط یک ویژگی ورودی

رگرسیون خطی ساده

Simple Linear Regression

E(Y) = Mean(Y)

E(Y|X) = ?



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

- مقدمه ای بر رگرسیون خطی

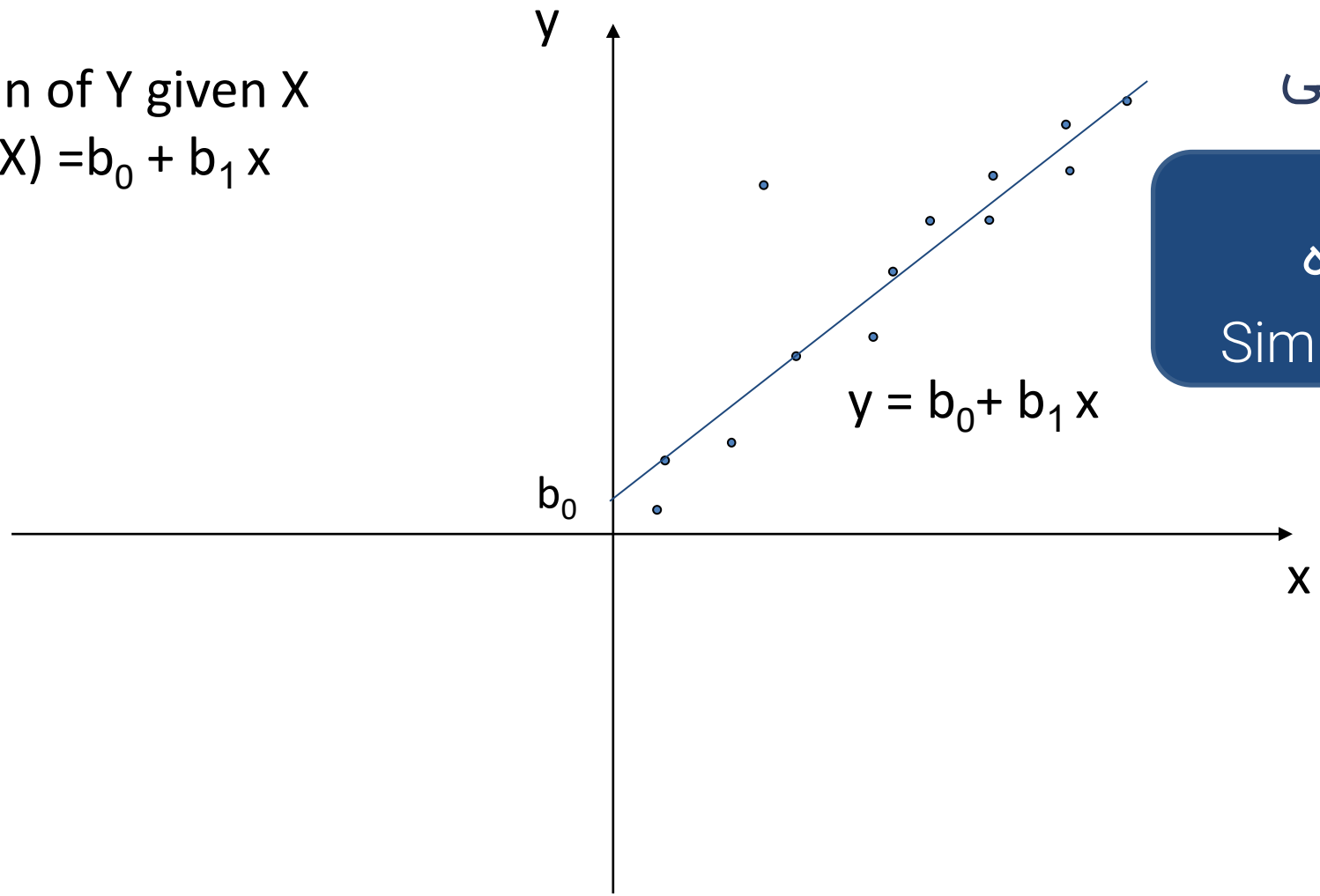
فقط یک ویژگی ورودی

رگرسیون خطی ساده

Simple Linear Regression

Mean of Y given X

$E(Y|X) = b_0 + b_1 x$



تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی

- مقدمه ای بر رگرسیون خطی

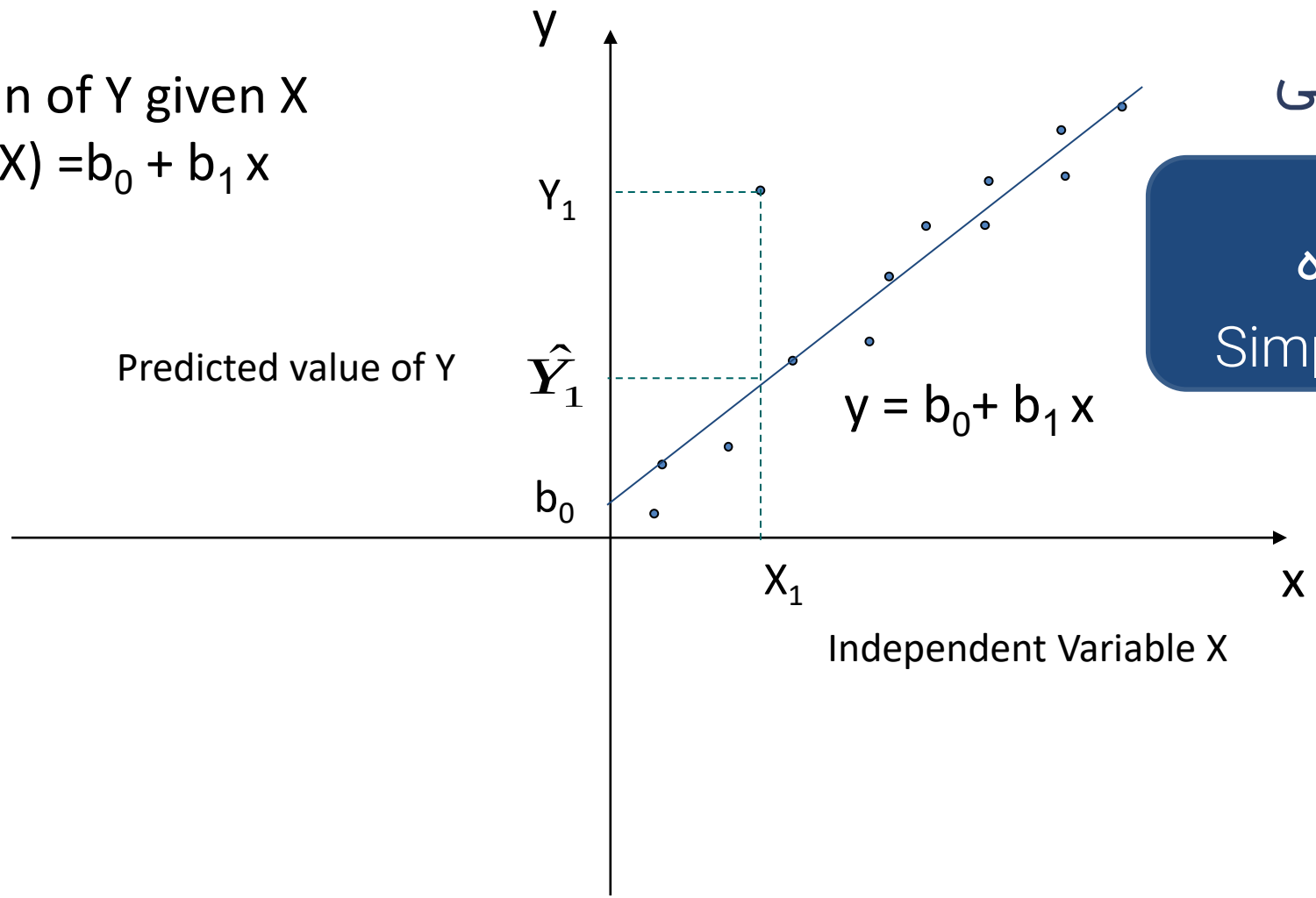
فقط یک ویژگی ورودی

رگرسیون خطی ساده

Simple Linear Regression

Mean of Y given X

$E(Y|X) = b_0 + b_1 x$



تولید محتوا: زهرا ذوالقدر

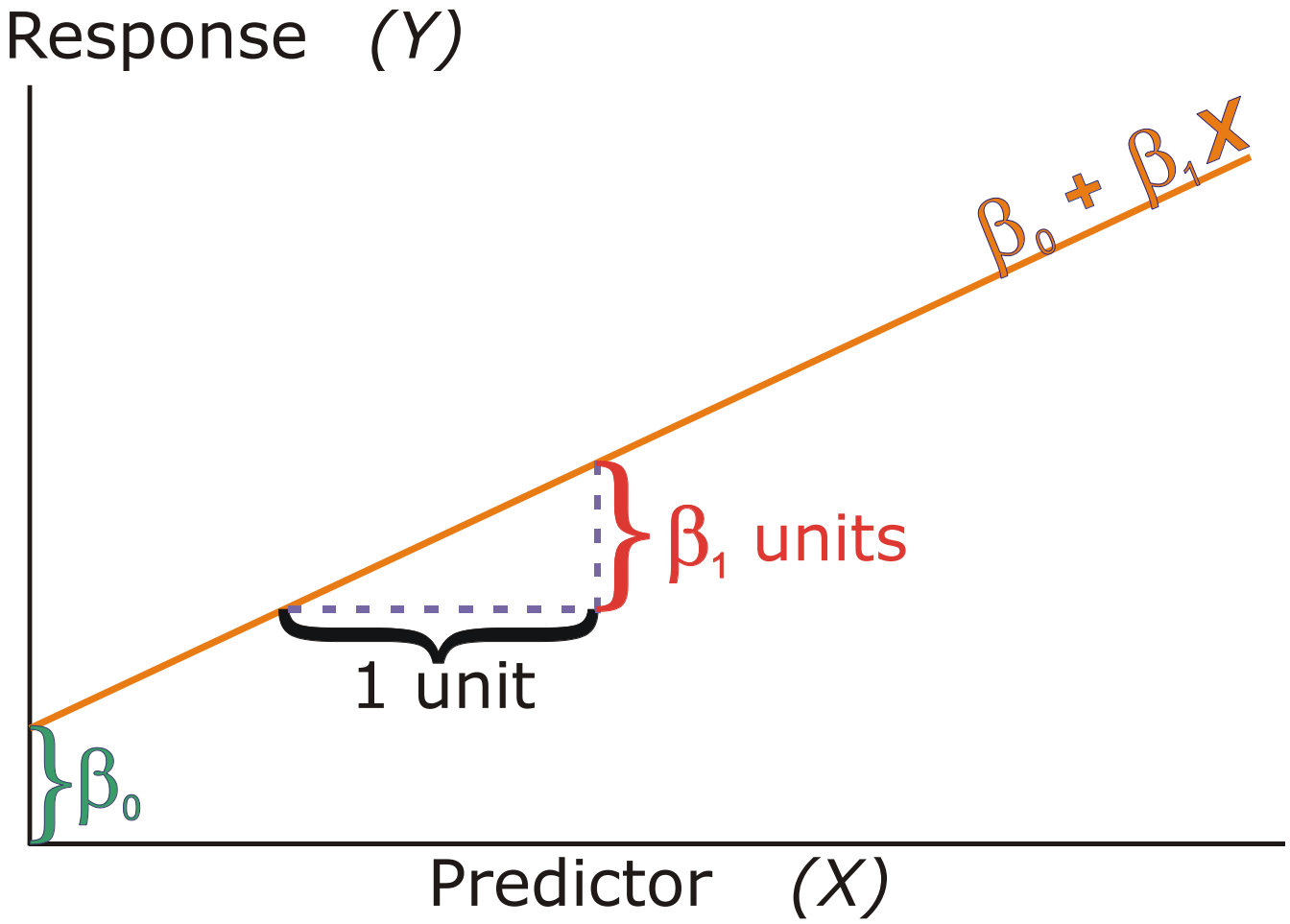
daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی



- مقدمه ای بر رگرسیون خطی

فقط یک ویژگی ورودی

رگرسیون خطی ساده

Simple Linear Regression

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

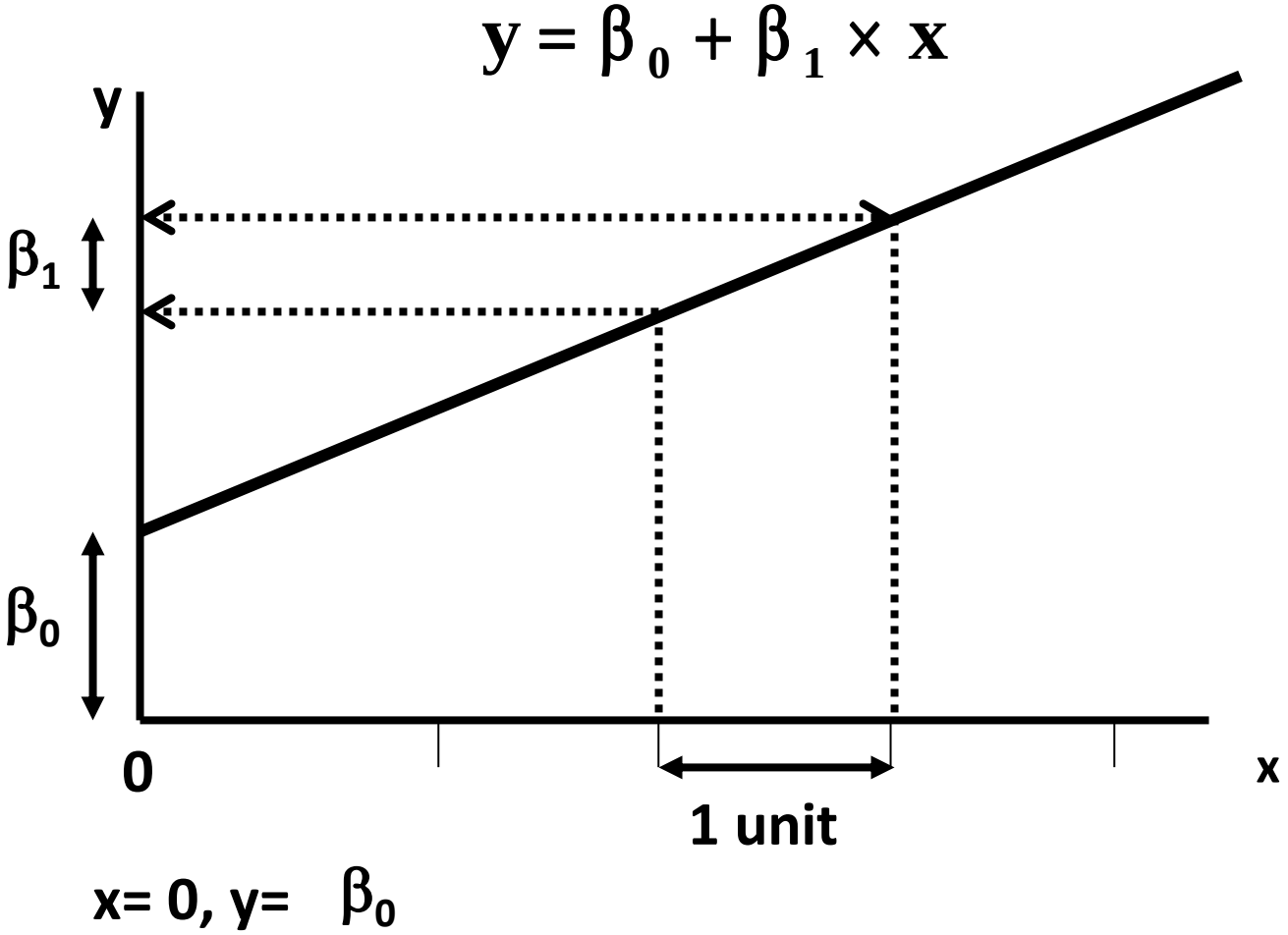
dayche.com | گروه دایکه

- مقدمه ای بر رگرسیون خطی

فقط یک ویژگی ورودی

رگرسیون خطی ساده

Simple Linear Regression



تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

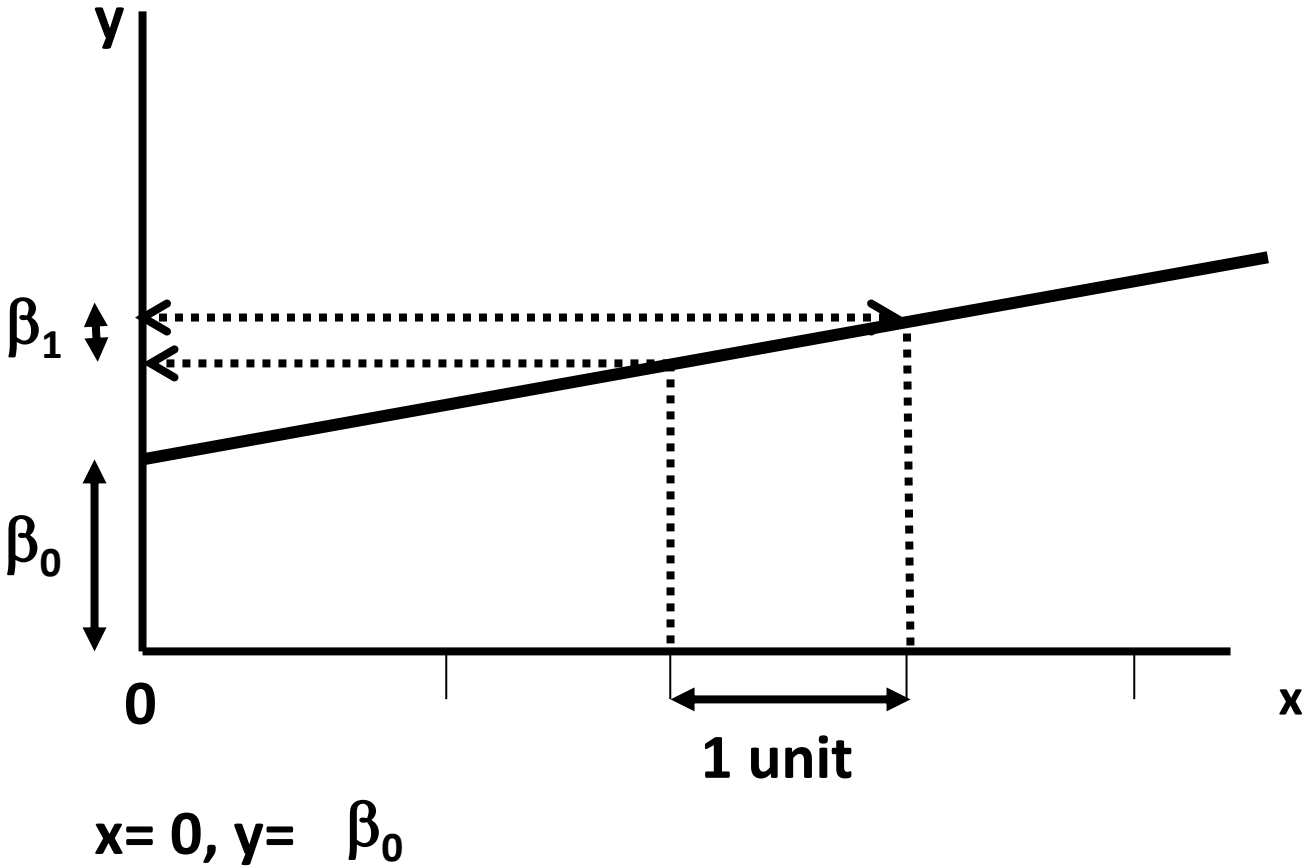
- مقدمه ای بر رگرسیون خطی

فقط یک ویژگی ورودی

رگرسیون خطی ساده

Simple Linear Regression

$$y = \beta_0 + \beta_1 \times x$$



تولید محتوا: زهرا ذوالقدر

daychegroup

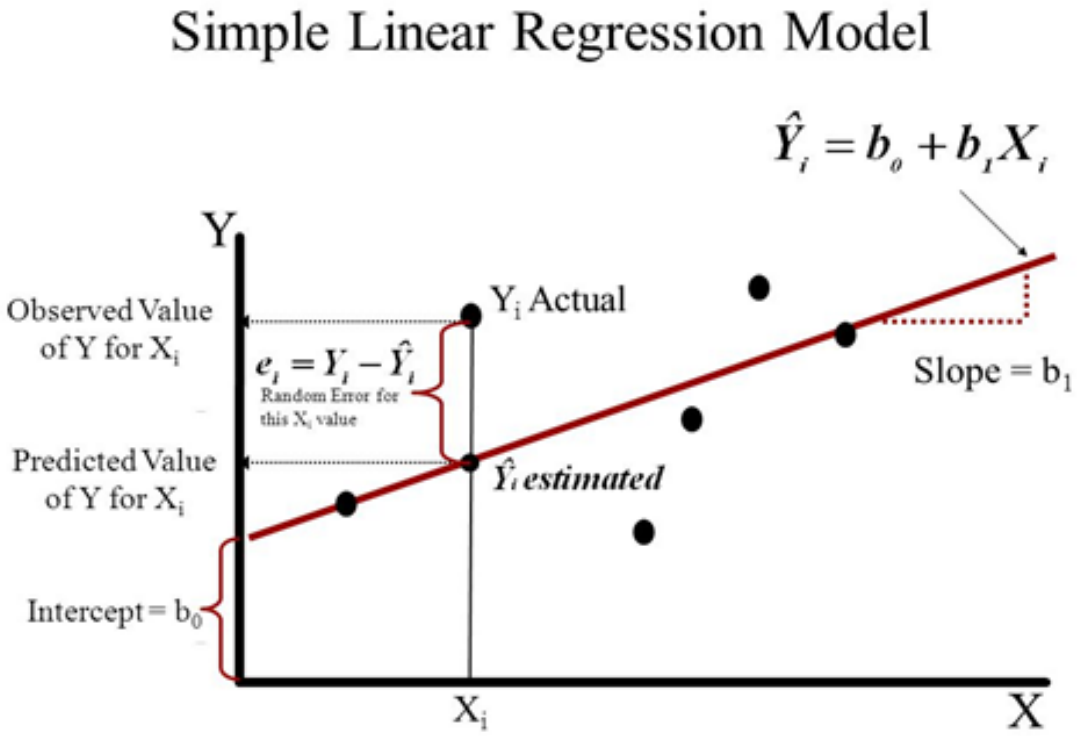
daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی

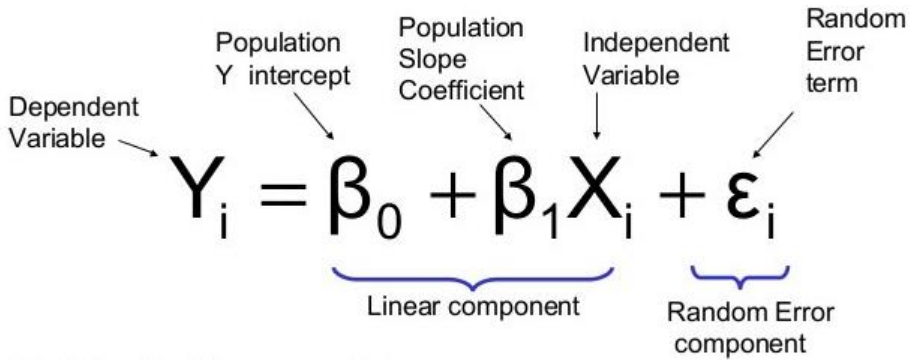
❑ مقدمه ای بر رگرسیون خطی



فقط یک ویژگی ورودی

رگرسیون خطی ساده

Simple Linear Regression



تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

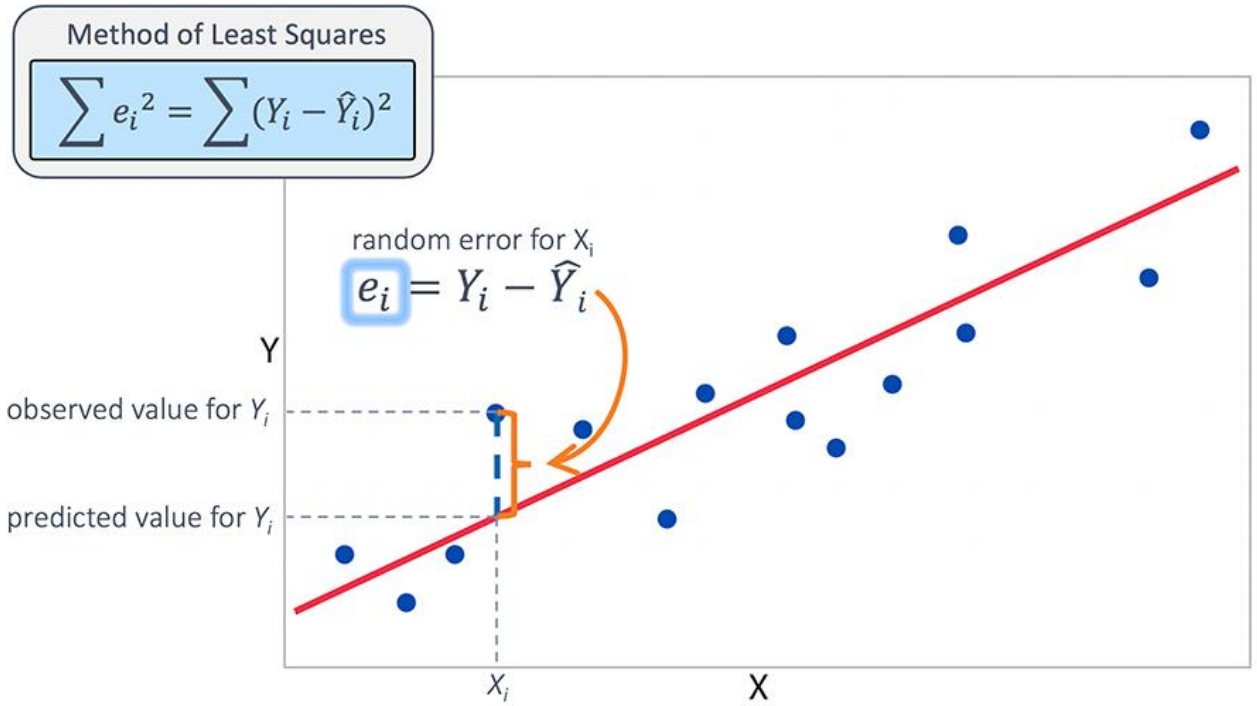
مدل های پیش بینانه – الگوریتم رگرسیون خطی

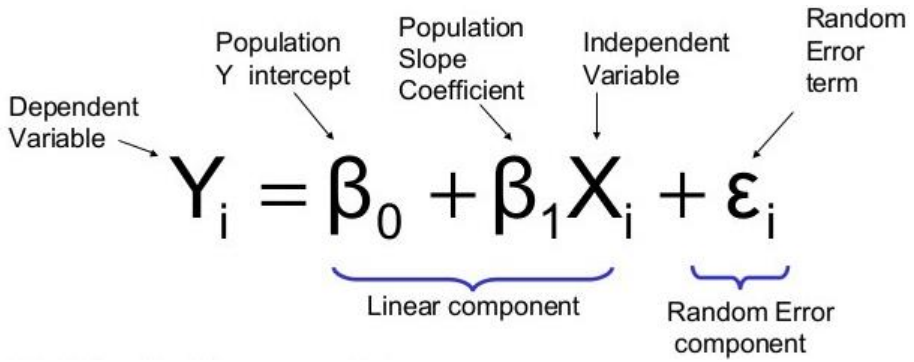
❑ مقدمه ای بر رگرسیون خطی

فقط یک ویژگی ورودی

رگرسیون خطی ساده

Simple Linear Regression





هدف مدل رگرسیون، یافتن بهترین مقدار برای پارامترهای مدل به منظور به حداقل رساندن مجموع مجذور خطاها می باشد.

تولید محتوا: زهرا ذوالقدر

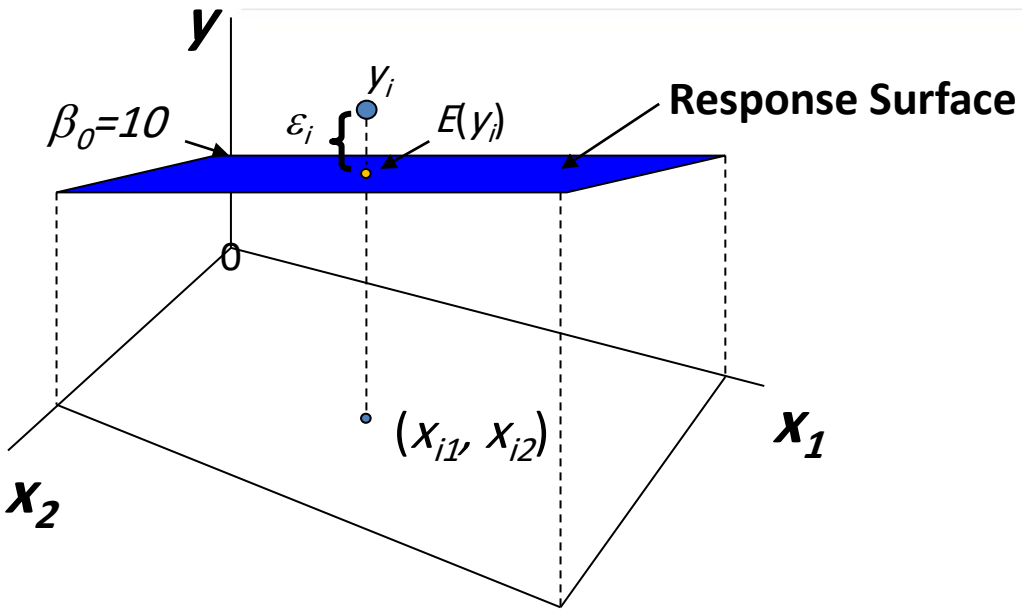
daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی



- مقدمه ای بر رگرسیون خطی

بیش از یک ویژگی ورودی

رگرسیون خطی چندگانه

Multiple Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

intercept — slopes — Random error

Dependent (response) variable

Independent (explanatory) variables

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

- مقدمه ای بر رگرسیون خطی

بیش از یک ویژگی ورودی

رگرسیون خطی چندگانه

Multiple Linear Regression

$$y = X\beta + \varepsilon$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی

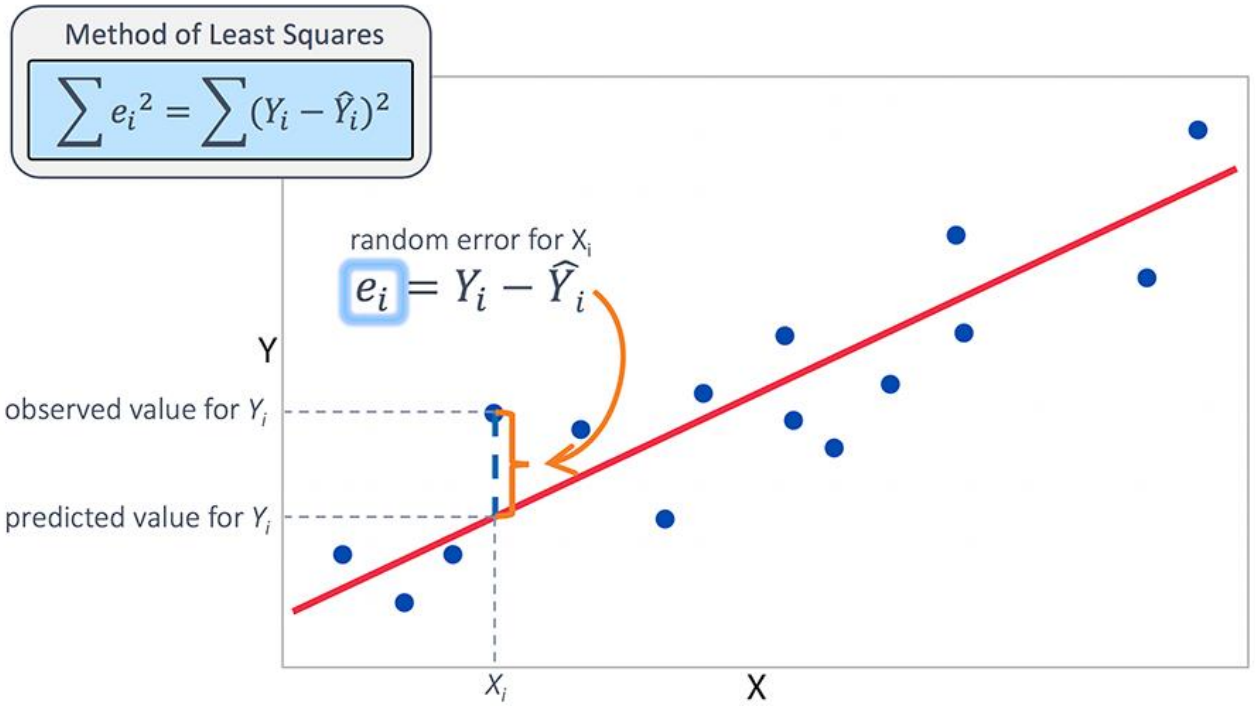
- مقدمه ای بر رگرسیون خطی

بیش از یک ویژگی ورودی

رگرسیون خطی چندگانه

Multiple Linear Regression

$$y = X\beta + \varepsilon$$



هدف مدل رگرسیون، یافتن بهترین مقدار برای پارامترهای مدل به منظور به حداقل رساندن مجموع مجذور خطاها می باشد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ برآورد پارامترهای مدل

- **روش حداقل مربعات معمولی (Ordinary Least Square Method - OLS)**

این روش به عنوان رایج ترین و پرکاربردترین رویکرد در برآورد پارامترهای مدل رگرسیون مورد استفاده قرار می گیرد. برای محاسبه ضرایب مدل، با استفاده از محاسبات ماتریسی در جبر خطی بهترین مقادیر برای پارامترهای مجهول مدل به شکلی یافت می شود که مجموع مربعات خطای مدل (اختلاف بین مقدار واقعی و مقدار خط رگرسیونی) کمینه شود.

$$\hat{\beta} = \arg\min_{\beta} S(\beta)$$

Where $S(\beta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \sum_{j=0}^{k} X_{ij}\beta_j)^2 = ||y - X\beta||^2$

در حجم داده های زیاد، محاسبات ماتریسی لازم در این روش نیاز به دسترسی به داده های کامل و حافظه کافی (Memmory) می باشد.

با مشتق گیری از تابع هدف نسبت به پارامترهای بتا، برآورد پارامترها انجام می پذیرد.

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

تولید محتوا: زهرا ذوالقدر

daychegroup

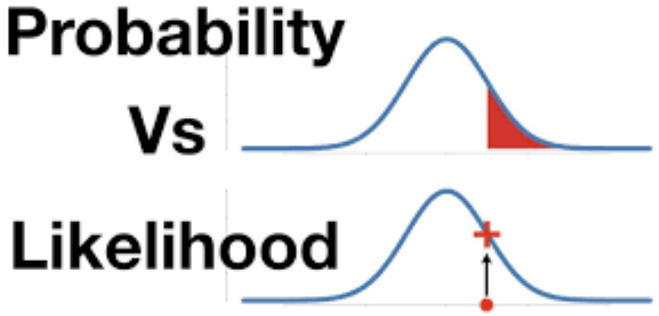
daychegroup

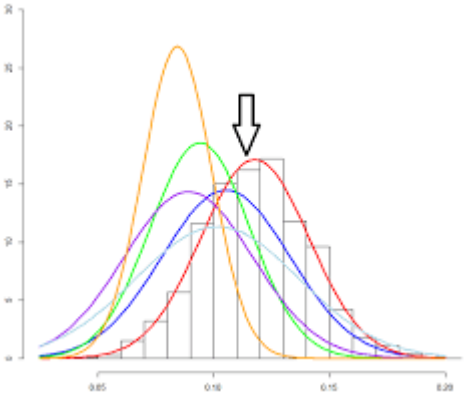
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی

❑ برآورد پارامترهای مدل





- روش تخمین بیشینه درستنمایی (Maximum Likelihood Estimation - MLE)

یکی از بهترین و مطمئن ترین روشهای برآورد در آمار، روش مبتنی بر تابع درستنمایی هست.

تابع درستنمایی: احتمال مشاهده نمونه های مشخصی از داده ها را، به شرط درستی یک فرضیه (مدل یا پارامترهای توزیع) درستنمایی می گویند. در صورتیکه نمونه ها را ثابت در نظر بگیریم با تغییر مقدار پارامتر در فضای مربوط به آن، تابع درستنمایی را خواهیم داشت.

$$\mathrm{L}(\theta|x) = \mathrm{P}(\mathrm{X} = \mathrm{x}|\theta) \xrightarrow{x_i \sim i.i.d} \mathrm{L}(\theta|x_i) = \prod_{i=1}^{n} P(x_i|\theta)$$

تفاوت تابع احتمال و تابع درستنمایی در نوع نگاه آنها به مسئله هست.

در تابع احتمال با ثابت نگه داشتن مقدار پارامتر، توزیع مقدار احتمال داده ها مورد بررسی قرار می گیرد.

در صورتیکه در تابع درستنمایی با ثابت نگه داشتن مقادیر داده، مقدار احتمال درستی پارامتر بررسی می شود.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ برآورد پارامترهای مدل

- روش تخمین بیشینه درستنمایی (Maximum Likelihood Estimation - MLE)

در برآورد ضرایب مدل رگرسیونی (h)، روش تخمین بیشینه درستنمایی (MLE) با در نظر گرفتن داده های ورودی X به مدل، به دنبال یافتن بهترین پارامترهای مدل (ضرایب بتا) است بطوری که بیشترین همخوانی را با داده های در نظر گرفته شده داشته باشد.

$$Maximize \sum_{i=1}^{n} \log(L(h|x_i)) = Minimize \sum_{i=1}^{n} -\log(P(x_i|h))$$

به علت کوچک بودن مقدار احتمال برای هر مقدار x معمولا از لگاریتم تابع درستنمایی در حل مسئله استفاده می شود و به همین دلیل تابع هدف این روش به عنوان Log – Likelihood نیز شناخته می شود.

ثابت می شود که در صورت برقراری فرض نرمال بودن توزیع خطا در مدل رگرسیون، برآورد حداقل مربعات یک برآورد MLE نیز محسوب می گردد.

فرآیند داده کاوی

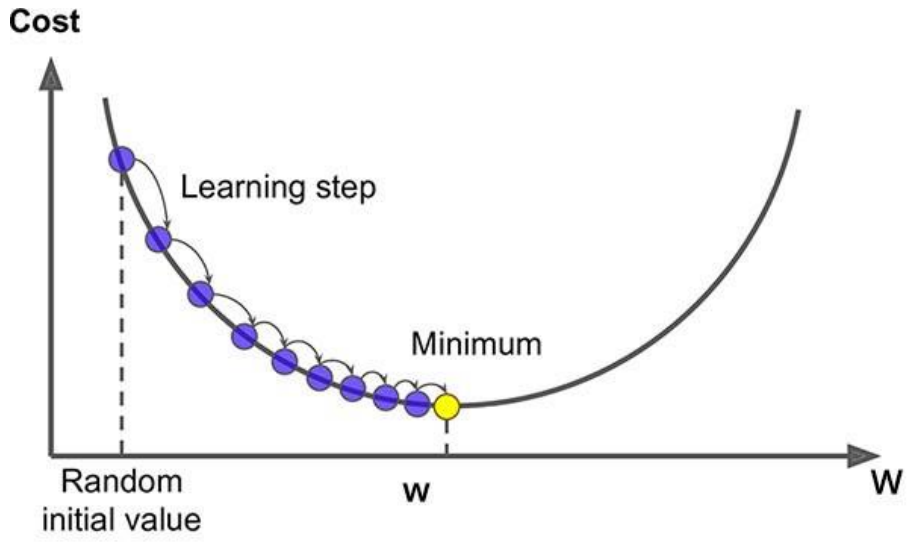
مدل های پیش بینانه – الگوریتم رگرسیون خطی

❑ برآورد پارامترهای مدل

- روش کاهش گرادیان (Gradient Decent Method)

یکی دیگر از روش های برآورد ضرایب مدل رگرسیونی استفاده از روش کاهش گرادیان می باشد که با در نظر گرفتن مقدار تصادفی اولیه برای هریک از ضرایب مدل و محاسبه میزان خطای پیش بینی در جهت کاهش خطا حرکت کرده و طی یک فرآیند تکراری به ضرایب بهینه همگرا می شود.

این روش در مواقعی که با حجم داده های زیادی در تعداد مشاهدات و ویژگی ها مواجه باشیم (نیاز به حافظه زیادی برای محاسبات ماتریسی خواهد داشت) گزینه مطلوبی خواهد بود.



برآورد ضرایب مدل در الگوریتم شبکه عصبی عموما از روش کاهش گرادیان استفاده می نماید.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ برآورد پارامترهای مدل

- روش تنظیم سازی (Regularization Method)

برخی روش های توسعه یافته روی مدل رگرسیون حداقل مربعات، یافتن ضرایب مدل را به گونه ای انجام می دهند که علاوه بر کاهش خطای پیش بینی بطور همزمان کاهش پیچیدگی مدل را نیز لحاظ می نمایند. این روش ها که تحت عنوان روش تنظیم سازی شناخته می شوند، با اضافه کردن مجموع نرم اول یا دوم ضرایب به تابع بهینه سازی، عملیات برآورد ضرایب را انجام می دهند.

$$\hat{\beta} = \arg\min \sum_{i=1}^{n} (y_i - \sum_{j=0}^{k} X_{ij}\beta_j)^2 + \sum_{j=0}^{k} \lambda |\beta_j|^q$$

مقدار λ در این رابطه به عنوان پارامتر تنظیم می باشد و مقادیر بزرگتر آن منجر به سادگی بیشتر مدل می شود؛ به این معنی که تعداد ویژگی های با مقدار بتای معنادار کمتر شده و مدل ساده تری خواهیم داشت. در نتیجه تنظیم سازی روشی برای جلوگیری از بیش برازشی می باشد.

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی

- برآورد پارامترهای مدل
  - روش تنظیم سازی (Regularization Method)
    - رگرسیون لاسو (Lasso Regression)

در صورتیکه مقدار q = 1 باشد، به عنوان L1-Regularization شناخته شده و ویژگی مهم این روش در این می باشد که منجر به صفر شدن مقدار بتا برای تعداد قابل توجهی از ویژگی های کم اهمیت می شود. به همین دلیل مدل بدست آمده با این روش تحت عنوان مدل تنک (Sparse) روشی مناسب برای انتخاب ویژگی در داده های با ابعاد بالا شناخته می شود.

    - رگرسیون ستیغی (Ridge Regression)

در صورتیکه مقدار q = 2 باشد، به عنوان L2-Regularization شناخته شده و منجر به کوچک شدن میزان بتا برای ویژگی های کم اهمیت می شود (نزدیک به صفر) و بعلت فرم بسته توان دوم آن نسبت به حالت L1 پیچیدگی محاسباتی کمتری دارد.

❑ آزمون های فرض مدل رگرسیون خطی

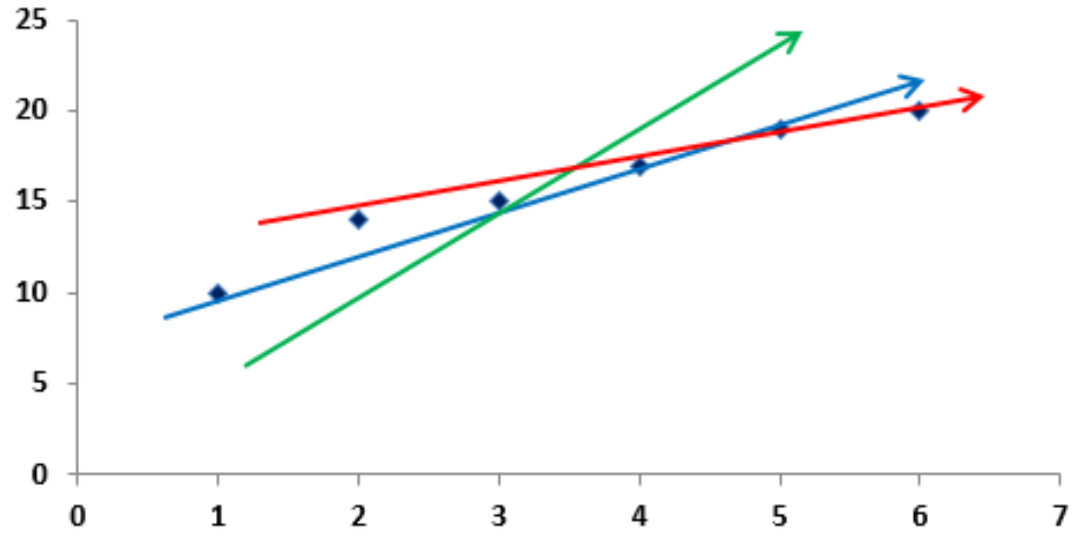
پس از برآورد ضرایب و ساخت مدل رگرسیونی، آزمون های فرض آماری جهت بررسی معناداری برازش مدل خطی و جزئیات بدست آمده بکار می رود. این آزمون ها در دو سطح معناداری مدل را مورد بررسی قرار می دهد:

- **معناداری برازش مدل خطی**

  بر اساس جدول تحلیل واریانس ANOVA

- **معناداری ضرایب مدل خطی**

  بر اساس آزمون t ضرایب برآورد شده



تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی

❑ آزمون های فرض مدل رگرسیون خطی

○ معناداری برازش مدل خطی

بر اساس جدول تحلیل واریانس ANOVA، پراکندگی (واریانس) مقادیر فیلد هدف به دو بخش تجزیه می شود. بخشی از پراکندگی که توسط مدل رگرسیونی توضیح داده می شود و بخش دیگر از آن که به عنوان جمله خطا در نظر گرفته می شود.

$$SST = SSR + SSE$$

صورت واریانس فیلد هدف

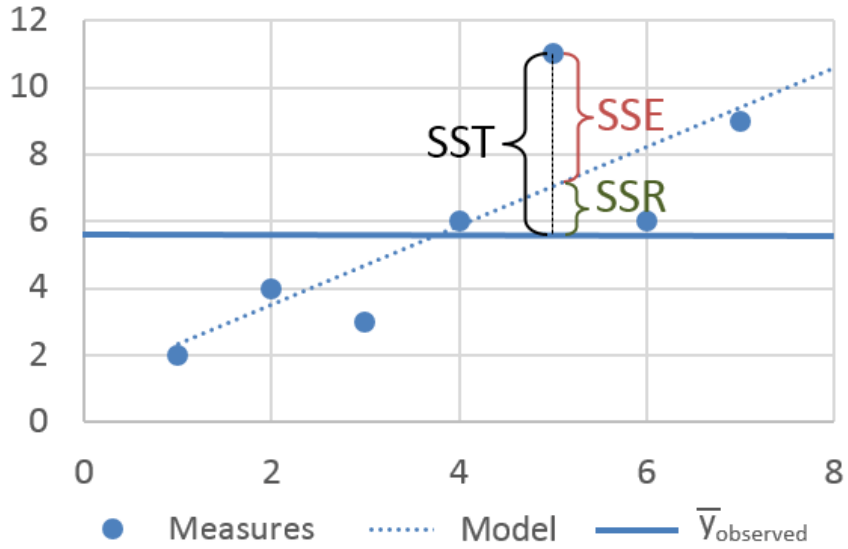
$$SST\,(Sum\,of\,Squared\,Total): \sum (y - \bar{y})^2$$

$$SSR\,(Sum\,of\,Squared\,Regression): \sum (\hat{y} - \bar{y})^2$$

$$SSE\,(Sum\,of\,Squared\,Error): \sum (y - \hat{y})^2$$

تابع هدف مدل رگرسیونی

بدیهی هست هرچقدر سهم SSR از مجموع مربعات کل داده ها (SST) بیشتر باشد، می توان نتیجه گرفت مدل بهتری خواهیم داشت.



❑ آزمون های فرض مدل رگرسیون خطی

- **معناداری برازش مدل خطی**

ساخت جدول تحلیل واریانس:

مقدار k برابر با تعداد ضرایب مدل رگرسیونی می باشد. در صورتیکه مقدار عرض از مبدا در مدل داشته باشیم و p تعداد ویژگی های ورودی به مدل باشد، مقدار K برابر با p+1 است.

منشاء تغییرات	درجه آزادی	مجموع مربعات	میانگین مربعات	آماره F
رگرسیون	k-۱	SSR	$MSR = \frac{SSR}{k-1}$	$F = \frac{MSR}{MSE}$
خطا	n-k	SSE	$MSE = \frac{SSE}{n-k}$	
کل	n-۱	SST		

مقدار n برابر با تعداد نمونه های مورد بررسی در مدل می باشد.

تولید محتوا: زهرا ذوالقدر

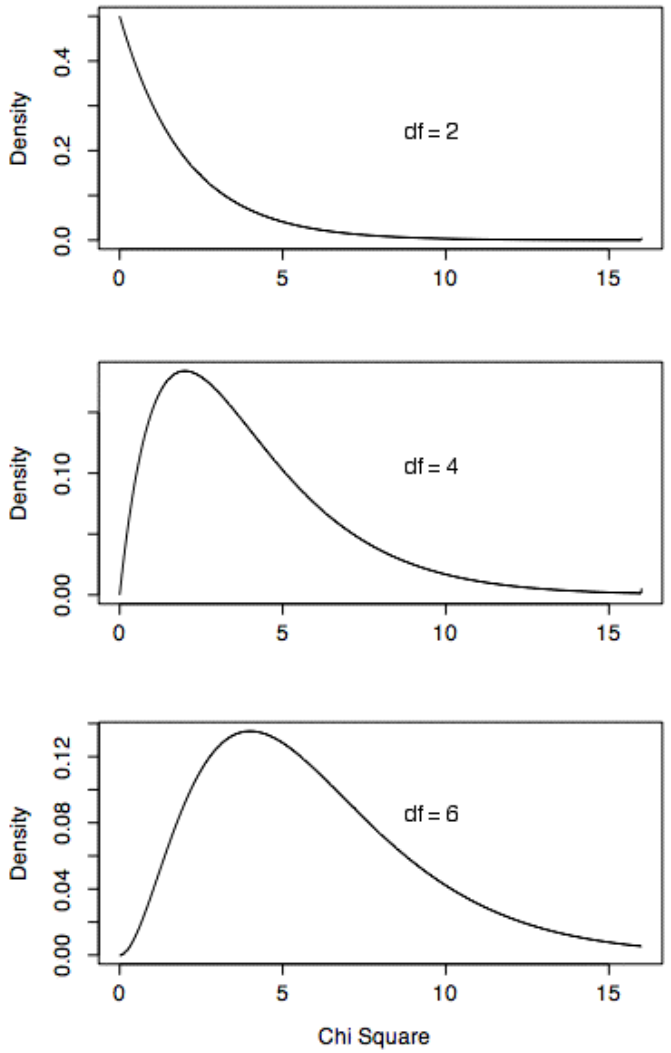
daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی



- آزمون های فرض مدل رگرسیون خطی
  - معناداری برازش مدل خطی

درجه آزادی: Degree of Freedom

مفهوم درجه آزادی به منابع پراکندگی مرتبط است. هر چقدر میزان درجه آزادی بیشتر باشد، انعطاف بیشتری در پراکندگی توزیع داده ها خواهیم داشت.

فرض کنید از شما خواسته شده تا 3 عدد دلخواه انتخاب کنید، بنابراین شما 3 درجه آزادی خواهید داشت تا هر عدد دلخواهی را برای این چالش در نظر بگیرید. حال اگر از شما بخواهند 3 عدد داخواه انتخاب کنید بطوریکه میانگین آنها برابر با مقدار فرضی 50 شود، آنگاه تنها دو درجه آزادی برای انتخاب دو عدد دلخواه خواهید داشت و عدد سوم بایستی مقداری در نظر گرفته شود که برابری مقدار میانگین با عدد 50 تضمین شود.

بطور کلی، به ازای برآورد هر پارامتر از مجموعه داده ها، یک واحد از درجه آزادی خطا کم می شود.

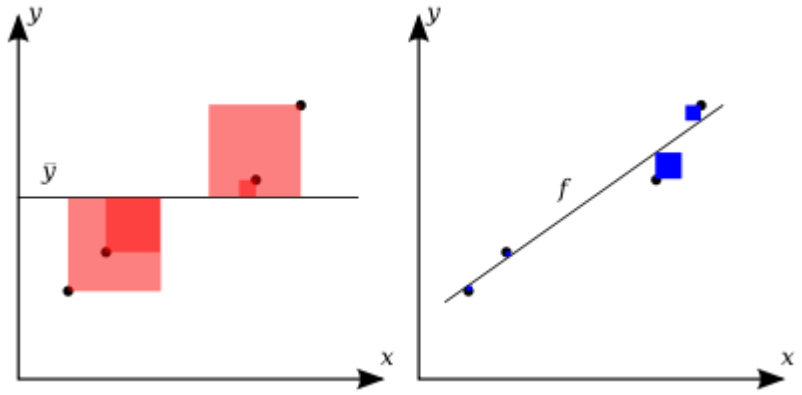
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ آزمون های فرض مدل رگرسیون خطی

○ معناداری برازش مدل خطی

ضریب تعیین R^2 : Coefficient of Determination

یکی از شاخص های مهم و پرکاربرد در تحلیل مدلهای رگرسیونی مقدار ضریب تعیین R^2 می باشد که نسبتی از پراکندگی کل فیلد هدف است که توسط مدل رگرسیونی توضیح داده می شود. در واقع این شاخص قدرت برازش مدل رگرسیون روی داده ها را اندازه گیری می کند.



$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$$

فرض کنید مقدار R^2 برای یک مدل 0.87 بدست آمده است. به این معنی می باشد که 87% از تغییرات فیلد هدف توسط مدل رگرسیونی کنترل و توضیح داده می شود که نشان دهنده میزان اثر بخشی ویژگی های ورودی به مدل می باشد.

❑ آزمون های فرض مدل رگرسیون خطی

- معناداری برازش مدل خطی

ضریب تعیین تعدیل شده $Adjusted\ R^2$

ضعف مهم شاخص R^2 اینست که با افزایش تعداد ویژگی های ورودی به مدل، بصورت کاذب افزایش می یابد. بنابراین با تعدیل آن نسبت به تعداد ویژگی های ورودی به مدل، مقدار مطمئنتری برای اندازه گیری قدرت برازش مدل خواهیم داشت.

$$Adjusted\ R^2 = 1 - \frac{\frac{SSE}{n-k}}{\frac{SST}{n-1}} = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

مقدار K برابر با تعداد پارامترهای برآورد شده مدل می باشد که برابر با تعداد ویژگی های ورودی به مدل (p) و پارامتر عرض از مبدا است.

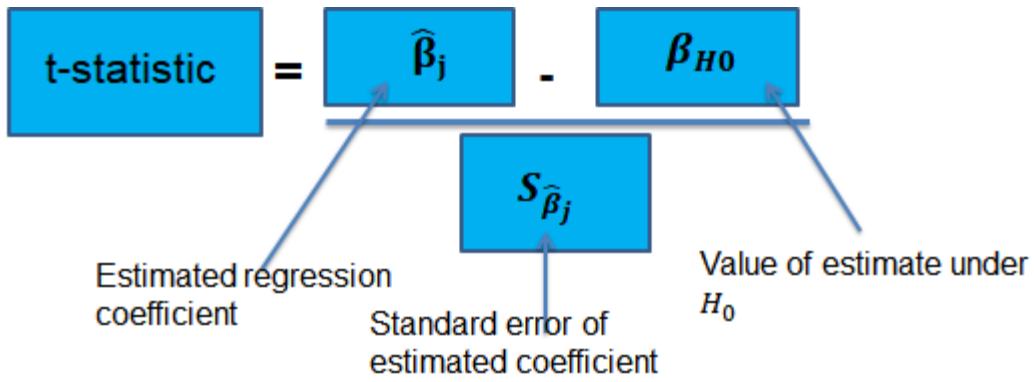
p ↑ | $N-p-1$ ↓ | $\frac{(1-R^2)(N-1)}{N-p-1}$ ↑ | $Adjusted\ R^2$ ↓

❑ آزمون های فرض مدل رگرسیون خطی

○ معناداری ضرایب مدل خطی

ضرایب مدل رگرسیونی برآورد شده دارای توزیع نرمال با میانگین مجهول β_i می باشند. بنابراین از طریق آزمون میانگین می توان فرض صفر بودن آنها را مورد بررسی قرار داد. با استفاده از آزمون t اینکار انجام می شود.

$$\begin{cases} H_0: \beta_i = 0 \\ H_1: \beta_i \neq 0 \end{cases}$$



در صورتیکه دلیلی بر رد فرض صفر بودن وجود نداشته باشد، می توان نتیجه گرفت ویژگی مربوط به آن ضریب دارای ارتباط معنادار خطی در مدل نمی باشد و با حذف آن مجددا مدلسازی شود.

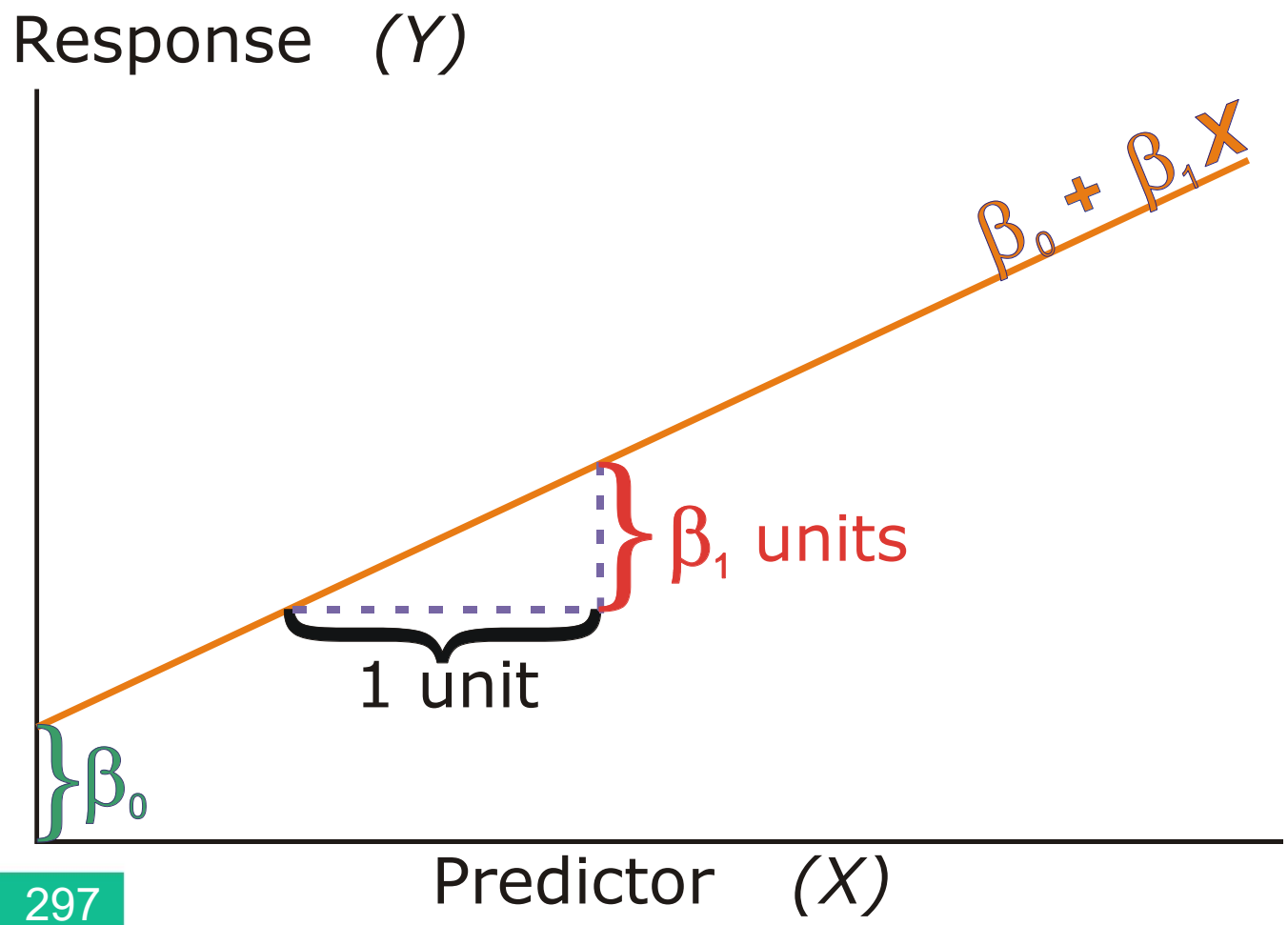
فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی

❑ آزمون های فرض مدل رگرسیون خطی

- معناداری ضرایب مدل خطی

در صورت معنادار بودن ضریب رگرسیونی، بر اساس مثبت یا منفی بودن آن می توان به نوع ارتباط مستقیم یا معکوس آن با فیلد هدف پی برد. همچنین اندازه ضریب نشان دهنده این می باشد، به ازای هر انحراف معیار تغییر در ویژگی مربوطه، به میزان ضریب بدست آمده در انحراف معیار فیلد هدف، مقدار فیلد هدف افزایش یا کاهش می یابد.



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی

- آزمون های فرض مدل رگرسیون خطی
  - مثال از خروجی آزمون های فرض مدل رگرسیونی

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.718[a]	.515	.500	18.639

a. Predictors: (Constant), Age1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11449.926	1	11449.926	32.958	.000[a]
	Residual	10769.710	31	347.410		
	Total	22219.636	32			

a. Predictors: (Constant), Age1

b. Dependent Variable: SBP1

مقدار P-Value در جدول آنالیز واریانس نشان دهنده معنادار بودن مدل خطی برازش داده شده می باشد.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	81.517	10.465		7.789	.000
	Age1	1.222	.213	.718	5.741	.000

a. Dependent Variable: SBP1

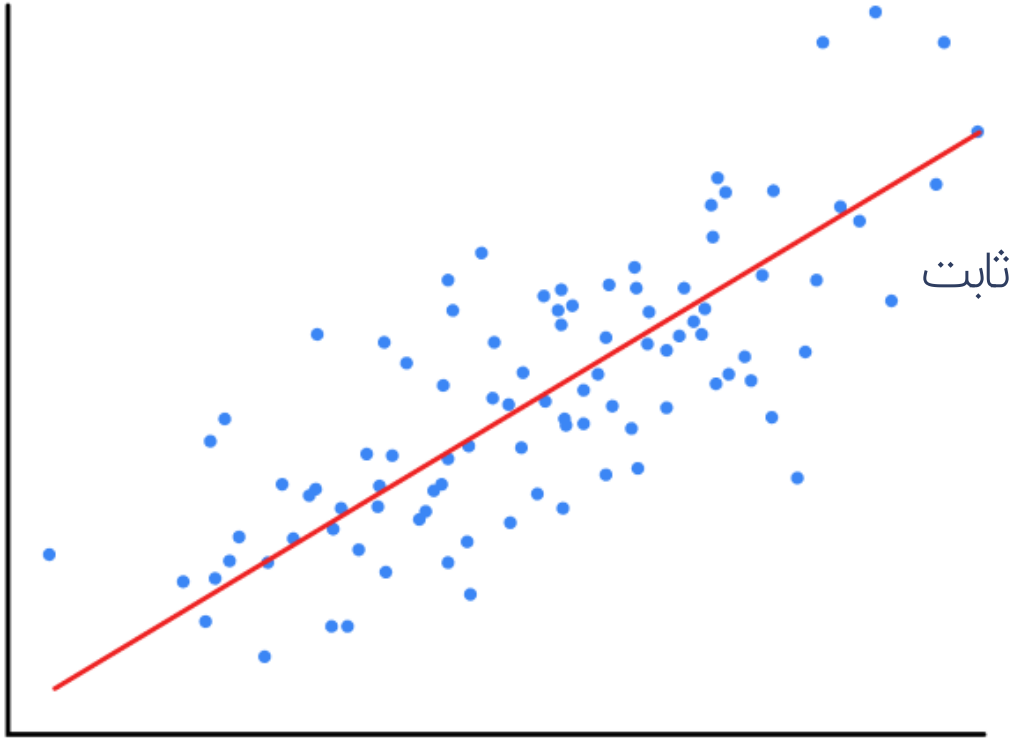
در صورتیکه از داده های استاندارد شده جهت مدلسازی استفاده شود، بدیهی است مقدار عرض از مبدا در مدل رگرسیونی برابر با صفر می باشد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی

❑ بررسی مفروضات مدل رگرسیون خطی



- وجود رابطه خطی بین ویژگی های ورودی و فیلد هدف
- توزیع نرمال خطاهای مدل رگرسیونی با میانگین صفر و واریانس ثابت
- عدم همبستگی بین خطاهای مدل رگرسیونی
- عدم هم خطی بین ویژگی های ورودی به مدل

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی

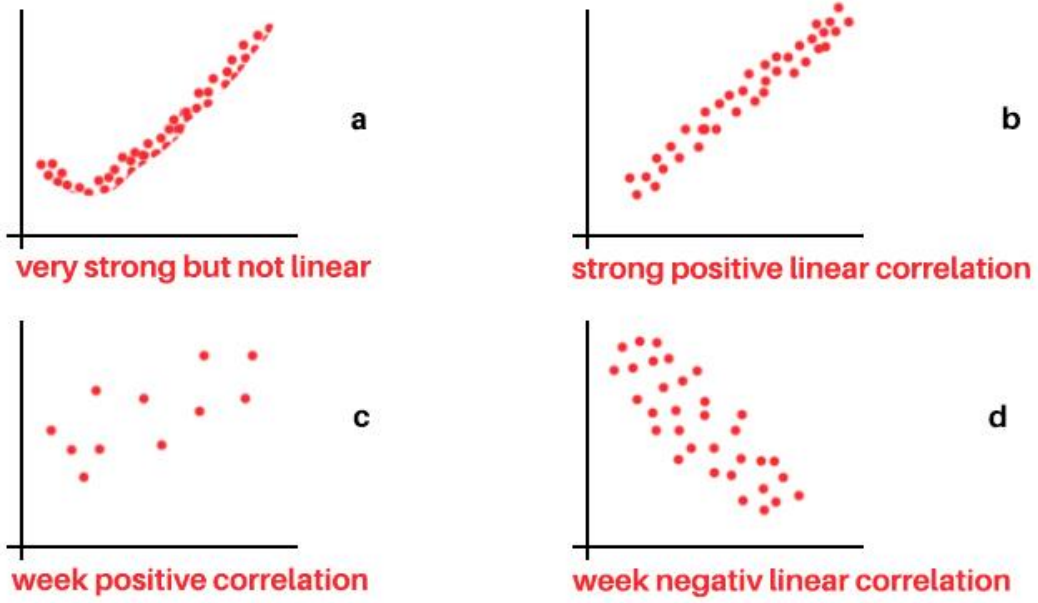
بررسی مفروضات مدل رگرسیون خطی

وجود رابطه خطی بین ویژگی های ورودی و فیلد هدف

یکی از فرضیات اساس مدل رگرسیون خطی، وجود ارتباط خطی بین ویژگی های ورودی با فیلد هدف می باشد.

بررسی این موضوع عمدتا از طریق رسم نمودار پراکنش بین ویژگی های ورودی و فیلد هدف و همچنین محاسبه شاخص هایی مانند آماره همبستگی پیرسن، اسپیرمن و ... انجام می شود.

در صورتی که رابطه خطی بین ویژگی و فیلد هدف مشاهده نشد، می توان با استفاده از توابع تبدیل لگاریتمی و ... روی ویژگی های ورودی، فرض رابطه خطی برای مدل رگرسیون را ایجاد نمود.



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

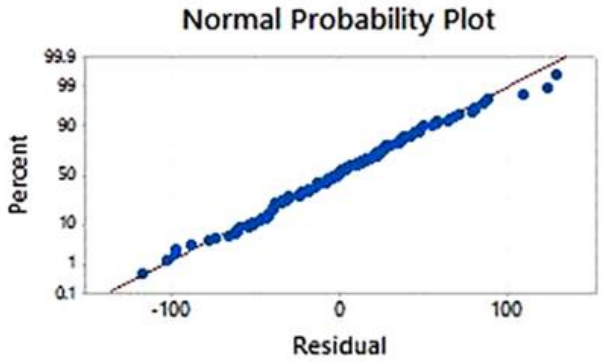
فرآیند داده کاوی

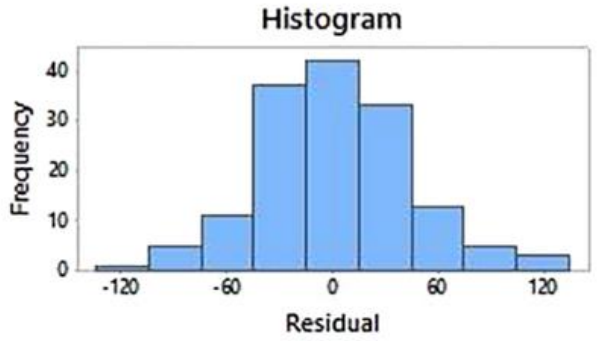
مدل های پیش بینانه – الگوریتم رگرسیون خطی

بررسی مفروضات مدل رگرسیون خطی

- **توزیع نرمال خطاهای مدل رگرسیونی با میانگین صفر و واریانس ثابت**

فرض دیگری که بعد از ساخت مدل رگرسیونی و محاسبه خطاهای مدل مورد بررسی قرار می گیرد، بررسی نرمال بودن توزیع آماری خطاهای مدل می باشد.





بررسی نرمال بودن توزیع از طریق رسم نمودار هیستوگرام یا نمودار چندک - چندک خطاهای مدل یا آزمون نیکویی برازش توزیع نرمال انجام می شود. همچنین می توان بررسی های فوق را روی داده های فیلد هدف نیز انجام داد، زیر دارای ارتباط خطی با خطاها می باشد.

درصورت عدم فرض نرمال بودن، با انجام برخی تبدیل ها از جمله تبدیل باکس – کاکس و ... روی فیلد هدف، می توان شرایط مناسب برای ساخت مدل رگرسیونی را ایجاد نمود.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

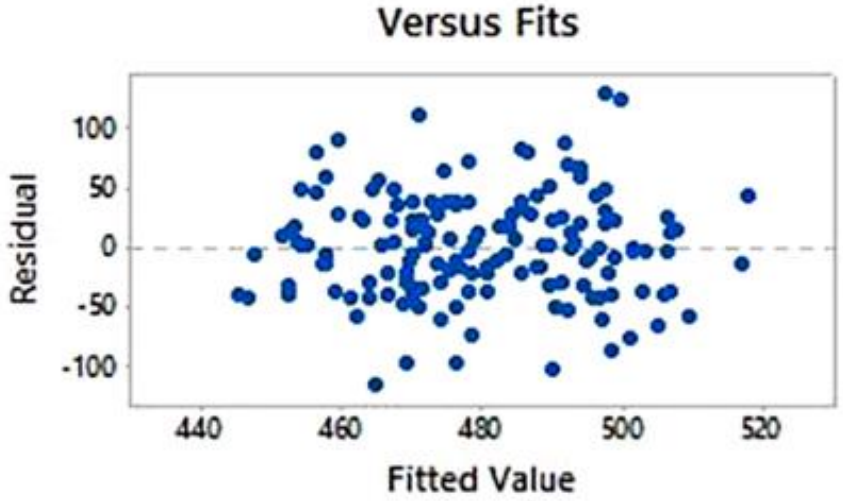
فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی

❑ بررسی مفروضات مدل رگرسیون خطی

○ توزیع نرمال خطاهای مدل رگرسیونی با میانگین صفر و واریانس ثابت

میانگین توزیع خطاهای مدل بایستی مقدار صفر باشد. در صورتی که مقدار میانگین خطاها بیشتر یا کمتر از صفر باشد، به این معنی می باشد که پیش بینی مدل ها دارای بیش/کم برآوردی (Over/Under Estimate) می باشد.



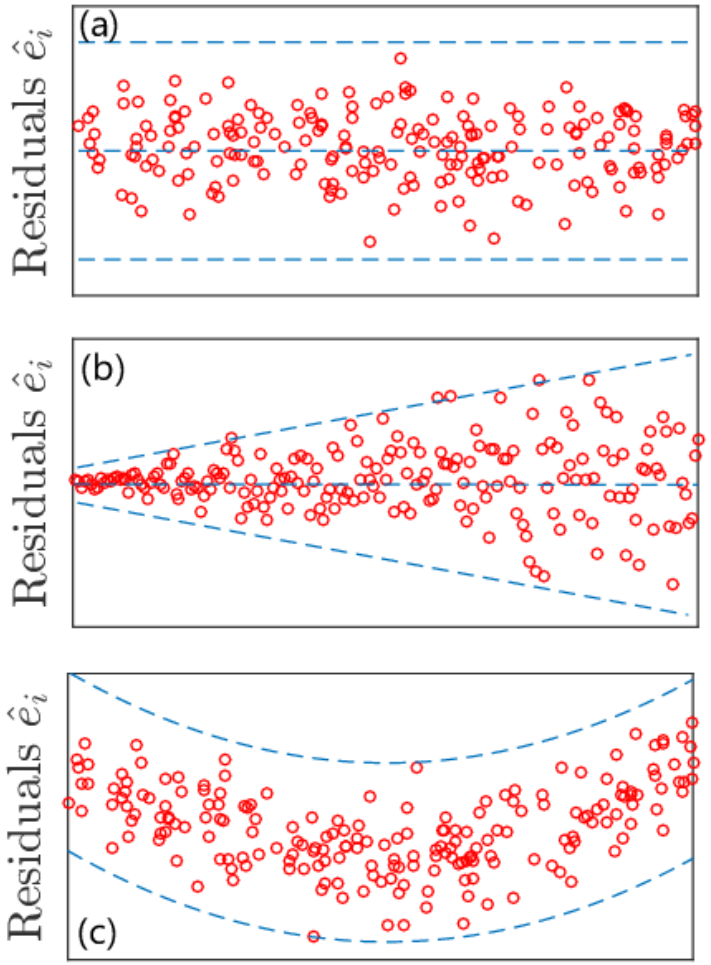
جهت بررسی میانگین خطا های مدل، کافیست نمودار پراکنش مقادیر خطا (باقیمانده ها) را در مقابل مقادیر پیش بینی شده مدل رسم کنیم. انتظار می رود داده ها در اطراف خط میانگین صفر پراکنده شده باشند.

در صورتی که مقادیر به سمت بالا یا پایین خط میانگین صفر، اریبی داشته باشد، می تواند به معنی کم/زیاد برآورد شدن مقدار عرض از مبدا مدل بوده و با اضافه کردن میزان اریبی به مدل، دقت پیش بینی را اصلاح نمود.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی



- **بررسی مفروضات مدل رگرسیون خطی**
  - **توزیع نرمال خطاهای مدل رگرسیونی با میانگین صفر و واریانس ثابت**

نکته بسیار مهم دیگر در بررسی توزیع آماری خطاهای مدل، وجود فرض ثابت بودن واریانس خطاها می باشد.

جهت بررسی واریانس خطا های مدل، کافیست نمودار پراکنش مقادیر خطا (باقیمانده ها) را در مقابل مقادیر پیش بینی شده مدل و همچنین ویژگی های ورودی رسم کنیم. انتظار می رود داده ها در اطراف خط میانگین صفر به صورت کاملا تصادفی و بدون هیچ الگویی پراکنده شده باشند.

در صورتی که با افزایش مقادیر پیش بینی، میزان خطای مدل روند کاهشی، افزایشی یا خطی داشته باشد، به معنی عدم ثبات واریانس بوده و نیاز به تبدیل های لگاریتمی، چند جمله ای و ... روی مقادیر فیلد هدف یا ویژگی های ورودی خواهیم داشت.

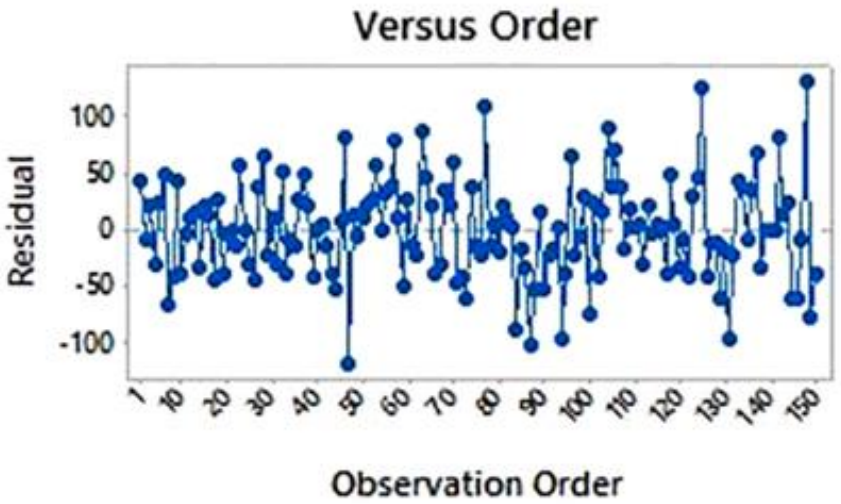
فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون خطی

بررسی مفروضات مدل رگرسیون خطی

- **ناهمبستگی بین خطاهای مدل رگرسیونی**

فرض دیگری که بایستی مورد توجه قرار گیرد، عدم همبستگی بین مقادیر متوالی (خودهمبستگی) و الگوهای معنادار در میزان خطای مدل است، زیرا برآورد واریانس و انحراف معیار ضرایب مدل به درستی انجام نمی شود.



جهت بررسی خودهمبستگی و روند در خطاهای پیش بینی، کافیست مقادیر خطاها را در مقابل ترتیب رکوردها رسم کنیم و با اتصال نقاط به بررسی آن بپردازیم.

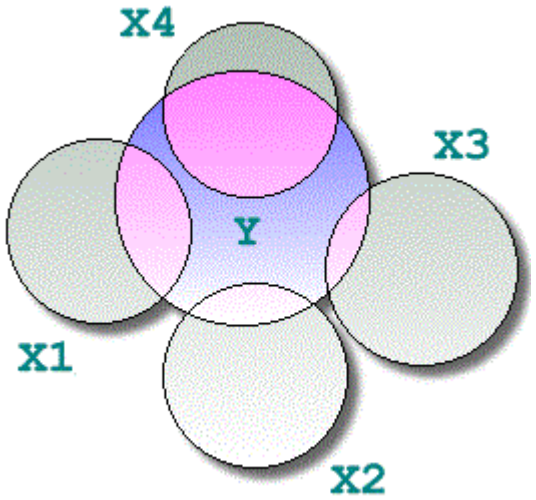
یکی از آماره های پرکاربرد در بررسی خودهمبستگی، آماره دوربین – واتسون می باشد. در صورتی که توزیع خطاها (باقیمانده ها) نرمال باشد، می توان از این آماره استفاده کرد. مقدار این آماره در بازه 0 تا 4 می باشد و مقدار 2 به معنی عدم وجود خودهمبستگی است. معمولا قرارگیری این آماره در بازه 1.5 تا 2.5 قابل قبول می باشد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

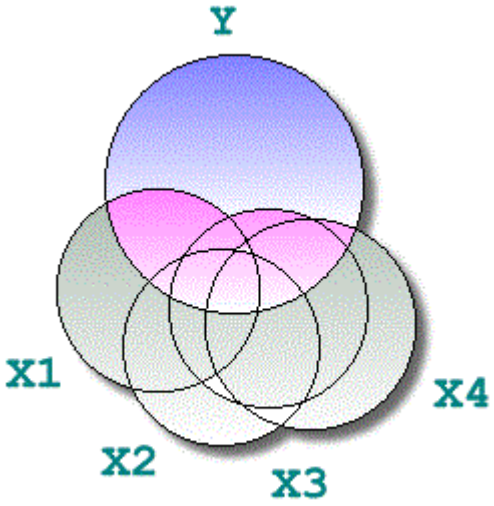
❑ بررسی مفروضات مدل رگرسیون خطی

- **عدم هم خطی/ هم خطی چندگانه بین ویژگی های ورودی به مدل (Collinearity/Multicollinearity)**

یکی از مفروضات مهم در برآورد حداقل مربعات (OLS) فرض مستقل بودن ورودی های مدل می باشد. نقض این فرض منجر به اریبی در برآورد ضرایب مدل می گردد.



ویژگی های ورودی مستقل از هم



ویژگی های ورودی دارای هم خطی

جهت بررسی این فرض، استفاده از نمودار پراکنش و ماتریس واریانس-کوواریانس بین ویژگی های ورودی کمک کننده می باشد. همچنین پس از ایجاد مدل رگرسیونی شاخص فاکتور تورم واریانس (Variance Inflation Factor – VIF) برای هر ویژگی محاسبه شده و مقادیر بیشتر از 10 نشان از مشکل هم خطی آن ویژگی با سایر ویژگی ها می باشد.

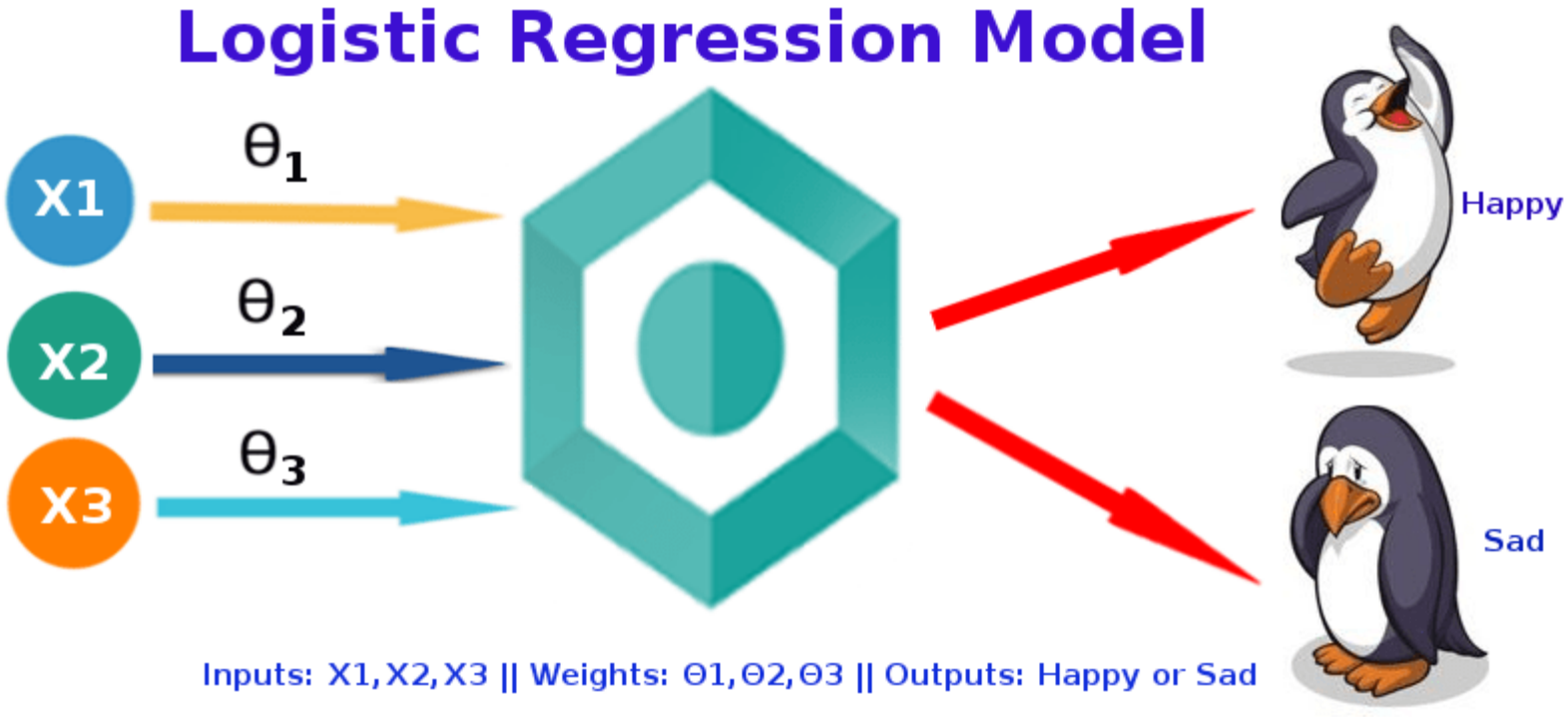
روش برخورد مناسب، حذف ویژگی، استفاده از روش PCA و یا روش تنظیم سازی (Regularization) در برآورد ضرایب مدل رگرسیونی می باشد.

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون لجستیک

❑ مقدمه ای بر رگرسیون لجستیک (Logistic Regression)

یکی از انواع الگوریتم های با یادگیری با نظارت از نوع رده بندی می باشد که بر خلاف رگرسیون خطی به دنبال بررسی رابطه و مدلسازی ویژگی های ورودی مستقل با فیلد هدف (پاسخ) از نوع کیفی می باشد.



بطور مثال:

- آیا ایمیل دریافت شده اسپم هست یا خیر؟
- آیا تومور دیده شده بدخیم هست یا خیر؟
- آیا مشتری ریزش می کند یا خیر؟
- و ...

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

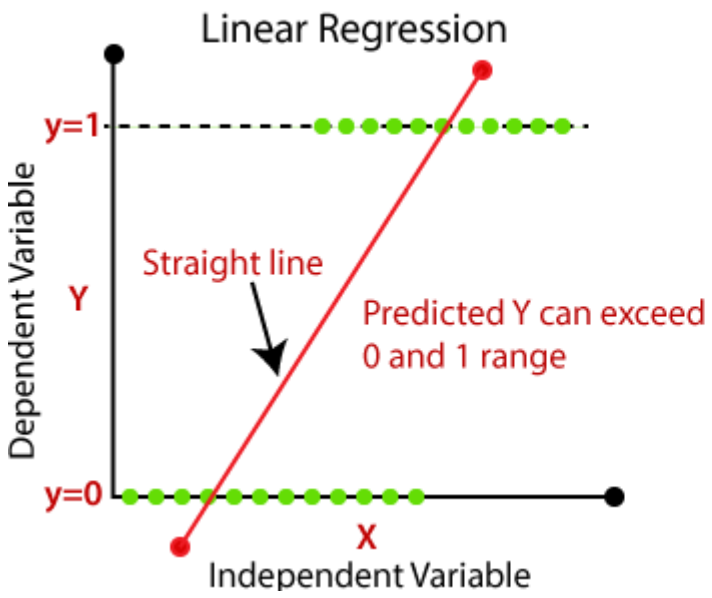
dayche.com | گروه دایکه

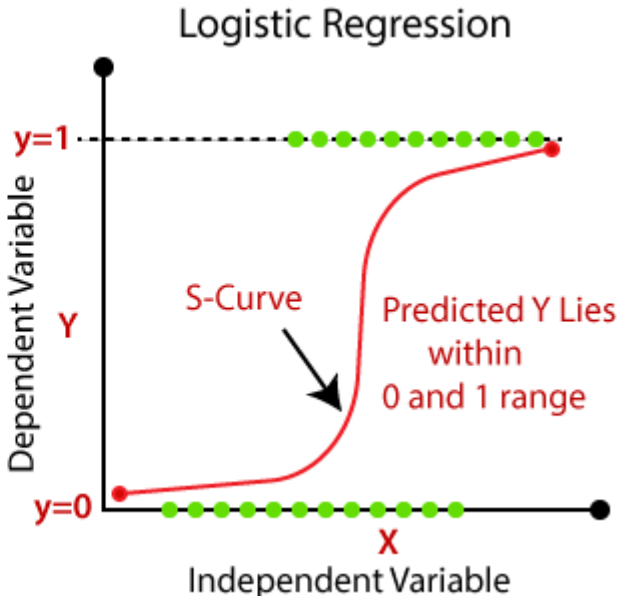
فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون لجستیک

❑ مقدمه ای بر رگرسیون لجستیک (Logistic Regression)

یکی از انواع الگوریتم های با یادگیری با نظارت از نوع رده بندی می باشد که بر خلاف رگرسیون خطی به دنبال بررسی رابطه و مدلسازی ویژگی های ورودی مستقل با فیلد هدف (پاسخ) از نوع کیفی می باشد.





با توجه به توزیع برنولی فیلد هدف خواهیم داشت:

$$\hat{y} = \mathrm{E}(Y|X = x) = \mathrm{P}(Y = 1|X = x) = \pi(x)$$

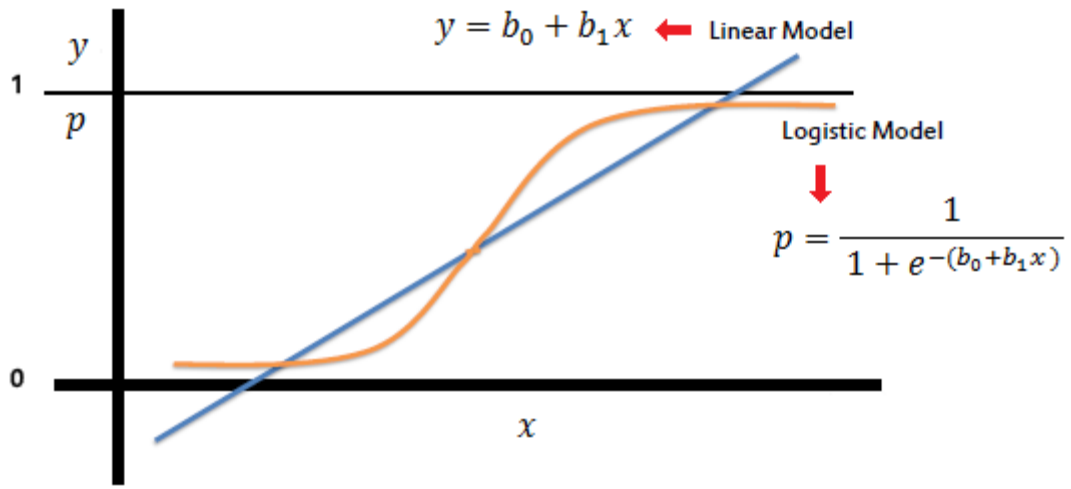
بر اساس رابطه فوق، مقدار پیش بینی برای برای فیلد هدف، بایستی مقدار احتمال در بازه 0 تا 1 باشد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ مقدمه ای بر رگرسیون لجستیک (Logistic Regression)

○ تابع لجستیک (سیگموید) Logistic (Sigmoid) Function

با توجه به اینکه مقدار $\pi(x)$ در بازه 0 تا 1 قرار می گیرد، بنابراین نیاز هست جهت پیش بینی مقدار y از تابعی مانند تابع لجستیک استفاده شود تا مقدار خروجی در بازه 0 تا 1 قرار گیرد. تابع لجستیک برای خط رگرسیون ساده به شکل زیر تعریف می گردد:



$$\sigma(t) = \frac{1}{1+e^{-t}} = \frac{e^t}{1+e^t}, \qquad t = b_0 + b_1 x$$

$$\hat{y} = \pi(x) = \frac{1}{1+e^{-(b_0+b_1 x)}} = \frac{e^{(b_0+b_1 x)}}{1+e^{(b_0+b_1 x)}}$$

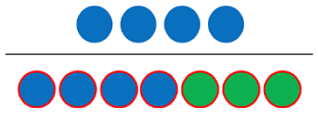
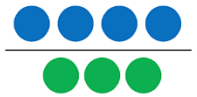
با توجه به پیچیدگی نمایش مدل برآورد y، علاقمند هستیم با تبدیل های مناسب نمایش آن را ساده تر کنیم. به این منظور از مفهومی به نام بخت (Odds) استفاده می کنیم.

❑ مقدمه ای بر رگرسیون لجستیک (Logistic Regression)

- مفهوم بخت (Odds)

محاسبه بخت بر اساس نسبت احتمال وقوع به احتمال عدم وقوع انجام می شود. بنابراین مقدار بخت، بر خلاف مقدار احتمال می تواند بزرگتر از یک نیز باشد. **بطور مثال** فرض کنید احتمال خوش خیم بودن غده سرطانی برای یک شخص 0.75 برآورد شده است، بنابراین مقدار بخت برای این فرد 3 می باشد.

Probability vs Odds

	Risk	Odds
Mathematically	$P(p) = \frac{p}{p+q}$	$O(p) = \frac{\frac{p}{p+q}}{\frac{q}{p+q}} = \frac{p(p+q)}{q(p+q)} = \frac{p}{q}$
Graphically		

با توجه به این مفهوم روابط زیر را خواهیم داشت:

$$Odds(\pi(x)) = \frac{\pi(x)}{1-\pi(x)} = \frac{\frac{e^{(b_0+b_1x)}}{1+e^{(b_0+b_1x)}}}{1-\frac{e^{(b_0+b_1x)}}{1+e^{(b_0+b_1x)}}}$$

$$Odds(\pi(x)) = \frac{\pi(x)}{1-\pi(x)} = e^{(b_0+b_1x)}$$

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

❑ مقدمه ای بر رگرسیون لجستیک (Logistic Regression)

○ تبدیل لوجیت (Logit Transformation)

لگاریتم طبیعی بخت به عنوان تبدیل لوجیت یا بعضا (Log-Odds) شناخته می شود. در رگرسیون لجستیک با استفاده از تبدیل لوجیت بر روی احتمال وقوع x (برآورد مقدار y) می توان رابطه خطی رگرسیون را به شکل زیر نمایش داد.

$$Logit\big(\pi(x)\big) = Ln\Big(Odds\big(\pi(x)\big)\Big) = Ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = Ln\big(e^{(b_0+b_1x)}\big) = b_0 + b_1x$$

پس از برآورد ضرایب مدل بر اساس روش حداکثر درستنمایی (MLE) در رگرسیون لجستیک، بر خلاف رگرسیون خطی، مقدار y بصورت مستقیم برآورد نمی شود. بلکه مقدار $Logit(\hat{y})$ بصورت ترکیب خطی از ویژگی های ورودی محاسبه شده و مقدار احتمال کلاس یک (کلاس مبنا) برای تصمیم گیری و رده بندی استفاده می شود.

$$\hat{y} = \mathrm{E}(Y|X=x) = \mathrm{P}(Y=1|X=x) = \pi(x) = \frac{1}{1+e^{-(b_0+b_1x)}}$$

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم رگرسیون لجستیک

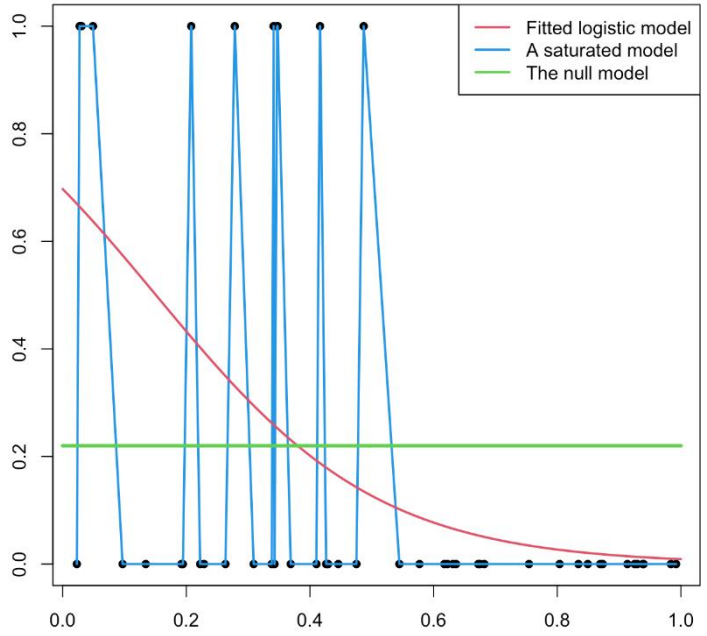
❑ آزمون های فرض رگرسیون لجستیک

در این بخش نیز بررسی آماری مدل برازش شده در دو سطح معناداری برازش مدل و ضرایب بدست آمده انجام می شود.

- **معناداری برازش مدل لجستیک**

یکی از روش های ارزیابی میزان برازش مدل در تحلیل رگرسیون لجستیک استفاده از نسبت درستنمایی (Likelihood Ratio) است که توزیع آماره آن χ^2 می باشد و معادل آماره F در تحلیل واریانس رگرسیون خطی است.

ایده اصلی این روش مقایسه میزان درستنمایی برای مدل برازش داده شده با مدل کامل یا اشباع شده (Saturated Model) است و شاخص انحراف Deviance گفته می شود که بایستی مینیمم شود.



$$Deviance = 2(\log\left(L(\widehat{\theta_s})\right) - \log\left(L(\widehat{\theta_m})\right))$$

$\log\left(L(\widehat{\theta_s})\right)=0$ ⟹

$$Deviance = -2\log\left(L(\widehat{\theta_m})\right)$$

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ آزمون های فرض رگرسیون لجستیک

در این بخش نیز بررسی آماری مدل برازش شده در دو سطح معناداری برازش مدل و ضرایب بدست آمده انجام می شود.

○ معناداری برازش مدل لجستیک

روش دیگری که مناسب بودن برازش مدل را مورد بررسی قرار می دهد، آزمون آماری Hosmer-Lemeshow یکی از پرکاربردترین آزمونهای نکویی برازش مدل لجستیک می باشد که با مقایسه نرخ مقادیر پیش بینی شده و واقعی در زیرگروه های مختلف از داده ها، مناسب بودن مدل را اندازه گیری می کند.

توزیع آماره این آزمون نیز از χ^2 پیروی می کند و فرض اولیه مورد آزمون، برابری نرخ مقادیر پیش بینی شده و واقعی است. بنابراین در صورتی که دلیل آماری بر رد فرض اولیه وجود نداشته باشد (مقدار P-Value بزرگتر از سطح معناداری باشد) نشان دهنده مناسب بودن مدل برازش داده شده می باشد.

❑ آزمون های فرض رگرسیون لجستیک

○ معناداری ضرایب مدل لجستیک

آزمون معناداری ضرایب مدل رگرسیون لجستیک، با استفاده از آماره والد (Wald) انجام می شود. این آزمون معادل آزمون t در رگرسیون خطی می باشد و فرض اولیه برابری مقدار ضرایب مدل با صفر را مورد آزمون قرار می دهد.

$$\begin{cases} H_0: \beta_i = 0 \\ H_1: \beta_i \neq 0 \end{cases}$$

$$W = \left(\frac{\beta_j}{s_{\beta_j}} \right)^2$$

در صورتیکه دلیلی بر رد فرض صفر بودن وجود نداشته باشد، می توان نتیجه گرفت ویژگی مربوط به آن ضریب دارای ارتباط معناداری در مدل نمی باشد و با حذف آن مجددا مدلسازی شود.

❑ آزمون های فرض رگرسیون لجستیک

○ معناداری ضرایب مدل لجستیک

نسبت بخت (Odds Ratio)

در صورتی که ضریب یک ویژگی بر اساس آزمون والد معنادار شناخته شود، تفسیر میزان اثرگذاری آن بر اساس شاخص نسبت بخت انجام می شود این شاخص به معنای نسبت بخت وقوع یک پیامد با فرض تعلق به گروه اول به بخت وقوع آن در صورت تعلق به گروه دوم است.

بطور مثال:

در صورتی که احتمال بازپرداخت به موقع تسهیلات در مردان، 0.75 و در زنان 0.8 باشد، بنابراین شاخص بخت مردان در بازپرداخت به موقع، برابر با 3 و در زنان 4 می شود. بر این اساس نسبت بخت مردان به زنان در بازپرداخت به موقع معادل 0.75 است. یعنی مردان نسبت به زنان 0.25 بخت کمتر در بازپرداخت به موقع تسهیلات خواهند داشت.

$$Odds = \frac{p_1}{1 - p_1}$$

$$Odds\ Ratio = \frac{p_1/(1 - p_1)}{p_2/(1 - p_2)}$$

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ آزمون های فرض رگرسیون لجستیک

- معناداری ضرایب مدل لجستیک

رابطه ضرایب مدل و نسبت بخت

تفسیر ضرایب مدل رگرسیون لجستیک بر اساس مفهوم نسبت بخت صورت می گیرد. ضرایب ویژگی های ورودی کمی به نسبت یک واحد افزایش در مقادیر آن و همچنین ضرایب ورودی های کیفی به نسبت تغییر رده (گروه) به رده مبنا (Base Category) تفسیر می گردد.

$if\ x = SEX\ with\ coding:$
$male = 1\ \&\ female = 0$

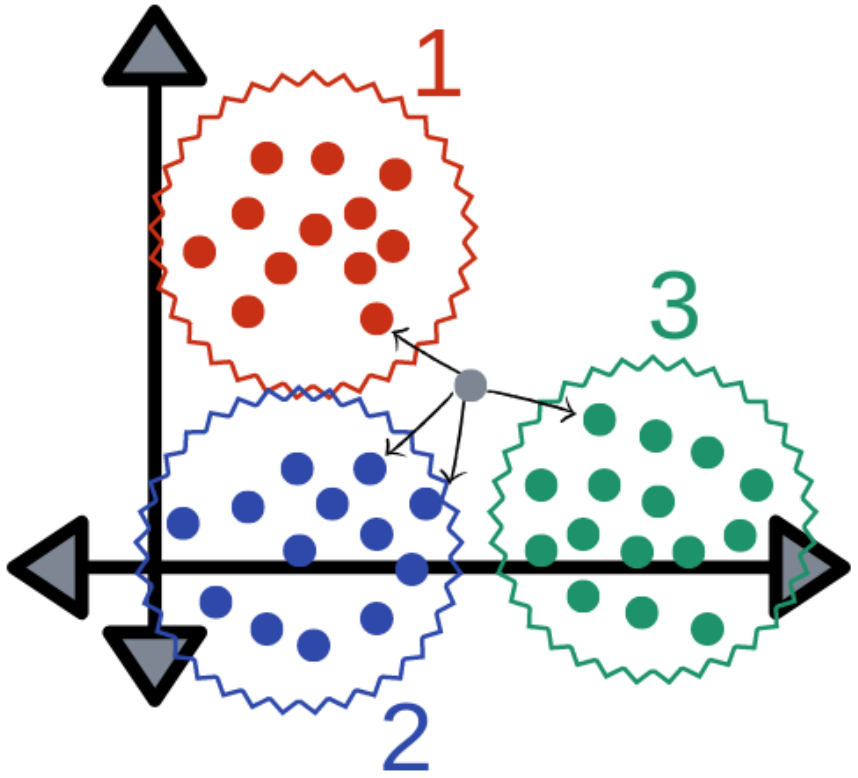
$$OR = \frac{Odds(male)}{Odds(female)} = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} = \frac{e^{(b_0+b_1*1)}}{e^{(b_0+b_1*0)}} = e^{b_1}$$

$if\ x = AGE$

$$OR = \frac{Odds(x+1)}{Odds(x)} = \frac{\frac{\pi(x+1)}{1-\pi(x+1)}}{\frac{\pi(x)}{1-\pi(x)}} = \frac{e^{(b_0+b_1*(x+1))}}{e^{(b_0+b_1*x)}} = e^{b_1}$$

❑ الگوریتم K – نزدیکترین همسایه K–Nearest Neighbors (KNN)

الگوریتم KNN یکی از انواع الگوریتم های جعبه سیاه با مدل یادگیری با نظارت می باشد که قابلیت بکارگیری در انواع مسائل رده بندی و رگرسیون را دارد.



ایده اصلی الگوریتم KNN اینست که چیزهای مشابه در نزدیکی هم قرار دارند.

به عبارتی با اندازه گیری میزان فاصله بین رکوردها، می توان این فرض را در نظر گرفت که رکوردهای نزدیکتر به هم دارای مشابهت هستند.

کبوتر با کبوتر، باز با باز

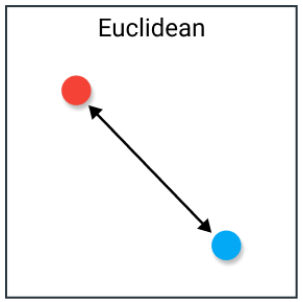
کند هم جنس با هم جنس پرواز

❑ الگوریتم K – نزدیکترین همسایه K–Nearest Neighbors (KNN)

معیارهای متفاوتی برای اندازه شباهت می توان در نظر گرفت:

- **فاصله اقلیدسی**

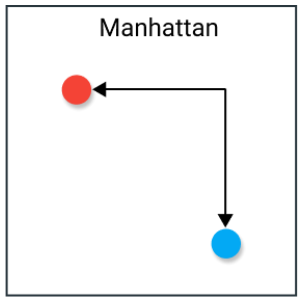
رایج ترین معیار فاصله که در صورت کم بودن تعداد ویژگی ها به علت سادگی و شهودی بودن آن، نتایج بسیار خوبی خواهد داشت.



$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- **فاصله منهتن**

درک شهودی و تجسم این معیار نسبت به فاصله اقلیدسی سخت تر است ولی معمولا در داده های با ابعاد بالا عملکرد خوبی نشان می دهد.



$$D(x,y) = \sum_{i=1}^{n}|x_i - y_i|$$

نکته مهمی که در اندازه فاصله بایستی در نظر داشت، هم مقیاس بودن مقادیر ویژگی های مورد بررسی می باشد.

تولید محتوا: زهرا ذوالقدر

daychegroup

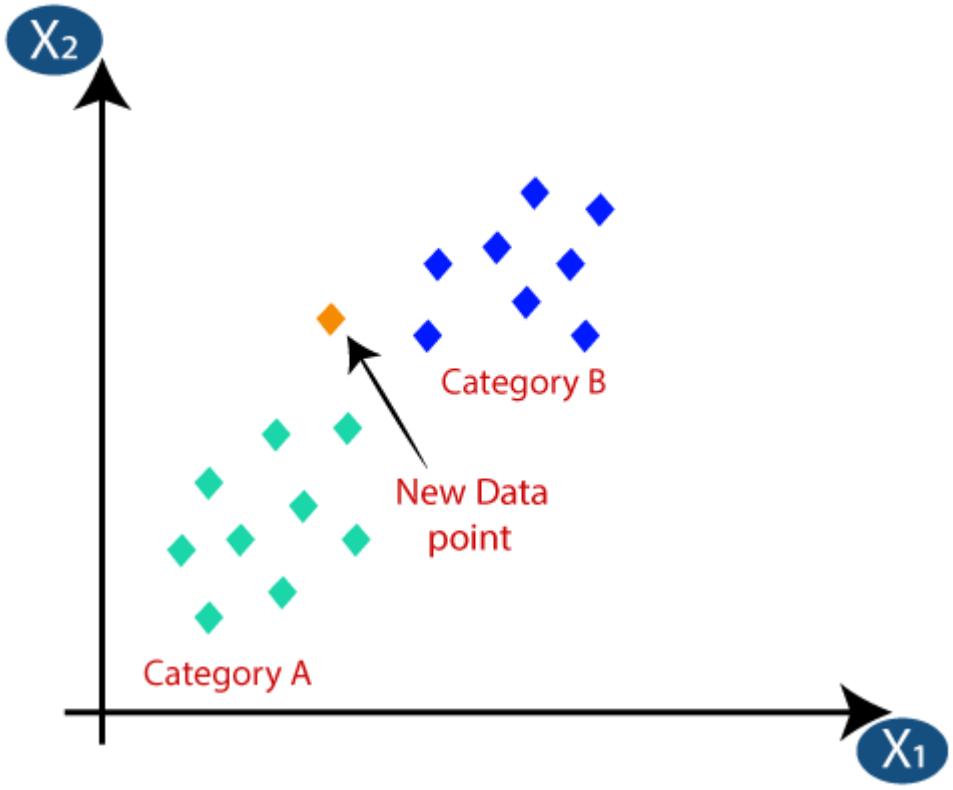
daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم k – نزدیکترین همسایه

❑ الگوریتم K – نزدیکترین همسایه K–Nearest Neighbors (KNN)



- نحوه عملکرد الگوریتم KNN:

1. تعیین مقدار K
2. محاسبه مقدار فاصله (معیار شباهت) برای رکورد مورد نظر با تمامی رکوردهای موجود
3. مرتب سازی شماره (اندیس) رکوردها بر اساس کوچکترین مقادیر فاصله
4. انتخاب K رکورد اول از لیست مرتب سازی شده
5. پیش بینی مقدار هدف
   - **مسئله رده بندی:** محاسبه مد (بیشترین فراوانی) کلاس هدف در K رکورد منتخب
   - **مسئله رگرسیون:** محاسبه میانگین (میانه) مقدار هدف در K رکورد منتخب

در مسائل رده بندی، معمولا از مقادیر فرد برای K استفاده می شود تا در رای گیری کلاس هدف، اطمینان از انتخاب کلاس داشته باشیم.

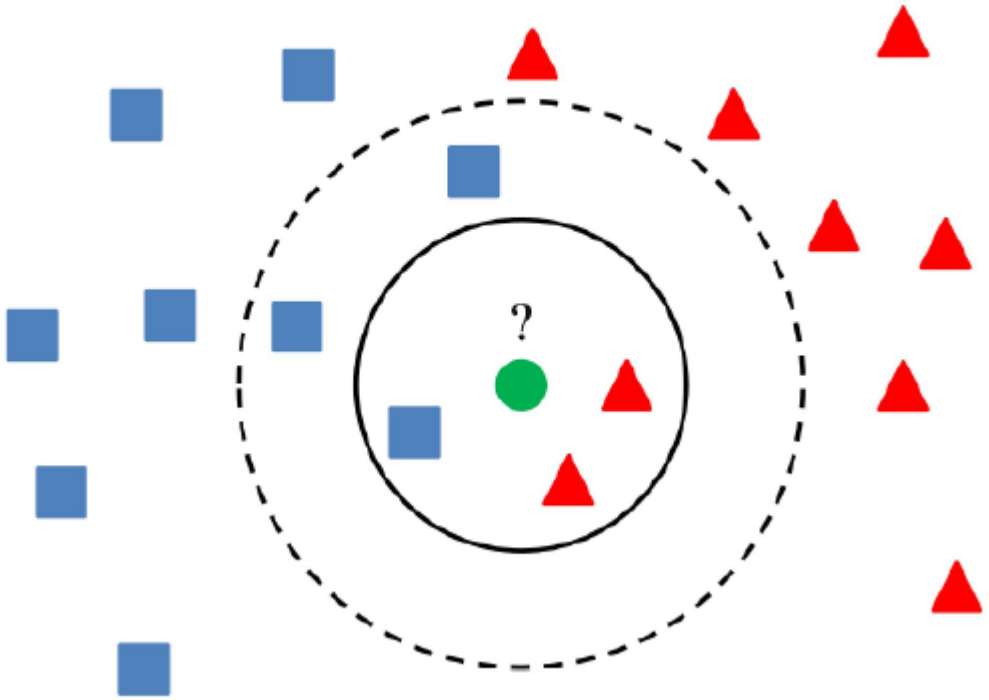
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم k – نزدیکترین همسایه

❑ الگوریتم K – نزدیکترین همسایه (KNN) K–Nearest Neighbors

تنها چالش مهم این الگوریتم تعیین مقدار K می باشد:



در صورتی که مقدار K کوچک انتخاب شود، منجر به نتایج ناپایدار می شود و با انتخاب مقادیر بزرگ K می توان ناحیه تصمیم گیری را هموارتر نمود و نتایج پایدارتری را ایجاد نمود؛ ولی در مسائل رده بندی نامتوازن ریسک نادیده گرفتن کلاس مینور (کلاس حداقلی) را افزایش داده و دچار خطای جدی در پیش بینی درست خواهد شد.

- **انتخاب K بهینه**

معمولا با انتخاب طرح اعتبارسنجی متقابل و برای مقادیر مختلف K، مدل KNN اجرا شده و مقدار میانگین خطای پیش بینی به ازای هر K محاسبه و در نمودار نمایش داده می شود. مقدار K با کمترین میزان خطا، به عنوان K بهینه انتخاب می گردد.

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم k – نزدیکترین همسایه

❑ الگوریتم K – نزدیکترین همسایه (KNN) K–Nearest Neighbors

یکی از کاربردهای مهم الگوریتم KNN، صرفا یافتن نزدیکترین موارد مشابه می باشد. بنابراین در این کاربرد، حضور فیلد هدف الزامی نیست و گزارش لیستی از K رکورد مشابه هدف مسئله می باشد.

مثال 1: فرض کنید تعداد 100 پرونده اظهارنامه مالیاتی متخلف (دارای عددسازی و مقادیر غیر واقعی) را شناسایی کرده ایم. با استفاده از این الگوریتم می توان به ازای هر یک از این پرونده ها، تعداد K پرونده مشابه از میان هزاران پرونده در صف ممیزی را استخراج نمود و با الویت بازرسی بالاتر، در زمانی سریعتر به سایر پرونده های متخلف دسترسی پیدا کنیم.

مثال 2: فرض کنید به عنوان یک اپلیکیشن پخش موزیک، مجموعه داده هایی متشکل از ژانر، ابزار موسیقی، نوازنده، خواننده و امتیاز کاربران را برای هر یک از ترانه های آرشیو خود در دسترس دارید. با استفاده از الگوریتم KNN می توان برای هر ترانه به تعداد K ترانه با نزدیکترین میزان شباهت را تعیین نمود و به عنوان یک سیستم پیشنهاد دهنده (Recommender Systems) به کاربران اپلیکیشن، ترانه پیشنهادی بعدی را معرفی کرد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ الگوریتم K – نزدیکترین همسایه (KNN) K–Nearest Neighbors

الگوریتم KNN از مجموعه مدلهای استنتاج مبتنی بر موارد (Case Based Reasoning) می باشد و بر خلاف سایر الگوریتم های دیگر هیچ مدلی بر روی داده ها برازش داده نمی شود. در نتیجه:

- به منظور اجرای الگوریتم برای هر رکورد، نیاز به حضور تمامی داده های موجود در حافظه (Memory) می باشد.
- اجرای الگوریتم در داده های با ابعاد بالا (رکوردها و ویژگی ها زیاد) بسیار کند خواهد شد. به همین دلیل جزو دسته مدلهای تنبل (Lazy) قرار می گیرد.

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم شبکه عصبی

❑ الگوریتم شبکه عصبی مصنوعی (Artificial Neural Network - ANN)

الگوریتم شبکه عصبی مصنوعی، با الهام از سیستم عصبی زیستی، با ایجاد شبکه ای از واحدهای پردازنده به نام نرون و اتصال آنها از طریق یال ها، عملیات محاسبه و ساخت مدل تصمیم گیری را انجام می دهد.

هر یال، مانند سیناپس ها در شبکه عصبی زیستی، سیگنال های ورودی را به یک نرون منتقل می کند. نرون ها پردازش لازم روی سیگنال ورودی را انجام داده و خروجی را که یک عدد است مجددا از طریق یال ها به نرون دیگر انتقال می دهد تا در نهایت به ناحیه تصمیم برسد.

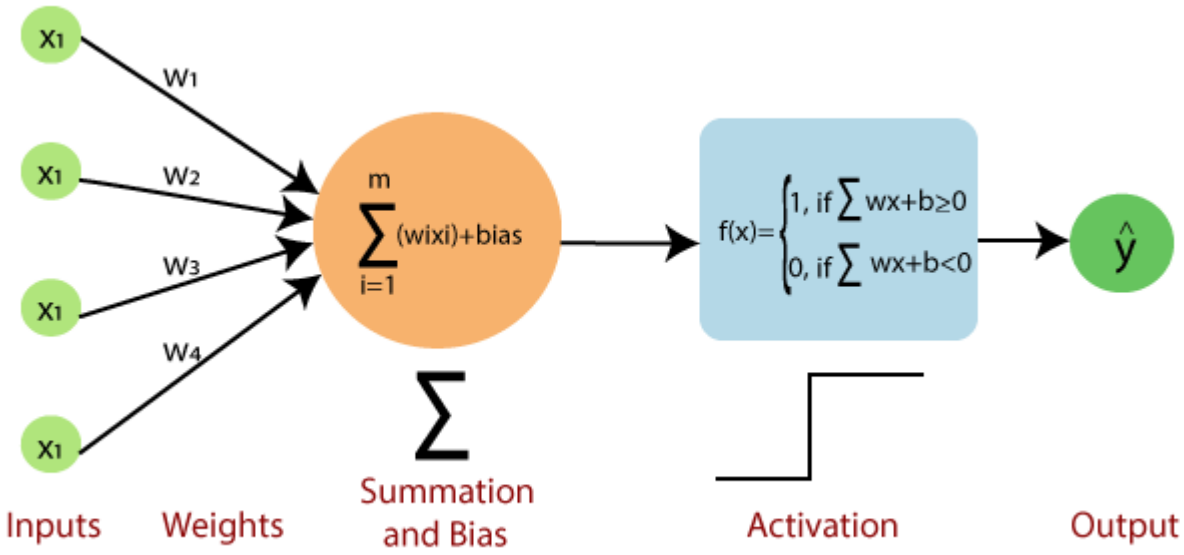
شبکه عصبی در فرآیند یادگیری با نظارت، قابلیت حل مسائل رگرسیون و رده بندی را ایجاد می نماید.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم شبکه عصبی

❑ پرسپترون (Perceptron)



پرسپترون، یک واحد از نرون های شبکه عصبی است که با دریافت مقادیر ورودی و انجام محاسبات برای حل مسائل رده بندی دودویی از طریق یادگیری با نظارت، در سال 1958 توسط فرانک روزنبلات معرفی شد. سیگنال های ورودی از طریق یال های وزن دهی شده وارد نرون شده و محاسبات پردازشی آن انجام می شود. می توان مقدار فیلد هدف را بصورت زیر نمایش داد:

$$y = \varphi\left(\sum_{i=1}^{n} w_i x_i + b\right) = \varphi(w^T x + b)$$

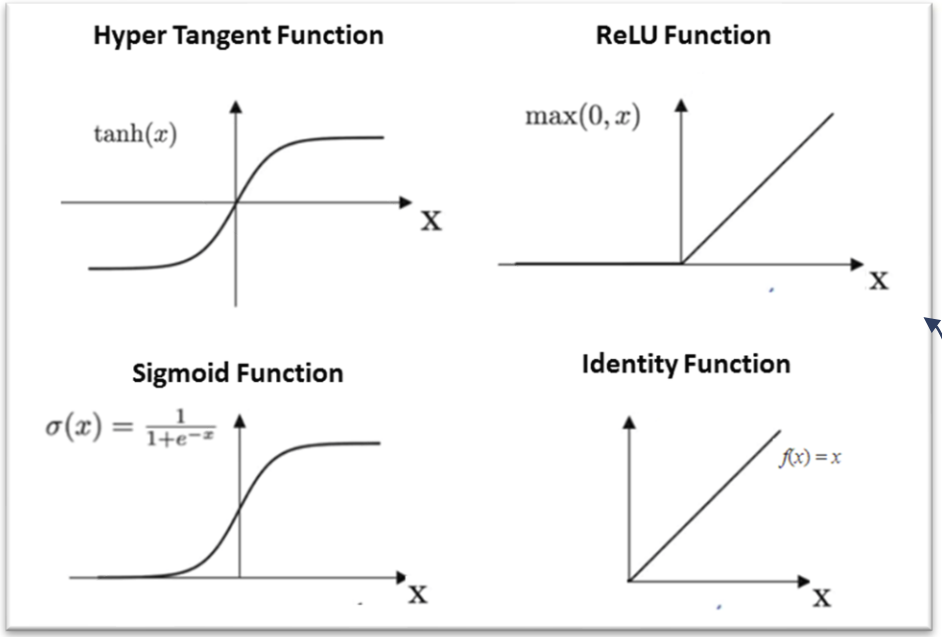
تابع φ به عنوان تابع فعالسازی Activation Function شناخته می شود. در واقع عملیات پردازشی یک پرسپترون، اعمال تابع فعالسازی Heaviside روی مجموع وزنی سیگنال های ورودی و ارائه خروجی صفر یا یک می باشد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم شبکه عصبی

❑ تابع فعال سازی (Activation Function)

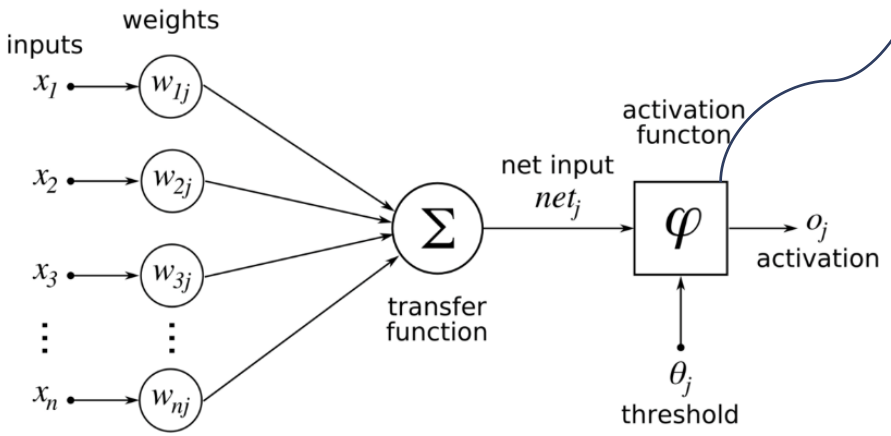


با توجه به نوع مسئله می توان از توابع فعالسازی متنوعی برای پرسپترون استفاده کرد.

در مسئله رده بندی، استفاده از تابع فعالسازی سیگموید (لجستیک) به عنوان یکی از توابع پرکاربرد در پرسپترون، مدل رگرسیون لجستیک را نتیجه می دهد.

همچنین در مسئله رگرسیون (پیش بینی) استفاده از تابع فعالسازی همانی در پرسپترون، منجر به مدل رگرسیون خطی می شود.

نکته مهم: استفاده از توابع فعال سازی خطی در شبکه عصبی، باعث می شود که شبکه عصبی فقط قادر به مدل سازی خطی باشد.

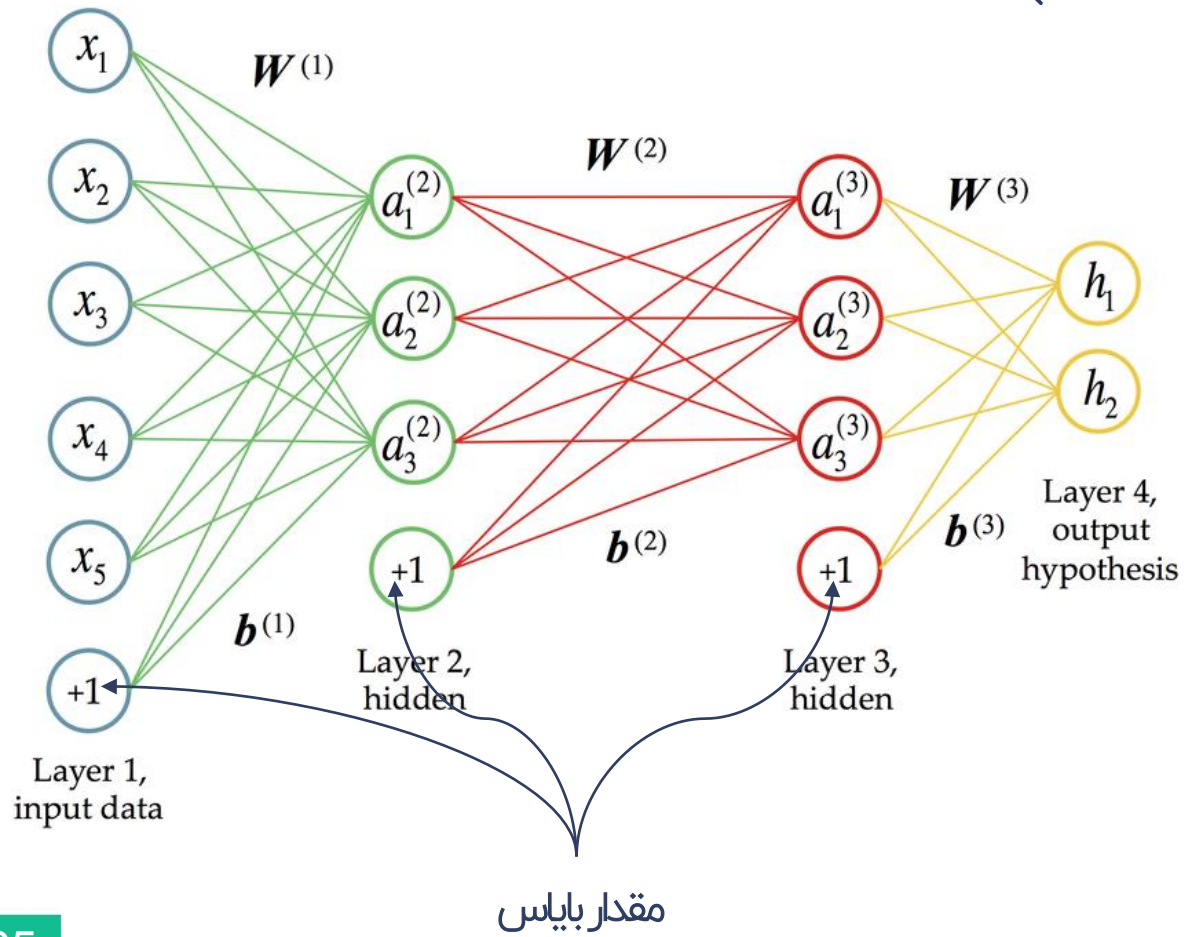


تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم شبکه عصبی

❑ پرسپترون چند لایه (Multilayer Perceptron-MLP)



الگوریتم شبکه عصبی پرسپترون چندلایه با هدف ایجاد قابلیت مدلسازی الگوهای غیرخطی و پیچیده توسعه داده شده است.

این توسعه، شامل اضافه کردن یک یا چند لایه پنهان به ساختار پرسپترون می باشد، بطوریکه در هر لایه یک یا چند نرون محاسباتی قرار داده شده است.

نحوه محاسبات در هر نرون، کاملا مشابه یک پرسپترون (با تابع فعالسازی غیر خطی) انجام می شود، بنابراین الگوریتم شبکه عصبی پرسپترون چند لایه، شامل تعداد زیادی مدل های مینیاتوری کوچک در ساختار خود می باشد که همدیگر را تغذیه می کنند و باعث می شود تا شبکه عصبی گوشه های مختلف از الگوهای غیرخطی در داده ها را شناسایی کند.

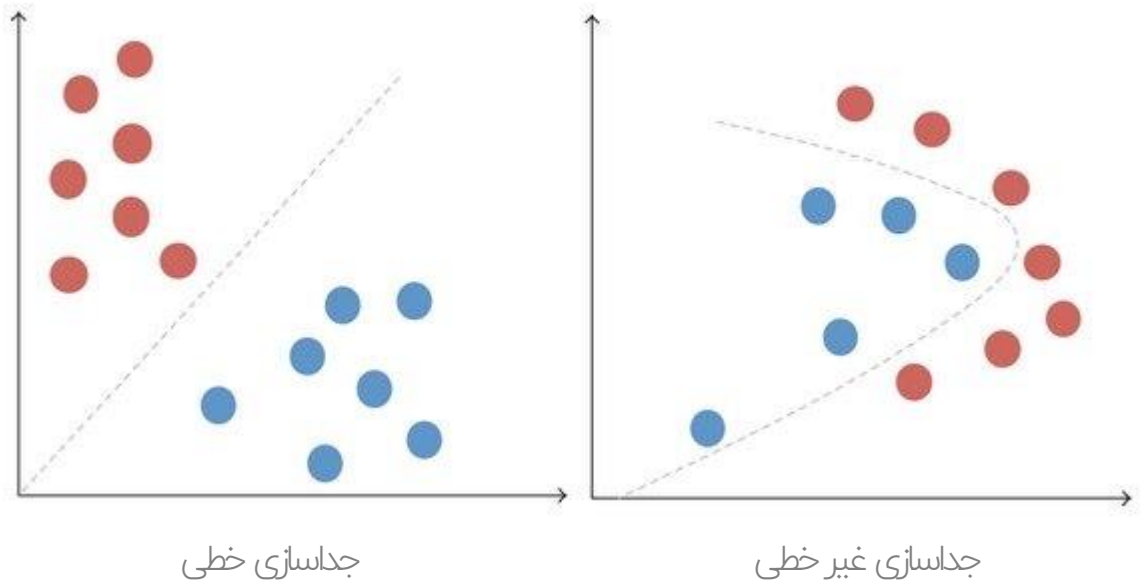
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ پرسپترون چند لایه (Multilayer Perceptron-MLP)

الگوریتم شبکه عصبی پرسپترون چندلایه با هدف ایجاد قابلیت مدلسازی الگوهای غیرخطی و پیچیده توسعه داده شده است.

این توسعه، شامل اضافه کردن یک یا چند لایه پنهان به ساختار پرسپترون می باشد، بطوریکه در هر لایه یک یا چند نرون محاسباتی قرار داده شده است.

نحوه محاسبات در هر نرون، کاملا مشابه یک پرسپترون (با تابع فعالسازی غیر خطی) انجام می شود، بنابراین الگوریتم شبکه عصبی پرسپترون چند لایه، شامل تعداد زیادی مدل های مینیاتوری کوچک در ساختار خود می باشد که همدیگر را تغذیه می کنند و باعث می شود تا شبکه عصبی گوشه های مختلف از الگوهای غیرخطی در داده ها را شناسایی کند.



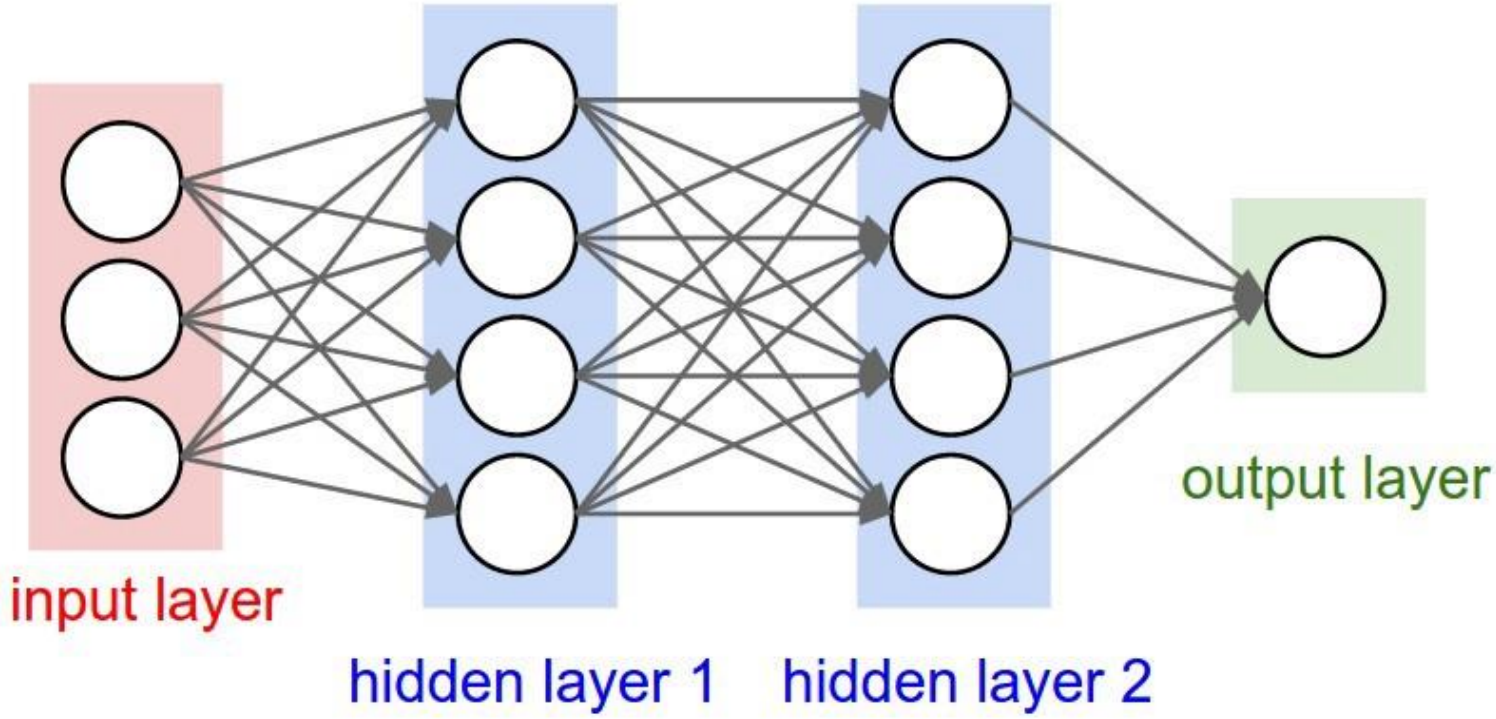
جداسازی غیر خطی　　جداسازی خطی

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم شبکه عصبی

- پرسپترون چند لایه (Multilayer Perceptron-MLP)

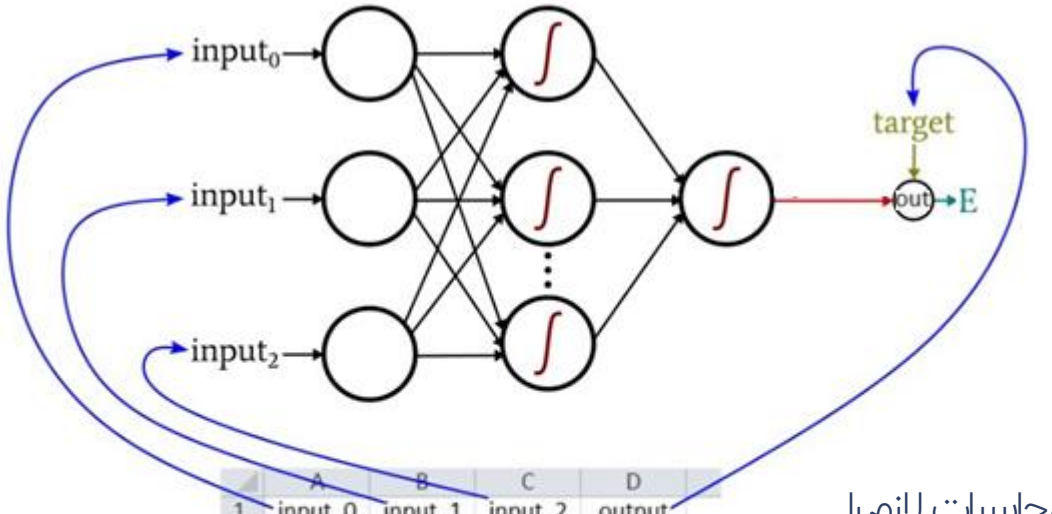


ویژگی مهم دیگری که در ساختار شبکه عصبی MLP وجود دارد، ماهیت پیش رونده بودن (Feed Forward) می باشد. این مفهوم به این معناست که جریان داده ها بصورت یک طرفه از سمت ورودی به سمت خروجی حرکت می کند.

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم شبکه عصبی

❑ پرسپترون چند لایه (Multilayer Perceptron-MLP)



	A	B	C	D
1	input_0	input_1	input_2	output
2	-4.5	4.5	-1	0
3	-4.5	-1.5	-5	0
4	4	4	-0.5	1
5	2.5	4	-2.5	1
6	-3	2.5	-5	0
7	5	-1.5	5	0

لایه ورودی

هریک از ویژگی های ورودی از طریق یک یال وزن دهی شده وارد شبکه عصبی می گردد. کلیه مقادیر ورودی قبل از ورود به شبکه، نیاز به نرمالسازی دارند که اغلب با روش Min-Max دامنه همه آنها در بازه (1و0) یا (1+، 1-) قرار می گیرد. کدگذاری فیلدهای کیفی نیز معمولا با روش One Hot انجام شده و برای هر کلاس از فیلد کیفی یک یال وزن دهی شده اختصاص داده می شود.

لایه های پنهان

هریک از نرون های لایه پنهان از تمام ورودی های لایه قبل مقدار گرفته و به عنوان یک پرسپترون محاسبات لازم را انجام داده و خروجی را از طریق یال های وزن دهی شده به تمامی نرون های لایه بعدی منتقل می کنند. در واقع در شبکه عصبی MLP هر نرون با تمامی نرون های قبل و بعد از خود در ارتباط مستقیم هست.

لایه خروجی

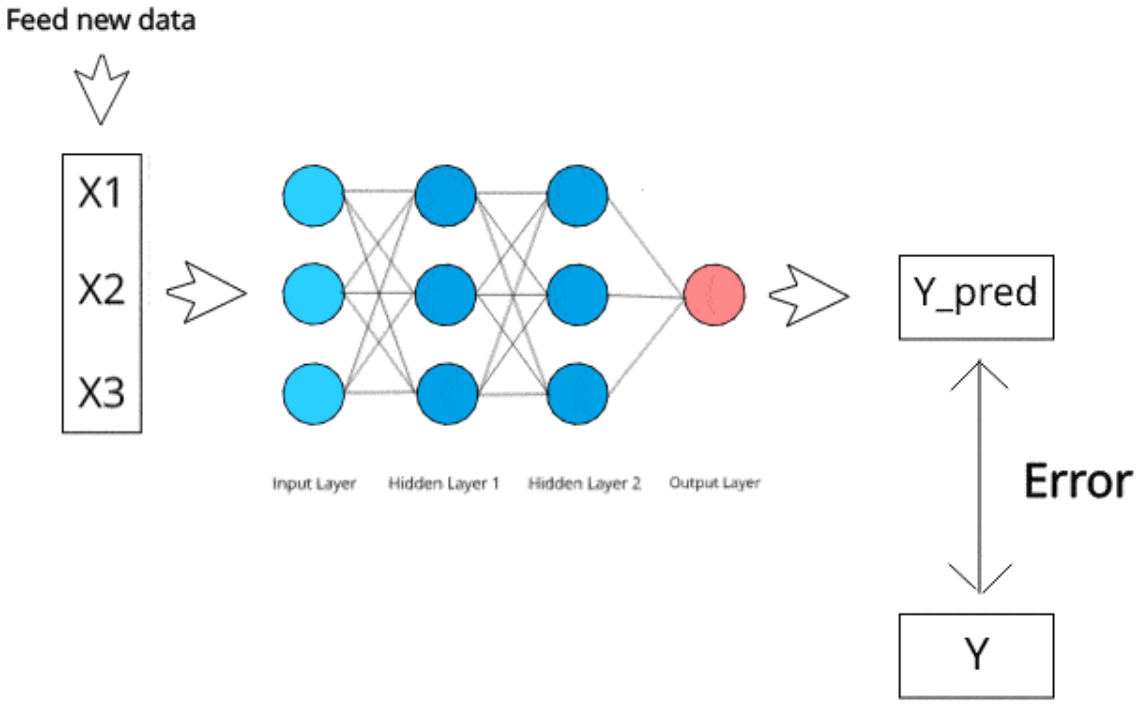
خروجی پرسپترون در این لایه با مقادیر فیلد هدف مقایسه شده و با اندازه گیری میزان خطا به اصلاح وزن یال ها پرداخته می شود.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم شبکه عصبی

❑ فرآیند یادگیری شبکه عصبی



در یادگیری مدل شبکه عصبی با تعریف تابع هزینه (زیان) مشابه مدل های رگرسیونی، به دنبال تعیین مقادیر وزن بهینه می باشیم بطوریکه تابع هزینه را مینیمم کند.

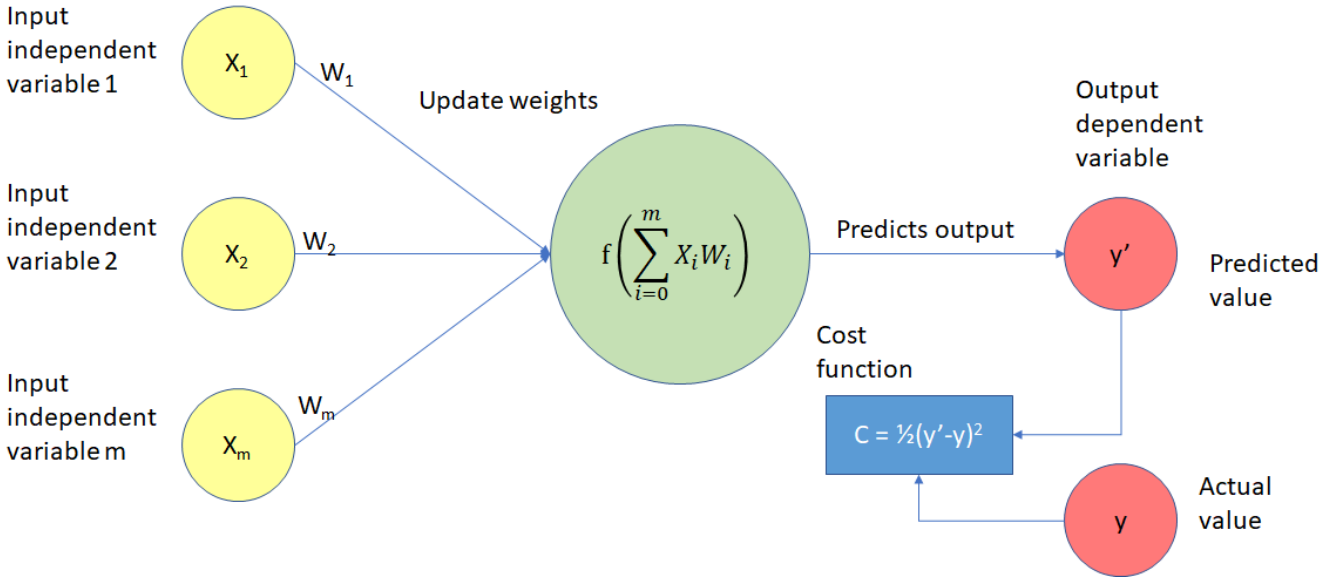
تفاوت مهم در شبکه عصبی MLP اینست که به علت وجود پرسپترون های متعدد (مدل های مینیاتوری کوچک) تغییرات جزئی در مقدار هر وزن، در مقادیر سایر پرسپترون ها نیز اثرگذار است و بهینه سازی تابع هزینه را پیچیده تر می کند. بهینه سازی تابع هزینه و برآورد وزن ها در این مدل بر اساس روش کاهش گرادیان (Gradient Decent) انجام می شود و جهت بروزرسانی وزن ها در لایه ها نیز از رویکرد پس انتشار (Backpropagation) استفاده می شود.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ فرآیند یادگیری شبکه عصبی

تابع هزینه (زیان) در شبکه عصبی

در مسئله رگرسیون، مشابه الگوریتم رگرسیون خطی از میانگین مجذور خطای مدل (MSE) استفاده می شود.



$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

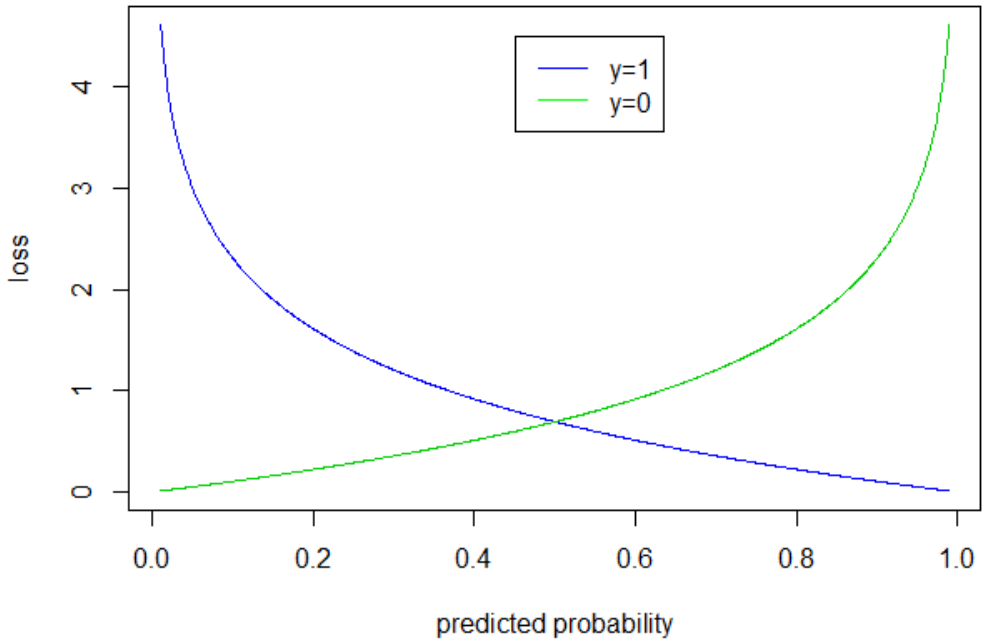
فرآیند داده کاوی

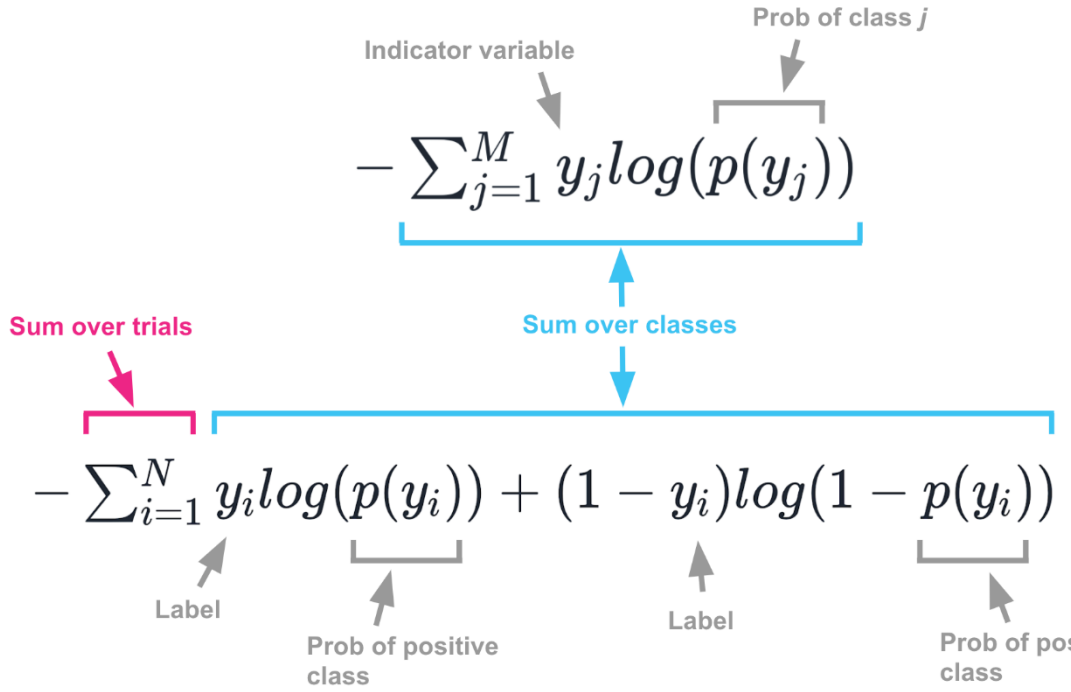
مدل های پیش بینانه – الگوریتم شبکه عصبی

❑ فرآیند یادگیری شبکه عصبی

تابع هزینه (زیان) در شبکه عصبی

در مسئله رده بندی، مشابه الگوریتم رگرسیون لجستیک از تابع کراس آنتروپی (Cross Entropy) استفاده می شود.



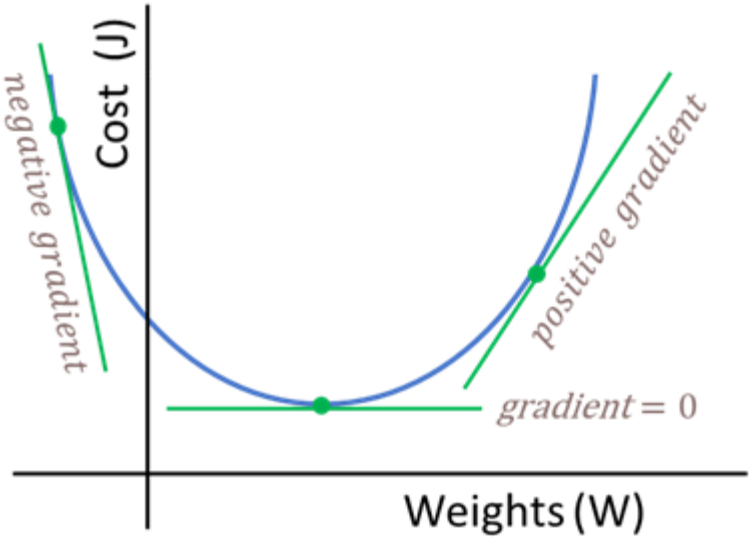


تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
گروه دایکه | dayche.com

❑ فرآیند یادگیری شبکه عصبی

- روش کاهش گرادیان (Gradient Decent Method)

این روش با در نظر گرفتن مقدار تصادفی اولیه برای هریک از ضرایب مدل (وزن ها) و محاسبه میزان خطای پیش بینی در جهت کاهش خطا حرکت کرده و طی یک فرآیند تکراری به ضرایب بهینه همگرا می شود. در واقع با تعیین گرادیان تابع هزینه که شامل برداری از مشتقات جزئی به تفکیک هر ضریب (وزن ها) است و حرکت در جهت بیشترین شیب، در یک فرآیند تکراری و گام به گام به سمت مینیمم تابع حرکت می کند.



$$Gradient\ L(w_1, w_2, \dots, w_n) = \left[\left[\frac{dL}{dw_1}\right], \left[\frac{dL}{dw_2}\right], \dots, \left[\frac{dL}{dw_n}\right]\right]$$

هر عنصر بردار گرادیان به ما نشان می دهد تغییر جزئی در هریک از مقادیر وزن ها به چه میزانی در جهت کاهش گرادیان تاثیر دارد. بنابراین می توان تعیین کرد کدام وزن و به چه مقداری تغییر نماید.

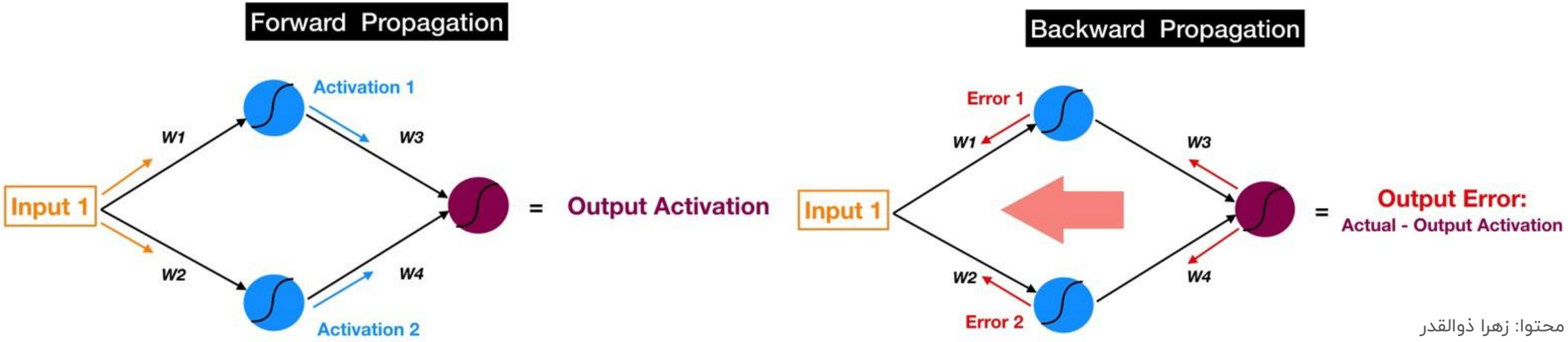
فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم شبکه عصبی

❑ فرآیند یادگیری شبکه عصبی

- روش پس انتشار (Backpropagation Method)

برخلاف جهت حرکت رو به جلوی شبکه عصبی MLP که مقادیر ورودی را لایه به لایه به سمت نرون های خروجی منتقل می کند، روش پس انتشار مقدار خطای محاسبه شده در نرون آخر را لایه به لایه به نرون های لایه ورودی منتقل می کند.



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ فرآیند یادگیری شبکه عصبی

- روش پس انتشار (Backpropagation Method)

الگوریتم یادگیری شبکه عصبی و اثر نرخ یادگیری

$Backpropagation:$

$D = \{x_i, y_i\}$

1. $Initialize\ W_{ij}^{k}\ ; where, i = Neuron\ from\ previous\ layer,$

   $j = Neuron\ from\ previous\ layer, k = next\ layer.$

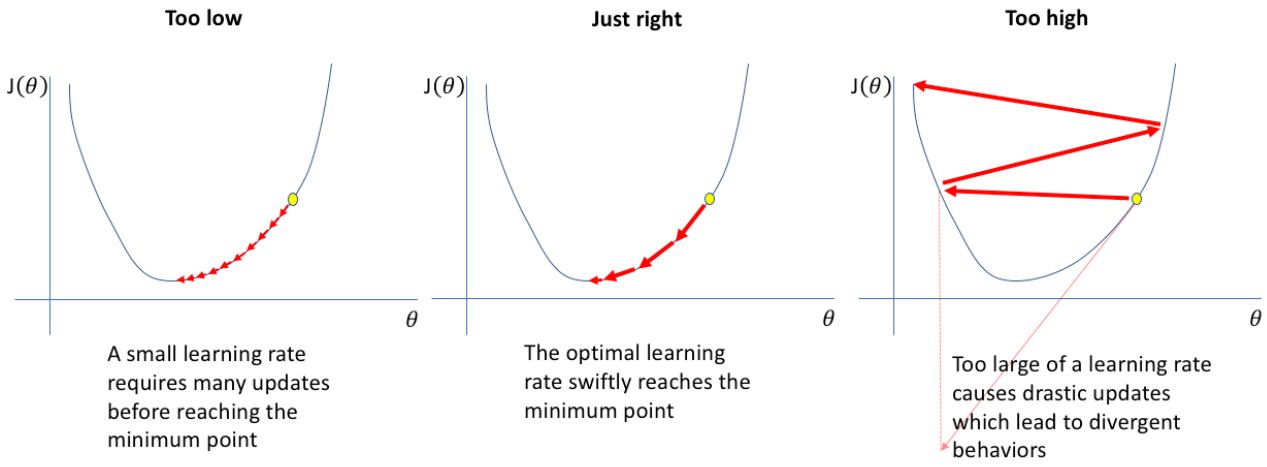
2. $for\ each\ x_i\ in\ D:$
   a. $Pass\ x_i\ forward\ through\ the\ network\ (forward\ pass)$
   b. $Compute\ the\ loss\ on\ y_i,\ L(y_i, \hat{y_i})$
   c. $Compute\ all\ the\ derivatives(\frac{\partial L}{\partial w})\ using\ chain\ rule$
   d. $Updates\ Weights\ from\ end\ of\ the\ network\ to\ the\ start$

$$(W_{ij}^{k})_{new} = (W_{ij}^{k})_{old} - \eta * \left(\frac{\partial L}{\partial W_{ij}^{k}}\right)_{(W_{ij}^{k})_{old}}$$

$where, \eta = learning\ rate$

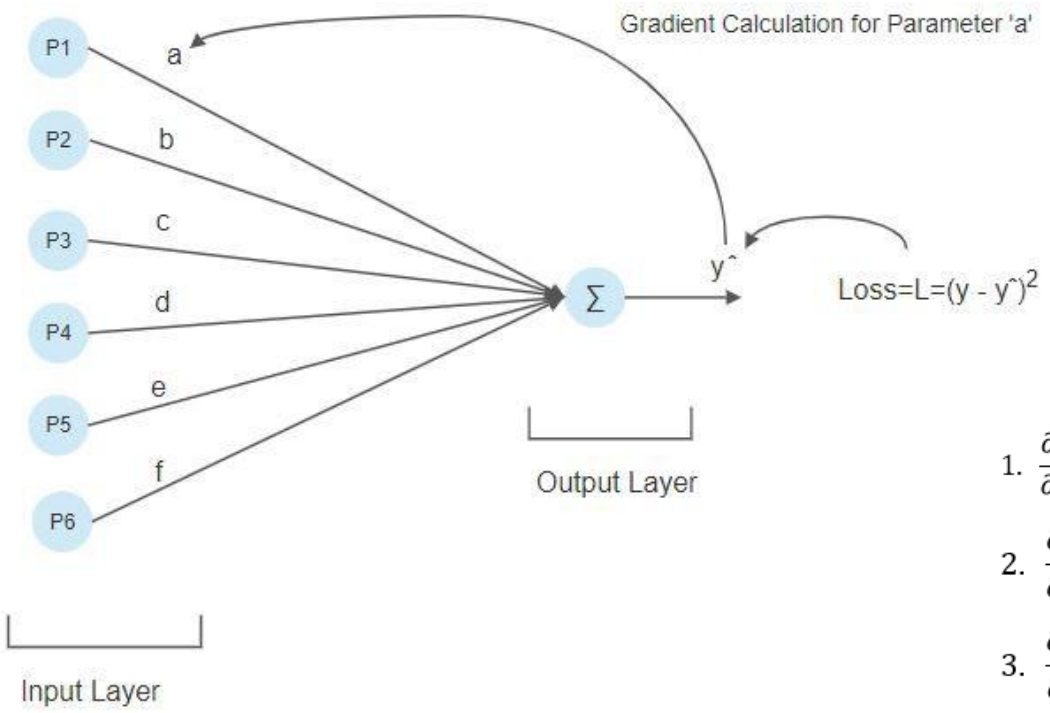
3. $Repeat\ step\ 2\ until\ convergence$

$$(W_{ij}^{k})_{new} \approx (W_{ij}^{k})_{old}$$



فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم شبکه عصبی



❑ فرآیند یادگیری شبکه عصبی

- روش پس انتشار (Backpropagation Method)

مثال: محاسبه روش پس انتشار در یک پرسپترون

$$L = (y - \hat{y})^2$$

$$\hat{y} = P_1 * a + P_2 * b + P_3 * c + P_4 * d + P_5 * e + P_6 * f$$

$$Weight\ matrix = W^* = [a \quad b \quad c \quad d \quad e \quad f]$$

$$Input\ matrix = X = [P_1 \quad P_2 \quad P_3 \quad P_4 \quad P_5 \quad P_6]$$

1. $\frac{\partial L}{\partial a} = \frac{\partial L}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial a} = -2(y - \hat{y}) * P_1$
2. $\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial b} = -2(y - \hat{y}) * P_2$
3. $\frac{\partial L}{\partial c} = \frac{\partial L}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial c} = -2(y - \hat{y}) * P_3$
4. $\frac{\partial L}{\partial d} = \frac{\partial L}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial d} = -2(y - \hat{y}) * P_4$
5. $\frac{\partial L}{\partial e} = \frac{\partial L}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial e} = -2(y - \hat{y}) * P_5$
6. $\frac{\partial L}{\partial f} = \frac{\partial L}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial f} = -2(y - \hat{y}) * P_6$

$$Gradient\ matrix = dW^* = \left[\frac{\partial L}{\partial a} \quad \frac{\partial L}{\partial b} \quad \frac{\partial L}{\partial c} \quad \frac{\partial L}{\partial d} \quad \frac{\partial L}{\partial e} \quad \frac{\partial L}{\partial f}\right]$$

$$(W^*)_{new} = (W^*)_{old} - \eta * dW^*$$

بروزرسانی مقادیر وزن — نرخ یادگیری

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ فرآیند یادگیری شبکه عصبی

- تنظیم پارامترهای مدل شبکه عصبی

تعداد لایه پنهان: هیچ قانون و رابطه ای برای تعیین تعداد لایه های مناسب شبکه عصبی وجود ندارد؛ ولی می توان ثابت کرد اغلب مسائل غیر خطی در رگرسیون و رده بندی با یک لایه پنهان قابلیت مدلسازی دارند و همچنین با دو لایه پنهان می توان هر شکل و فرمی از روابط غیر خطی را مدل کرد.

تعداد نرون های پنهان: تعیین دقیق تعداد نرون ها در لایه های پنهان نیز، هیچ راهکار محاسباتی ندارد و برای هر مسئله بایستی بر اساس آزمون و خطا و ارزیابی نتایج مدل با تنظیم پارامترهای مختلف بدست آید.

در برخی از منابع، قواعدی برای تنظیم اولیه شبکه عصبی معرفی شده و بر اساس ارزیابی نتایج اولیه، می توان به کم یا اضافه کردن نرون ها و لایه های شبکه عصبی پرداخت:

- تعداد نرون های لایه پنهان بایستی در بازه تعداد نرون های خروجی و ورودی باشد.
- تعداد نرون های لایه پنهان میانگین تعداد نرون های لایه ورودی و خروجی باشد.
- تعداد نرون ها در لایه های پنهان طوری تعیین گردد که ساختار هرمی شکل شبکه عصبی از ورودی به خروجی حفظ شود.

❑ فرآیند یادگیری شبکه عصبی

- تنظیم پارامترهای مدل شبکه عصبی

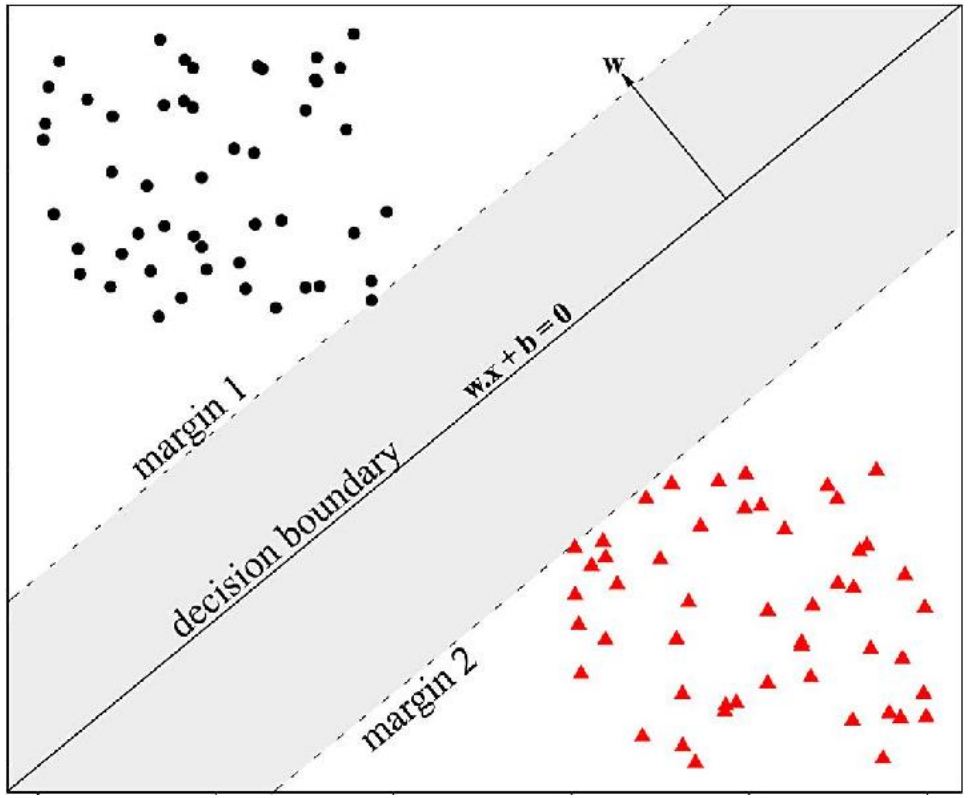
دوره (Epoch): در یادگیری مدل های شبکه عصبی، مجموعه داده های آموزشی بیش از یک بار در فرآیند یادگیری وارد شده و در تعیین وزن های مدل نقش بازی می کنند. به هر بار تکرار مجموعه داده یک چرخه یا دوره (Cycle or Epoch) گفته می شود افزایش این مقدار باعث می شود، هر رکورد در مجموعه داده های آموزشی، بارها در فرآیند یادگیری مدل وارد شده و منجر به بهبود نتایج مدل شود.

اندازه بسته (Batch Size): در یادگیری مدل های شبکه عصبی، می توان با تقسیم مجموعه داده های آموزشی به تعداد برابر از بسته هایی در اندازه حداقل یک رکورد تا حداکثر به تعداد کل داده های آموزشی، فرآیند بروزرسانی وزن های مدل در روش کاهش گرادیان را انجام داد. به این معنی که پس از اندازه گیری خطای پیش بینی در هر بسته از داده های آموزشی، محاسبات کاهش گرادیان انجام شده و وزنهای مدل با روش پس انتشار اصلاح می شوند.

مثال: فرض کنید 100 رکورد آموزشی داریم و تنظیمات مدل شبکه عصبی با اندازه بسته 5 و تعداد 1000 دوره انجام شده است. بنابراین در دوره اول از ورودی داده ها به شبکه عصبی، وزن های مدل به تعداد 20 مرتبه اصلاح می شوند و در پایان فرآیند یادگیری مدل، وزن های نهایی مدل پس از 20 هزار مرتبه بروزرسانی بدست می آیند.

❑ الگوریتم ماشین بردار پشتیبان (Support Vector Machine – SVM)

- مقدمه ای بر الگوریتم SVM



این الگوریتم با رویکرد یادگیری با نظارت در حل مسائل رده بندی و رگرسیون مورد استفاده قرار می گیرد. البته بیشترین موارد استفاده آن در مسائل رده بندی می باشد و برای اولین بار در سال 1995 توسط Vapnik برای مسئله رده بندی دودویی معرفی شد.

توسعه این الگوریتم برای حل مسائل چند کلاسه، رگرسیون و حتی خوشه بندی (یادگیری بدون نظارت) نیز انجام شده است:

SVR: Support Vector Regression

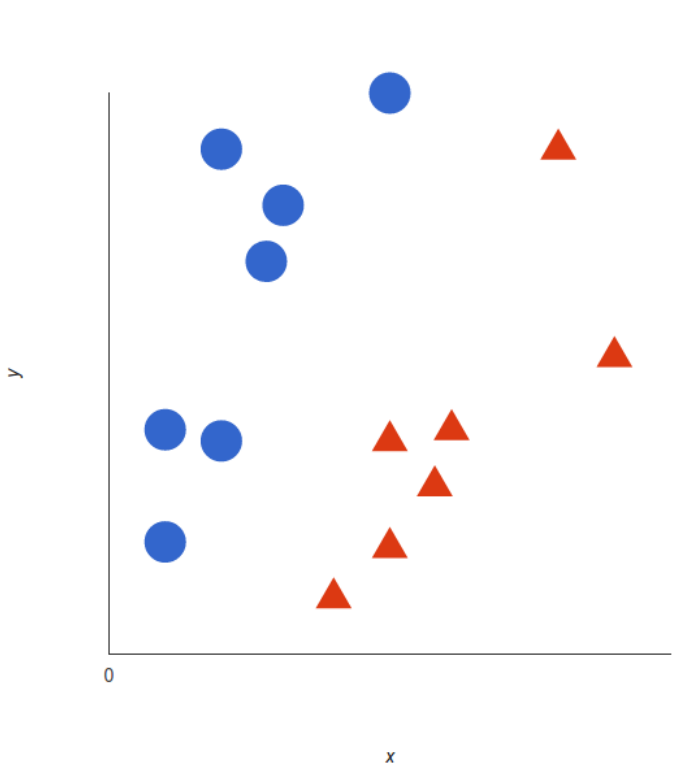
SVC: Support Vector Clustering

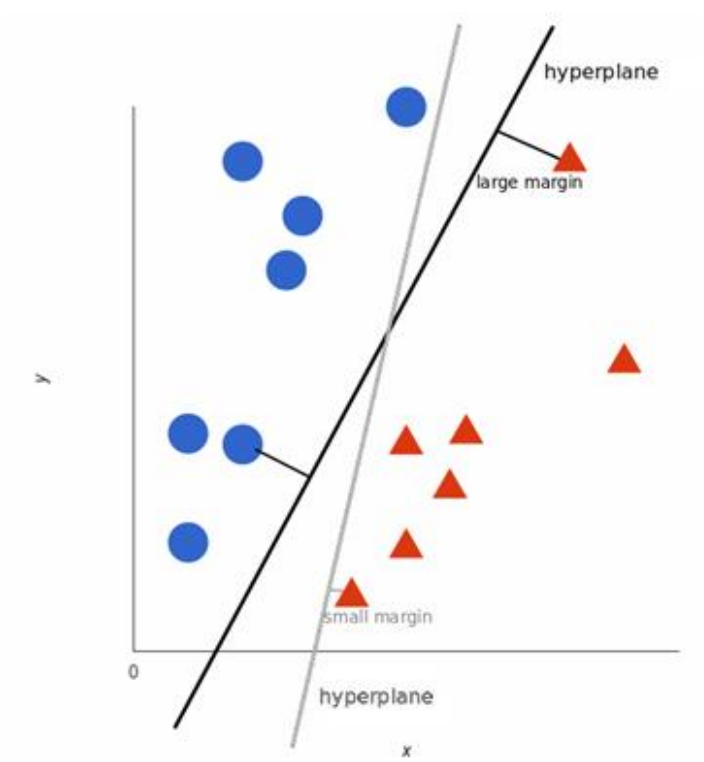
ایده اصلی و متمایز این الگوریتم نسبت به سایر الگوریتم ها، یافتن خط جداکننده و رده بند به شکلی است که حداکثر فاصله از رکوردهای کلاس های هدف را داشته باشد.

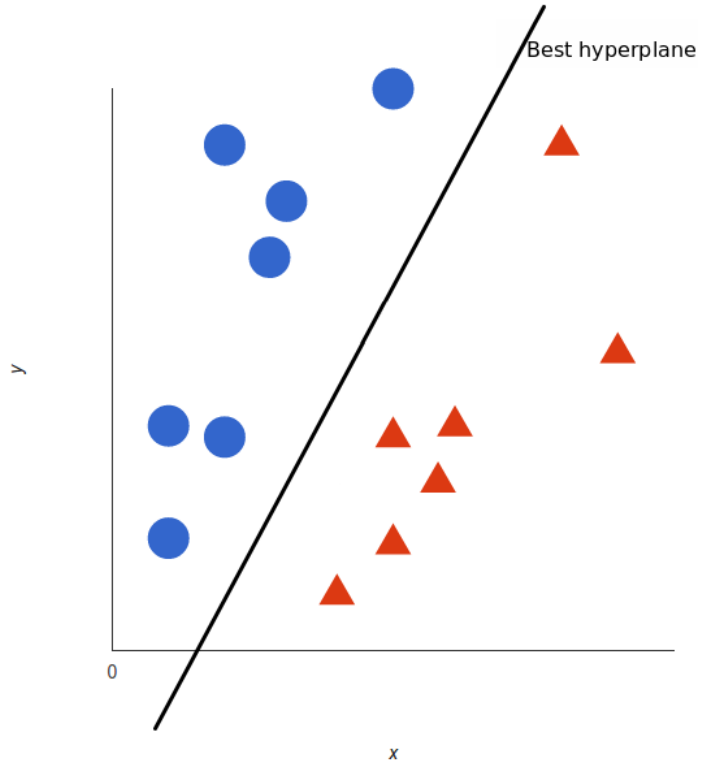
فرآیند داده کاوی

مدل های پیش بینانه – الگوریتم ماشین بردار پشتیبان

- الگوریتم ماشین بردار پشتیبان (Support Vector Machine – SVM)
  - مقدمه ای بر الگوریتم SVM

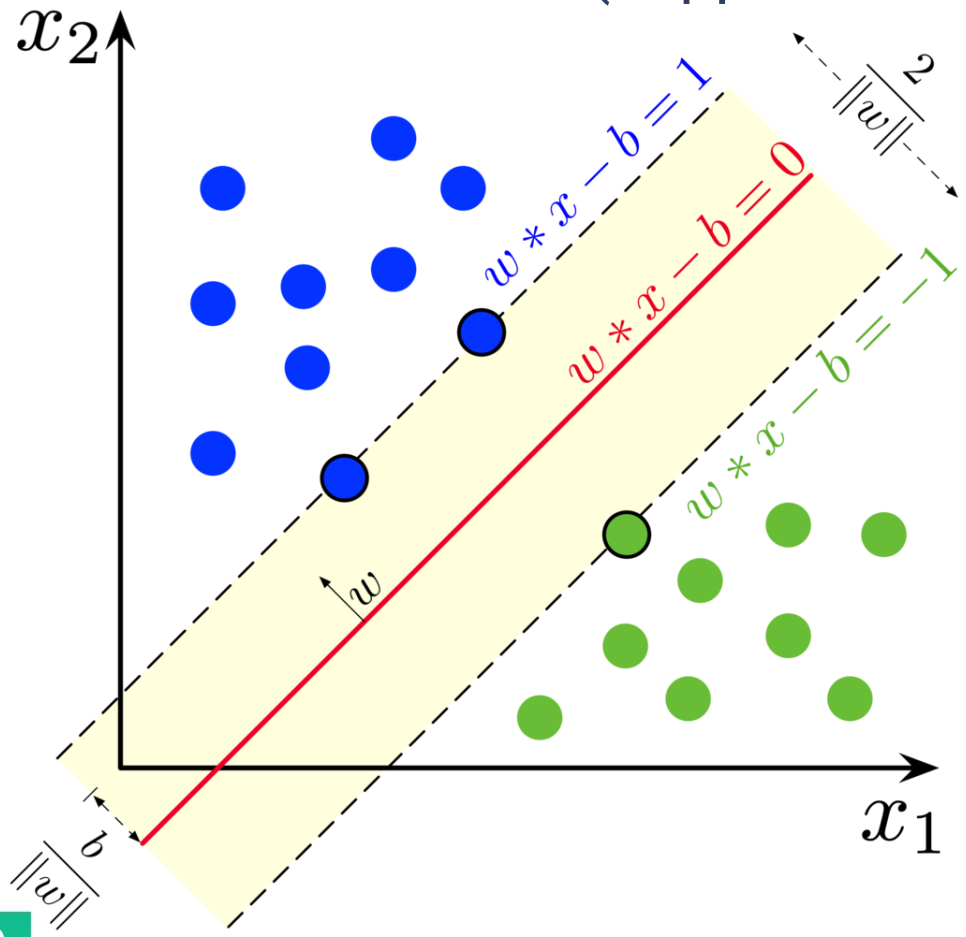






تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ الگوریتم ماشین بردار پشتیبان (Support Vector Machine – SVM)



- مقدمه ای بر الگوریتم SVM

فرض کنید مجموعه داده ها شامل بردارهای $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$ باشد بطوریکه بردار p – بعدی X ویژگی های ورودی مدل و Y فیلد هدف و دارای دو مقدار 1+ و 1- می باشد.

می توان معادله خط (ابرصفحه) جداکننده را به فرم زیر نمایش داد:

$$w^T * x_i - b = 0$$

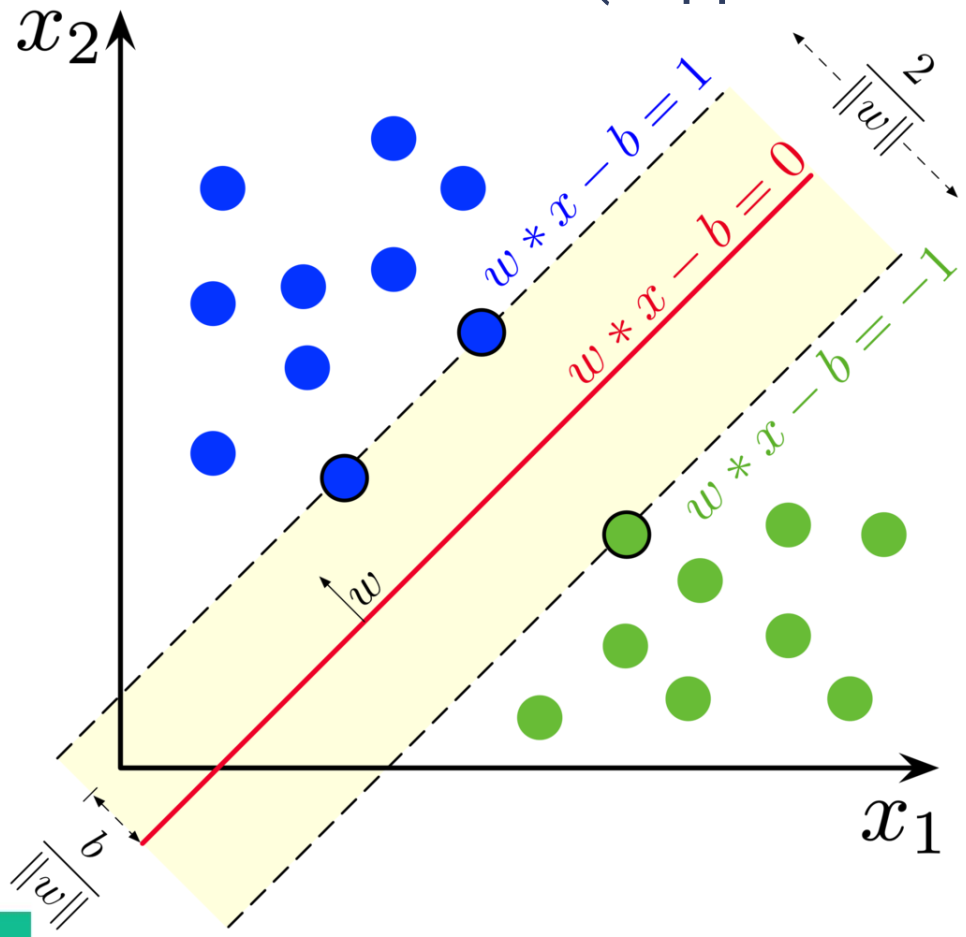
این خط (ابرصفحه) مرز تصمیم گیری برای تعلق داده های ورودی به کلاس های فیلد هدف است؛ در الگوریتم SVM به دنبال کاهش ریسک تصمیم گیری از طریق ایجاد بیشترین فاصله خط مرزی با نقاط مرزی کلاس های مجاور می باشد. بنابراین ناحیه تصمیم گیری به فرم زیر نمایش داده می شود، بطوریکه فاصله بین دو خط (ابرصفحه) ماکسیمم گردد:

$$w^T * x_i - b \geq 1, \quad \text{if } y_i = 1$$

$$w^T * x_i - b \leq -1, \quad \text{if } y_i = -1$$

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ الگوریتم ماشین بردار پشتیبان (Support Vector Machine – SVM)



- مقدمه ای بر الگوریتم SVM

با ضرب مقادیر Y در معادله خطوط مرزی، می توان در قالب یک معادله به شکل زیر نمایش داد:

$$y_i(w^T * x_i - b) \geq 1, \qquad for\, 1 \leq i \leq n$$

بر اساس مفهوم بردار نرمال w در هندسه تحلیلی ثابت می شود، فاصله بین دو خط مرزی که به عنوان حاشیه (Margin) شناخته می شود، برابر با 2 برابر معکوس نرم دوم w می باشد.

با توجه به هدف ماکسیمم کردن این فاصله می توان مسئله را به صورت تابع بهینه سازی زیر در نظر گرفت:

$$Min\ \|w\| \quad or \quad Min\ \frac{1}{2}\|w\|^2 \quad or \quad Min\ \frac{1}{2}w^T w$$

$$subject\ to \quad y_i * \mathrm{f}(x_i) \geq 1, \qquad if \quad f(x_i) = w^T x_i - b$$

پس از حل مسئله و برآورد مقادیر w و b، مشاهداتی از X که روی خطوط مرزی قرار می گیرند و یا بسیار نزدیک به آنها هستند، بردارهای پشتیبان می گویند.

تولید محتوا: زهرا ذوالقدر

daychegroup

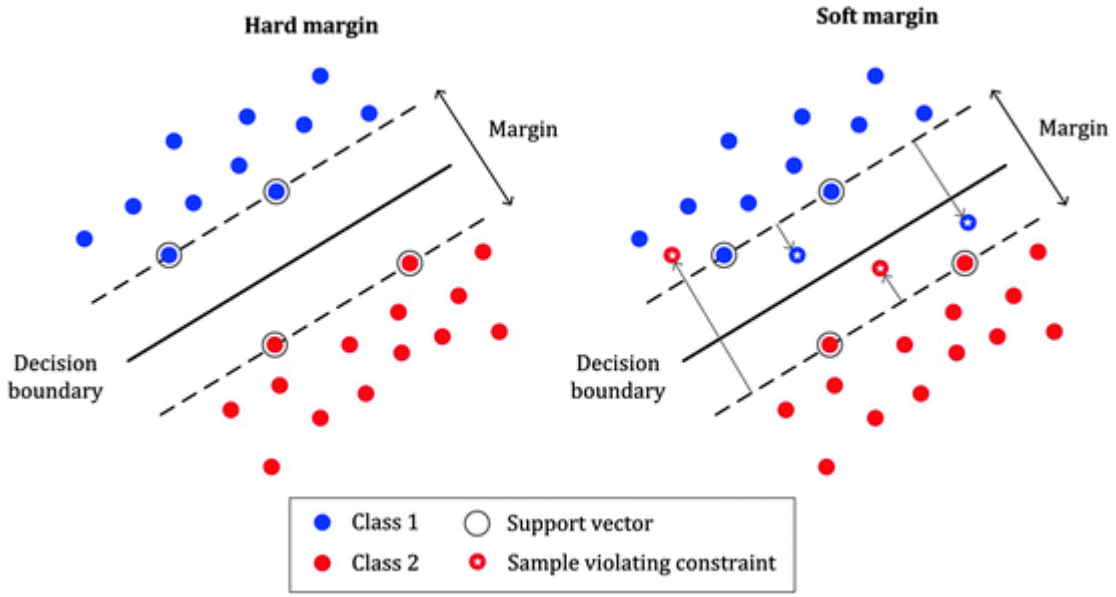
daychegroup

dayche.com | گروه دایکه

❑ الگوریتم ماشین بردار پشتیبان (Support Vector Machine – SVM)

- حاشیه سخت در مقابل حاشیه نرم

در داده های واقعی معمولا به سختی می توان با استفاده از یک خط جداکننده با حاشیه های سخت گیرانه، فاصله های مرزی مطمئنی ایجاد نمود. بنابراین با پذیرفتن مقداری خطا در تصمیم گیری می توان حاشیه های نرم تری برای رده بندی ایجاد نمود:



$$w.x_i+b \geq 1-\xi_i \qquad for\ y_i=+1$$

$$w.x_i+b \leq -1+\xi_i \qquad for\ y_i=-1$$

$$combining\ above\ two\ equation\ ,it\ can\ be\ written\ as$$

$$y_i(w.x_i+b)-1+\xi_i \geq 0 \qquad for\ y_i=+1,-1$$

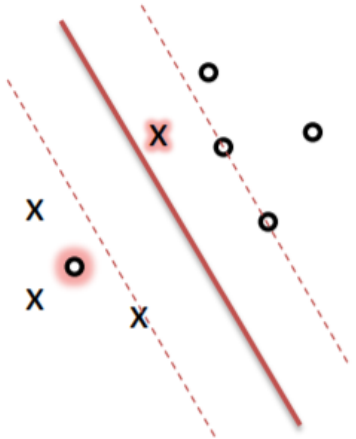
$$\xi_i \geq 0 \qquad for\ i=1..l$$

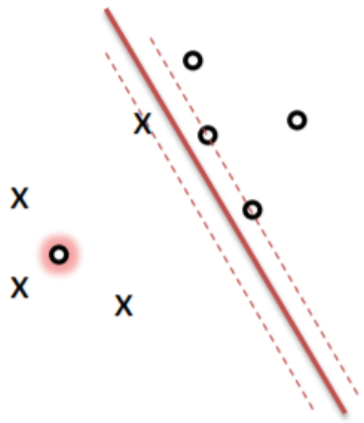
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ الگوریتم ماشین بردار پشتیبان (Support Vector Machine – SVM)

- حاشیه سخت در مقابل حاشیه نرم

در داده های واقعی معمولا به سختی می توان با استفاده از یک خط جداکننده با حاشیه های سخت گیرانه، فاصله های مرزی مطمئنی ایجاد نمود. بنابراین با پذیرفتن مقداری خطا در تصمیم گیری می توان حاشیه های نرم تری برای رده بندی ایجاد نمود:





پارامتر تنظیم

$$\min_{\boldsymbol{w},w_0,\{\xi_i\}_{i=1}^N} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{s.t.} \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right) \geq 1 - \xi_i \quad i = 1, \dots, N$$

$$\xi_i \geq 0$$

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

❑ الگوریتم ماشین بردار پشتیبان (Support Vector Machine – SVM)

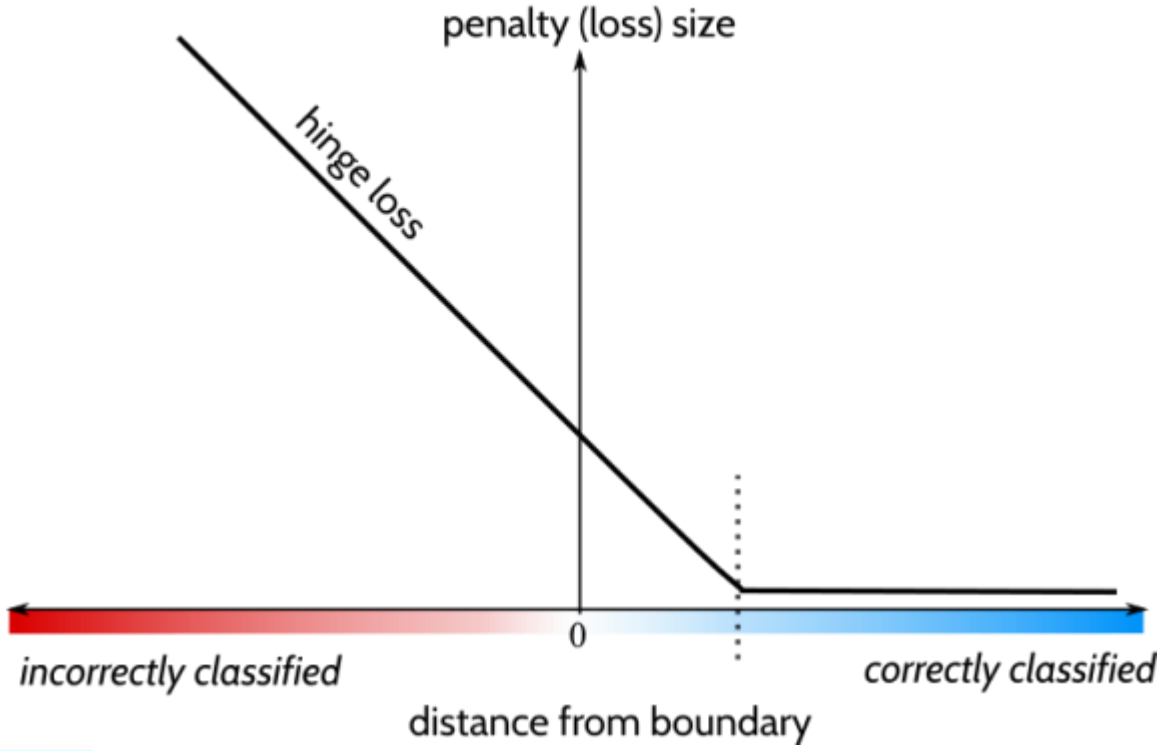
- **تابع هزینه (زیان) الگوریتم SVM (Hinge Loss)**

برای تعریف تابع هزینه در الگوریتم SVM، برای خطاهای پیش بینی شده طبق رابطه زیر مقدار هزینه (زیان) تعریف می شود:

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases}$$

$$\rightarrow \quad Hinge\ Loss = \sum_{i=1}^{n} \max(0, 1 - y_i * f(x_i))$$

طبق تعریف تابع هزینه الگوریتم SVM، مقدار هزینه برای خطای پیش بینی، به میزان دور شدن از ناحیه تصمیم مدل، به صورت خطی افزایش می یابد.

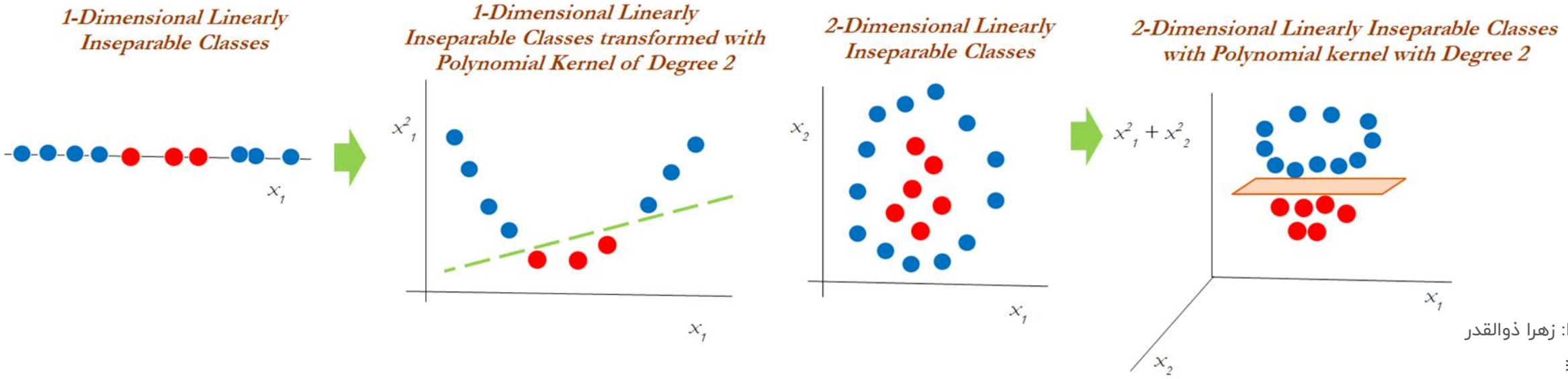


تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ الگوریتم ماشین بردار پشتیبان (Support Vector Machine – SVM)

- رده بندی الگوهای غیرخطی با استفاده از Kernel

در صورتی که الگوهای موجود در داده ها قابلیت جداسازی بصورت خطی نداشته باشد، با استفاده از توابع تبدیل کرنل جهت نگاشت داده های غیرخطی به فضای ویژگی های جدید به طوری که قابلیت تفکیک خطی داشته باشد، انجام می شود.



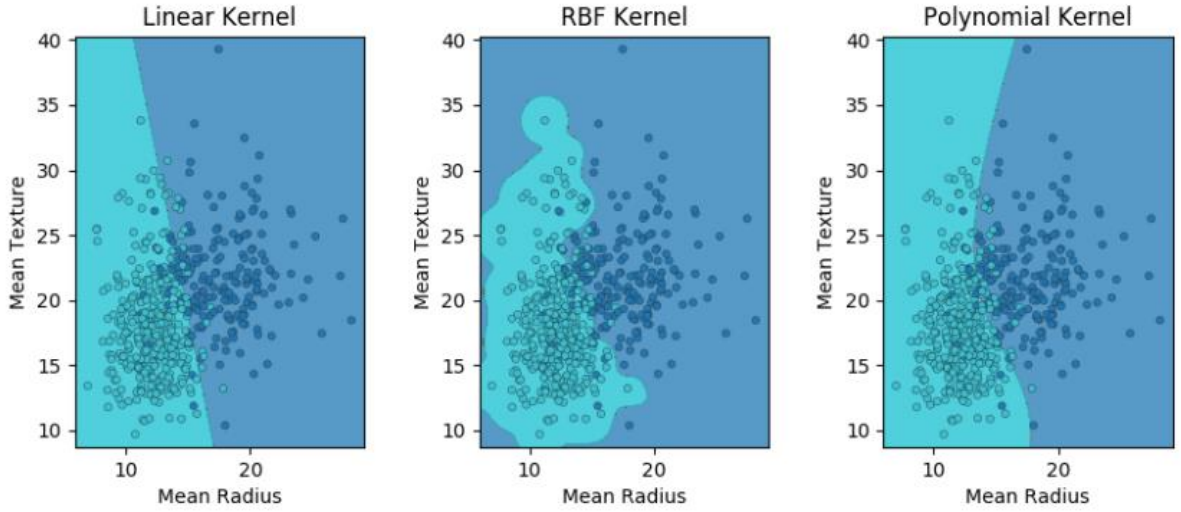
تولید محتوا: زهرا ذوالقدر
group
daychegroup
dayche.com | گروه دایکه

❑ الگوریتم ماشین بردار پشتیبان (Support Vector Machine – SVM)

- رده بندی الگوهای غیرخطی با استفاده از Kernel

در صورتی که الگوهای موجود در داده ها قابلیت جداسازی بصورت خطی نداشته باشد، با استفاده از توابع تبدیل کرنل جهت نگاشت داده های غیرخطی به فضای ویژگی های جدید به طوری که قابلیت تفکیک خطی داشته باشد، انجام می شود. انتخاب تابع کرنل، با بکارگیری توابع مختلف و مقایسه شاخص های ارزیابی انجام می شود.

بنابراین نمی توان بر اساس روش های محاسباتی یا قواعد از پیش تعیین شده، تابع کرنل خاصی را برای یک مجموعه داده یا مسئله تعیین نمود.



Kernels	Formula
linear	$k(x, y) = x.y$
sigmoid	$k(x, y) = \tanh(ax.y + b)$
polynomial	$k(x, y) = (1 + x.y)^d$
RBF	$k(x, y) = \exp(-a\|x - y\|^2)$
exponential RBF	$k(x, y) = \exp(-a\|x - y\|)$

❑ الگوریتم ماشین بردار پشتیبان (Support Vector Machine – SVM)

○ بکارگیری روش تنظیم سازی (Regularization term in SVM)

الگوریتم SVM معمولا در مدلسازی داده های با ابعاد بالا و همچنین مجموعه داده هایی که تعداد ویژگی ها بیشتر از رکوردهای آموزشی است، عملکرد خوبی نسبت به سایر الگوریتم ها دارد. بنابراین استفاده از روش های تنظیم سازی برای برآورد ضرایب مدل که منجر به سادگی بیشتر آن می شود، در چنین مسائلی یکی از روش های پر کاربرد می باشد.

$$L1\ Regularization\ Loss = C\sum_{i=1}^{n}\max\left(0,1-y_i*\mathrm{f}(x_i)\right)+\lambda\|W\|_1$$

$$L2\ Regularization\ Loss = C\sum_{i=1}^{n}\max\left(0,1-y_i*\mathrm{f}(x_i)\right)+\lambda\|W\|_2$$

❑ الگوریتم ماشین بردار پشتیبان (Support Vector Machine – SVM)

○ بکارگیری الگوریتم SVM در حل مسائل چند کلاسه

الگوریتم های رده بند دودویی برای حل مسائل چند کلاسه، معمولا از دو رویکرد استفاده می کنند:

○ یک در مقابل بقیه (One vs Rest – OvR)

در این حالت مدل رده بند چند کلاسه به تعداد کلاس های فیلد هدف ساخته می شود و هر بار یکی از کلاس ها در مقابل بقیه کلاس ها بصورت مدل دودویی آموزش داده می شود. در نهایت کلاسی که با بیشترین احتمال عضویت بدست آید به عنوان خروجی پیش بینی در نظر گرفته می شود.

این رویکرد عموما در الگوریتم هایی که مقدار احتمال تعلق به یک کلاس را بر می گردانند، مانند الگوریتم های لجستیک یا پرسپترون، مورد استفاده قرار می گیرد و در مواقعی که تعداد رکوردهای مجموعه داده بسیار زیاد باشد، می تواند روش کندی باشد.

❑ الگوریتم ماشین بردار پشتیبان (Support Vector Machine – SVM)

- **بکارگیری الگوریتم SVM در حل مسائل چند کلاسه**

الگوریتم های رده بند دودویی برای حل مسائل چند کلاسه، معمولا از دو رویکرد استفاده می کنند:

- **یک در مقابل یک (One vs One – OvO)**

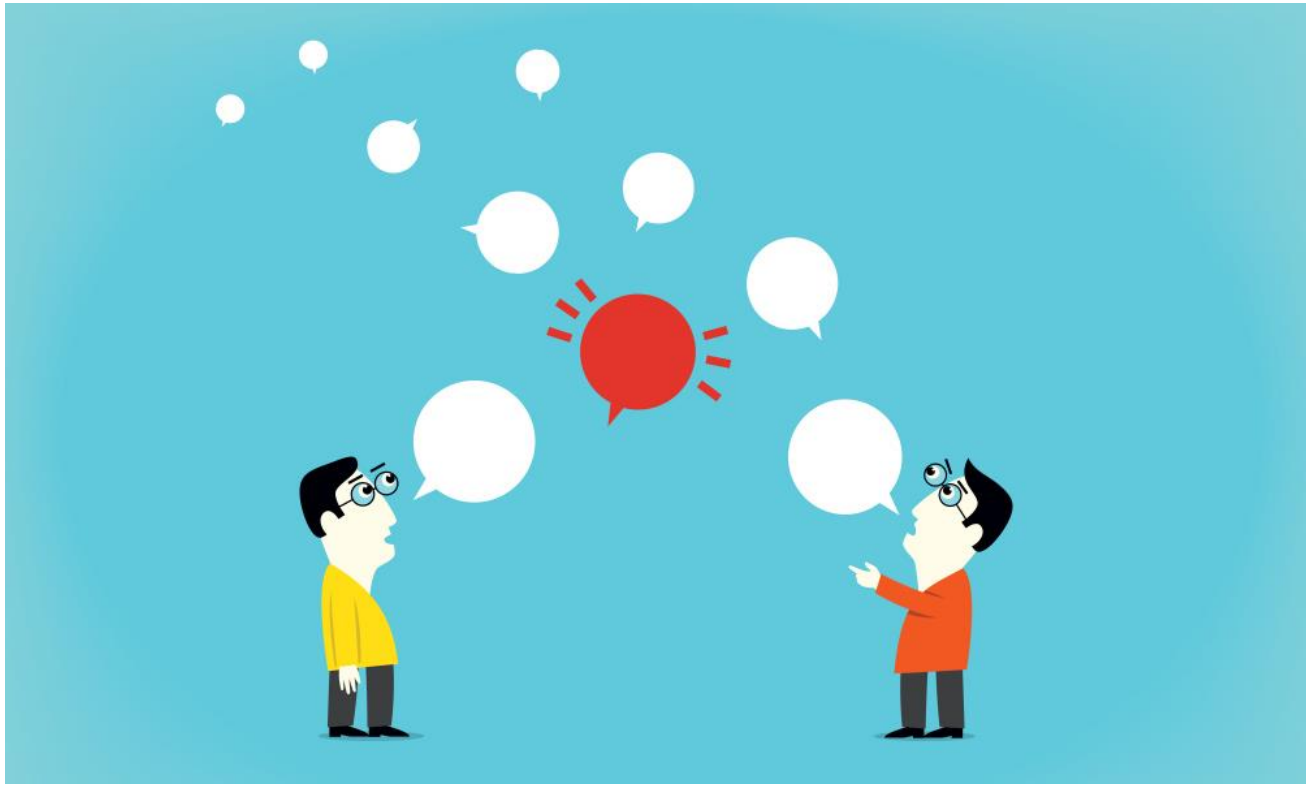
در این حالت به تعداد حالت های ممکن تقابل بین کلاس ها، مدل دودویی ساخته می شود و به ازای هر مدل، یک کلاس خروجی یا احتمال تعلق به آن محاسبه می شود. در نهایت با رای گیری روی کلاس های پیش بینی شده توسط تمام مدل هایی دودویی و یا جمع احتمال تعلق هر کلاس در تمامی مدل ها و انتخاب کلاس نهایی بر اساس مقدار ماکسیمم احتمال تجمیعی آن، پیش بینی انجام می شود.

در این حالت تعداد مدل های ساخته شده بیشتر از رویکرد قبلی می باشد، اما به علت انتخاب زیر مجموعه هایی از داده های آموزشی که منجر به کاهش رکوردها می شود، در الگوریتم هایی مانند SVM و یا سایر الگوریتم های مبتنی بر کرنل، بیشتر مورد استفاده قرار می گیرد.

فرآیند داده کاوی

مدل های پیش بینانه – یادگیری گروهی

❑ یادگیری گروهی در مدلسازی (Ensemble Learning)



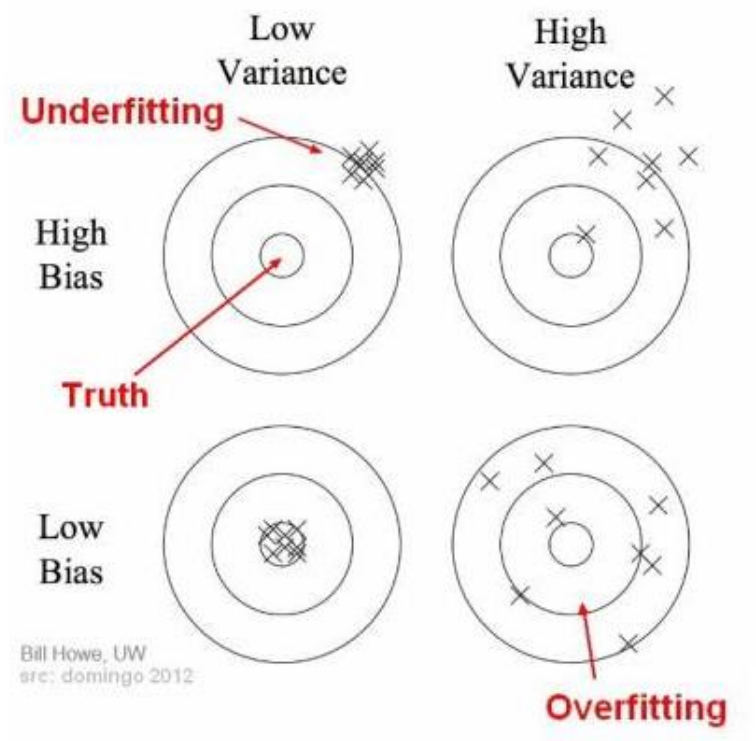
بر خلاف رویکرد یادگیری که تا کنون گفته شده و سعی بر ایجاد یک مدل قوی و جامع بوده است، رویکرد یادگیری گروهی به دنبال استفاده از برآیند تعداد زیادی مدل ضعیف است، بطوری که قادر به پیش بینی با دقت بالا و قوی شود.

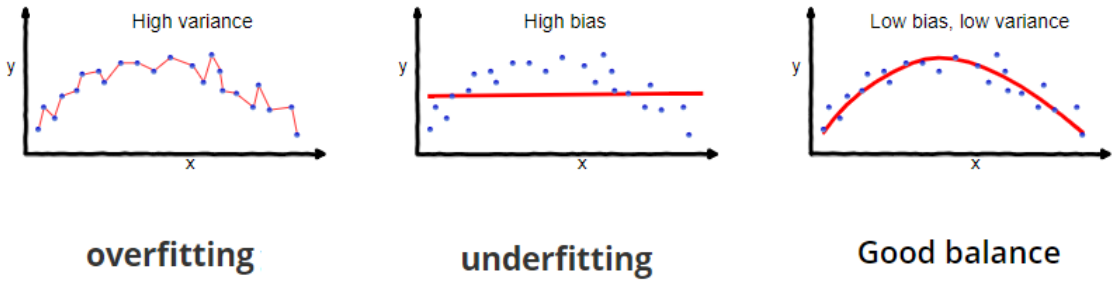
"حتی ضعیف ها هم وقتی متحد شوند، قوی می شوند." – فردریش شیلر

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – یادگیری گروهی





❑ یادگیری گروهی در مدلسازی (Ensemble Learning)

چرا یادگیری گروهی می تواند منجر به نتایج خوب شود؟

- Bias – Variance Trade-off

منابع خطا

سوگیری (Bias): ناشی از وجود فرضیات ساده سازی در مدلسازی است که مدل توانایی شناسایی روابط پیچیده بین ویژگی های ورودی و مقادیر خروجی را از دست می دهد. (کم برازشی)

واریانس: ناشی از پیچیدگی و انعطاف بیش از حد مدل است که منجر به حساسیت بالا به تغییرات جزئی در داده های آموزشی و خطای زیاد در داده های آزمایشی می شود. (بیش برازشی)

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – یادگیری گروهی

❑ یادگیری گروهی در مدلسازی (Ensemble Learning)

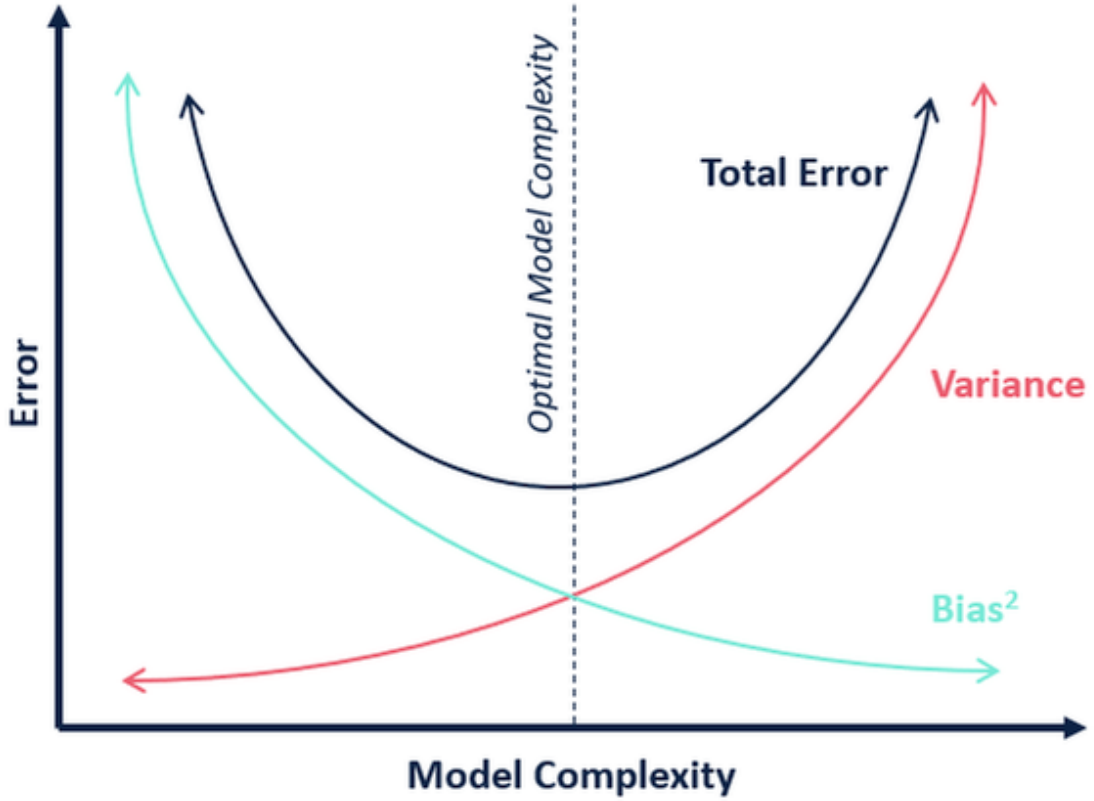
چرا یادگیری گروهی می تواند منجر به نتایج خوب شود؟

- Bias – Variance Trade-off

منابع خطا

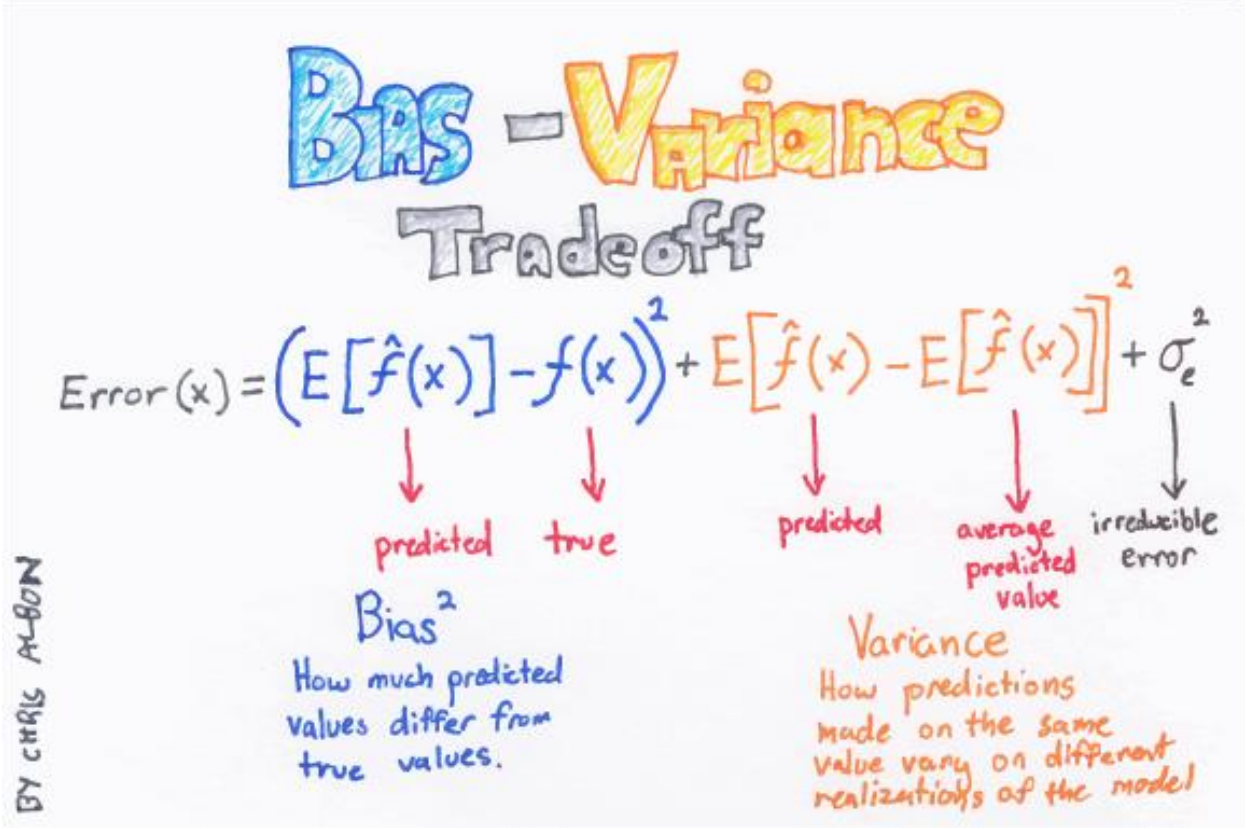
سوگیری (Bias): ناشی از وجود فرضیات ساده سازی در مدلسازی است که مدل توانایی شناسایی روابط پیچیده بین ویژگی های ورودی و مقادیر خروجی را از دست می دهد. (کم برازشی)

واریانس: ناشی از پیچیدگی و انعطاف بیش از حد مدل است که منجر به حساسیت بالا به تغییرات جزئی در داده های آموزشی و خطای زیاد در داده های آزمایشی می شود. (بیش برازشی)



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ یادگیری گروهی در مدلسازی (Ensemble Learning)



چرا یادگیری گروهی می تواند منجر به نتایج خوب شود؟

- Bias – Variance Trade-off

بصورت تئوری انتظار می رود، تمام الگوریتم های مدلسازی، در حین یادگیری از داده های آموزشی به مدلی با میزان سوگیری (بایاس) کم و همچنین مقدار پراکندگی در خطا (واریانس) کم دست پیدا کنند. ولی در خیلی از مواقع دستیابی همزمان به هر دو هدف امکان پذیر نیست. به این حالت مبادله بایاس – واریانس گفته می شود.

$$MSE\left(\hat{\theta}\right) = E\left(\theta - \hat{\theta}\right)^2 = Bias(\hat{\theta})^2 + Variance(\hat{\theta})$$

❑ یادگیری گروهی در مدلسازی (Ensemble Learning)

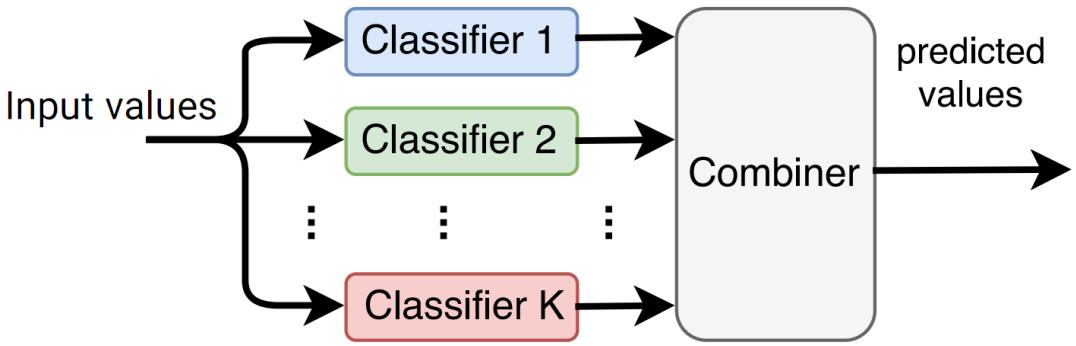
چرا یادگیری گروهی می تواند منجر به نتایج خوب شود؟

یادگیری گروهی از طریق ترکیب چندین مدل آموزش داده شده برای پیش بینی مقادیر هدف استفاده می کند. بسته به رویکرد یادگیری گروهی، مدل های آموزش داده شده (مدل های پایه) می توانند از:

الگوریتم های متفاوت، یا

الگوریتم یکسان در داده های آموزشی متفاوت و یا

الگوریتم یکسان در داده های آموزشی یکسان با وزن های متفاوت بدست آیند.



همچنین ترکیب (ادغام) پیش بینی مدل های پایه نیز با دو روش قابل انجام هست:

خلاصه سازی آماری (استفاده از رای گیری ساده یا وزنی، میانگین، میانه)

استفاده از Metaclassifier (مدل متا)

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

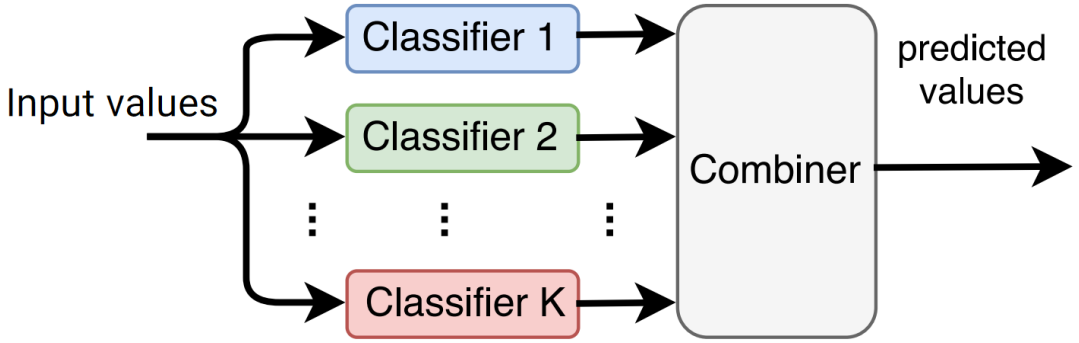
❑ یادگیری گروهی در مدلسازی (Ensemble Learning)

چرا یادگیری گروهی می تواند منجر به نتایج خوب شود؟

مزایای یادگیری گروهی

افزایش کارایی مدل ها: هریک از مدل های انفرادی ممکن است در زیرمجموعه ای از داده ها و الگوهای موجود دارای عملکرد خوب یا بد باشند. ادغام آنها امکان ایجاد یک مدل قوی برای شرایط مختلف را با کاهش بایاس (سوگیری) به وجود می آورد.

افزایش پایداری نتایج: استفاده از شاخص های خلاصه سازی آماری در یادگیری گروهی، امکان کاهش واریانس مقادیر پیش بینی در مدل های انفرادی و در نتیجه افزایش اطمینان و پایداری نتایج را فراهم می سازد.



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – یادگیری گروهی

❑ یادگیری گروهی در مدلسازی (Ensemble Learning)

رویکردهای متنوعی در پیاده سازی یادگیری گروهی وجود دارد، ولی عموما سه رویکرد رایج در یادگیری گروهی مورد استفاده قرار می گیرد:

Stacking

با هدف بهبود پیش بینی مدل

Bagging

با هدف کاهش واریانس مدل

Boosting

با هدف کاهش بایاس مدل

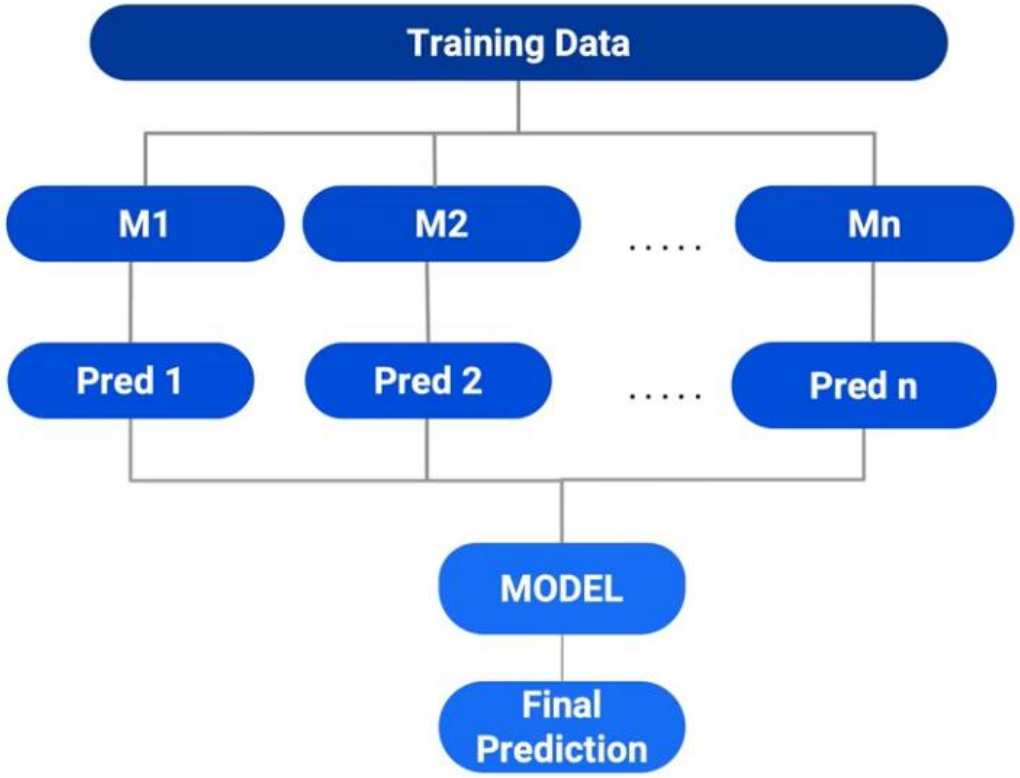
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – یادگیری گروهی

❑ رویکرد Stacking

رویکرد Stacking در یادگیری گروهی بر اساس ترکیب مدل های پایه با الگوریتم های متفاوت انجام می شود. بنابراین در این رویکرد عمدتا از الگوریتم های با ساختارهای مختلف استفاده می شود و با ادغام نتایج آنها، پیش بینی نهایی حاصل می شود.



ایده اصلی در این رویکرد اینست که ساختارهای متفاوت از الگوریتم های پایه توانایی تشخیص الگوهای متفاوتی در داده ها را دارند که ادغام آنها منجر به بهبود تصمیم گیری و پیش بینی می شوند.

در ادغام نتایج مدل های پایه در رویکرد Stacking، معمولا از یک مدل جدید با عنوان Metaclassifier استفاده می شود و پیش بینی مدل های پایه به عنوان ورودی مدل متا در نظر گرفته می شود.

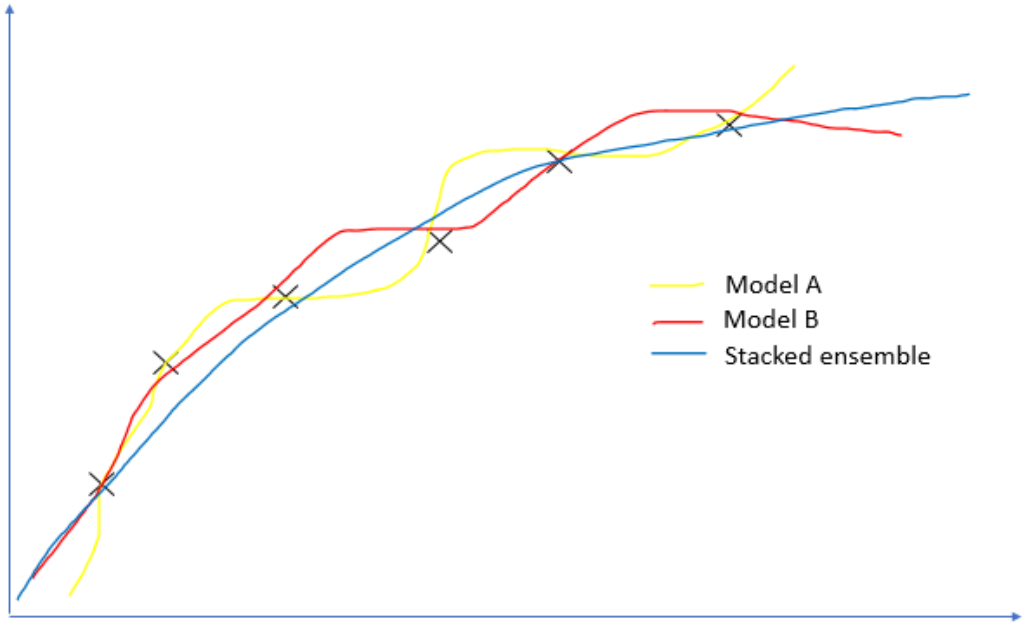
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – یادگیری گروهی

❑ رویکرد Stacking

رویکرد Stacking معمولا با هدف بهبود کارایی مدل و کاهش بایاس و واریانس مدل مورد استفاده قرار می گیرد؛ اما بطور ویژه اثر زیادی در کاهش واریانس و بیش برازشی دارد.



استفاده از مدل های پایه با ساختار های متفاوت و همچنین خروجی های متفاوت، به خصوص در مسائل رگرسیون (که تفاوت در پیش بینی مقادیر هدف توسط مدل های پایه زیاد است) از جمله مواردی هست که رویکرد Stacking با کاهش بایاس به ارتقای مدل کمک می کند.

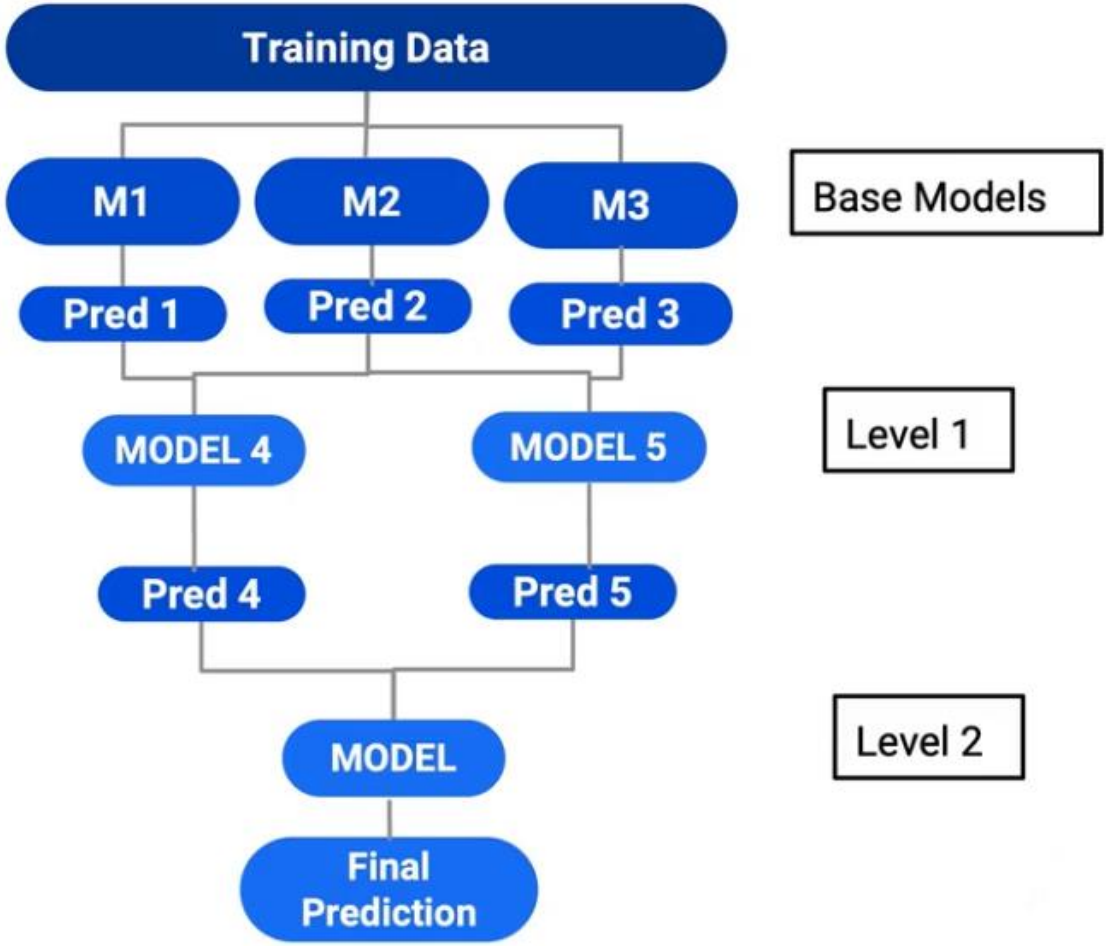
در مثال روبرو، مدل A و مدل B مدل های با مقدار بایاس کم و واریانس بالا هستند که با دقت بسیار زیاد بر روی داده های نمونه دچار بیش برازش شده اند. رویکرد Stacking توانایی کاهش واریانس مدل را فراهم می سازد تا مدل مقاوم تر و پایدارتری یا ایجاد نماید.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – یادگیری گروهی

❑ رویکرد Stacking



ادغام پیش بینی مدل های پایه، می تواند در چندین لایه یا سطح انجام پذیرد. در این حالت در لایه اول ادغام، به جای استفاده از یک مدل متا، می توان از چندین مدل استفاده نمود و سپس پیش بینی های حاصل از هر مدل متای لایه اول را، در لایه دوم وارد مدل متای نهایی نمود.

نکته دیگری که در رویکرد Stacking وجود دارد، استقلال بین آموزش مدل های مختلف می باشد، به این معنی که هر یک از مدل ها دارای فرضیات مربوط به خود هستند. به همین دلیل این رویکرد در چارچوب یادگیری گروهی مستقل قرار می گیرد.

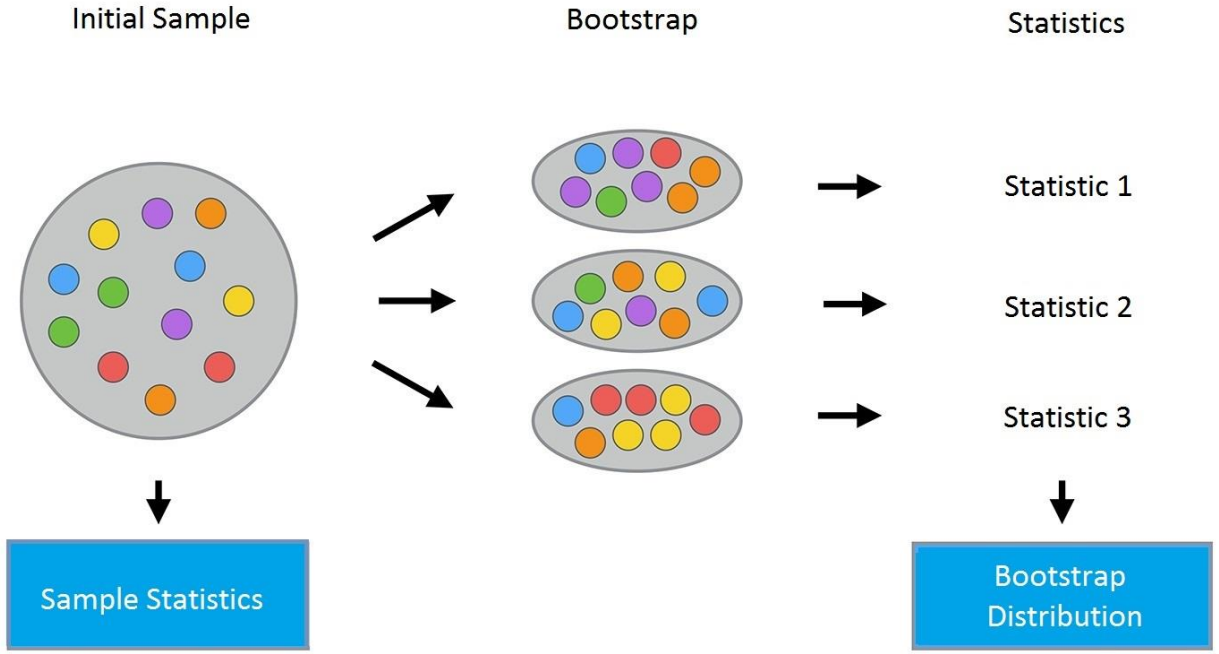
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – یادگیری گروهی

❑ رویکرد Bagging – (Bootstarp AGGregatING)

روش رایج در رویکرد Bagging ساخت مدل های پایه با استفاده از الگوریتم یکسان در داده های آموزشی متفاوت می باشد. بطوریکه نمونه ها آزمایشی متعددی با روش بوت استرپ از داده های آموزشی ساخته شده و یک الگوریتم یکسان با استفاده از آنها آموزش داده می شود.



روش بوت استرپ (Bootstrap)

یک روش آماری برای محاسبه پارامترهای جامعه از طریق نمونه گیری با جایگذاری می باشد. در تخمین پارامترهای آماری می توان با ایجاد نمونه های متعدد از طریق نمونه گیری با جایگذاری و میانگین گرفتن از برآوردهای هر نمونه به برآورد پارامتر مجهول دست یافت.

در رویکرد Bagging، با این روش مجموعه داده های آزمایشی متعددی ایجاد شده و در ساخت مدل های پایه استفاده می شود.

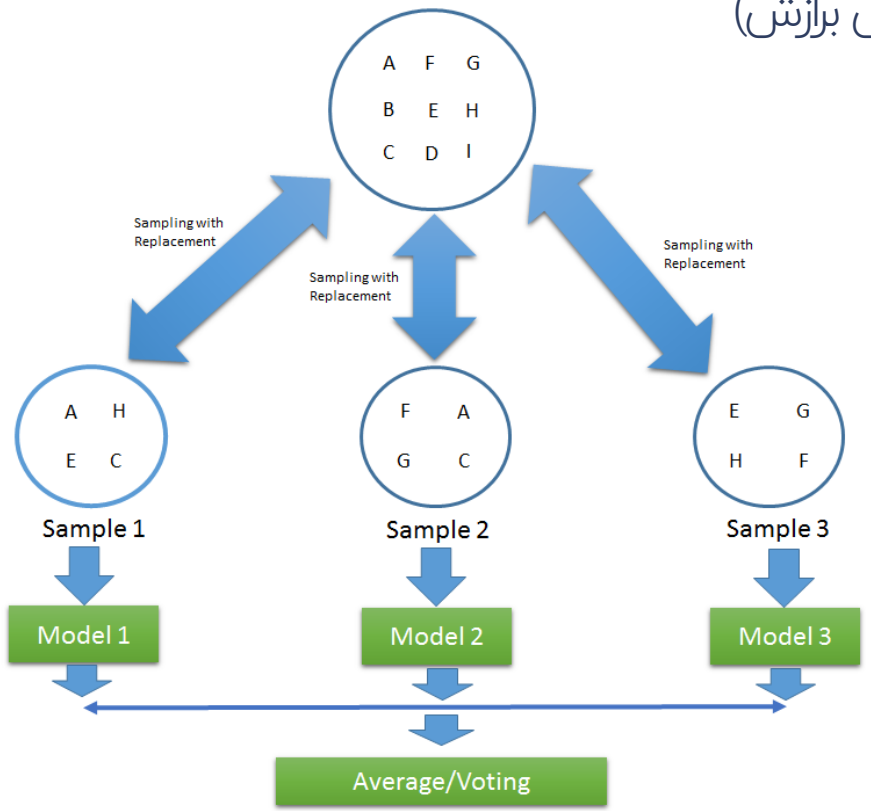
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – یادگیری گروهی

❑ رویکرد Bagging – (Bootstarp AGGregatING)

کارکرد اصلی روش Bagging در کنترل و کاهش واریانس خطای پیش بینی است. بنابراین در مواقعی از این رویکرد استفاده می شود که بایاس مدل های پایه کم و واریانس آنها زیاد باشد. (مدل های با پیچیدگی زیاد و دارای بیش برازش)



افزایش تعداد مدل های پایه در این رویکرد، به علت استفاده از میانگین یا مد در نتیجه نهایی منجر به بیش برازش نمی شود و عموما تا جایی که منجر به بهبود در نتایج نگردد، تعداد مدل ها افزایش می یابد.

رویکرد Bagging نیز در چارچوب یادگیری گروهی مستقل قرار می گیرد و از ویژگی های آن در پیاده سازی، موازی سازی آموزش مدل های پایه می باشد.

به همین دلیل در آموزش مدل های رویکرد Bagging از حداکثر CPU استفاده شده و زمان آن برابر با آموزش یک مدل پایه می باشد.

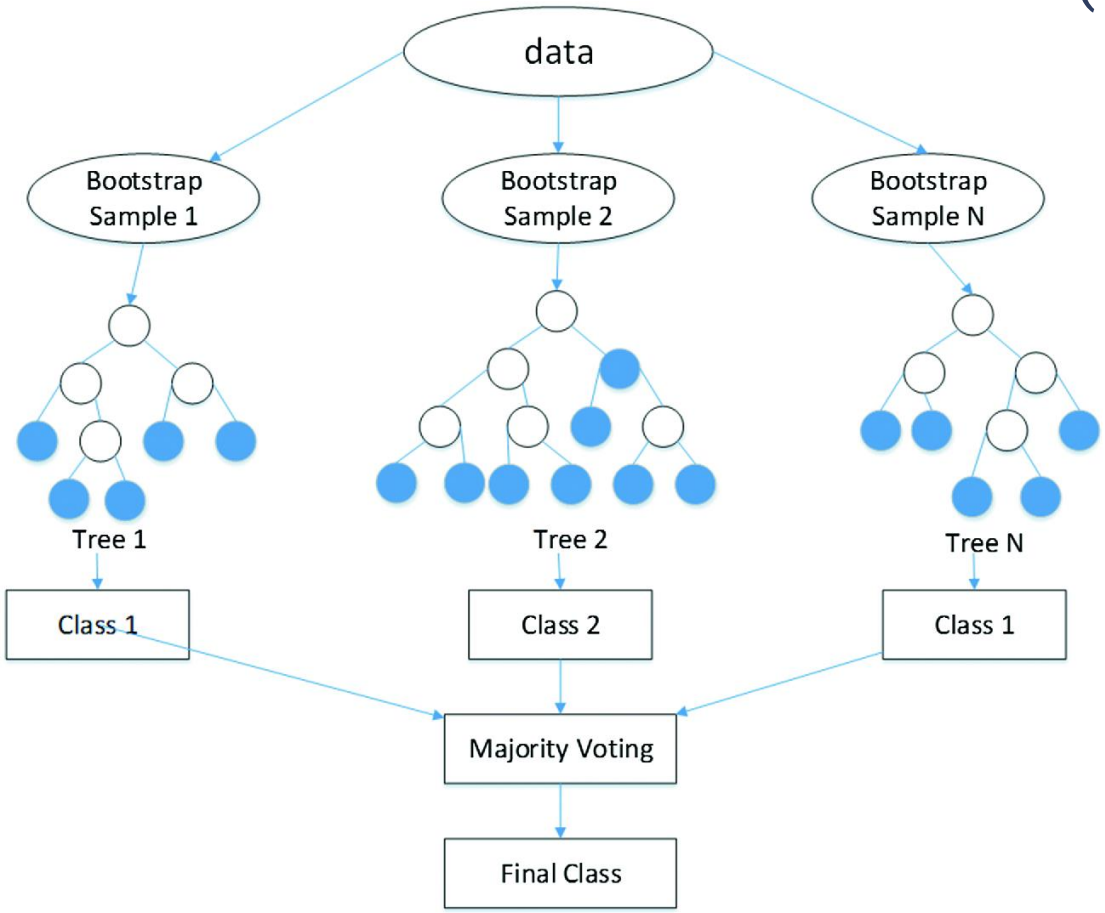
تولید محتوا: زهرا ذوالقدر

فرآیند داده کاوی

مدل های پیش بینانه – یادگیری گروهی

❑ رویکرد Bagging – (Bootstarp AGGregatING)

- **الگوریتم جنگل تصادفی (Random Forest)**



این الگوریتم یکی از محبوب ترین و رایج ترین الگوریتم های یادگیری گروهی با رویکرد Bagging می باشد که در دامنه وسیعی از مسائل رده بندی و رگرسیون، عملکرد بسیار خوبی داشته است.

الگوریتم جنگل تصادفی در چارچوب رویکرد Bagging با در نظر گرفتن الگوریتم درخت تصمیم هرس نشده (با عمق زیاد) به عنوان مدل پایه، تعداد بسیار زیادی مدل بیش برازش شده دارای بایاس کم ایجاد نموده و با رای گیری یا میانگین گیری از پیش بینی مدل های پایه از بیش برازش کلی مدل جلوگیری می کند.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

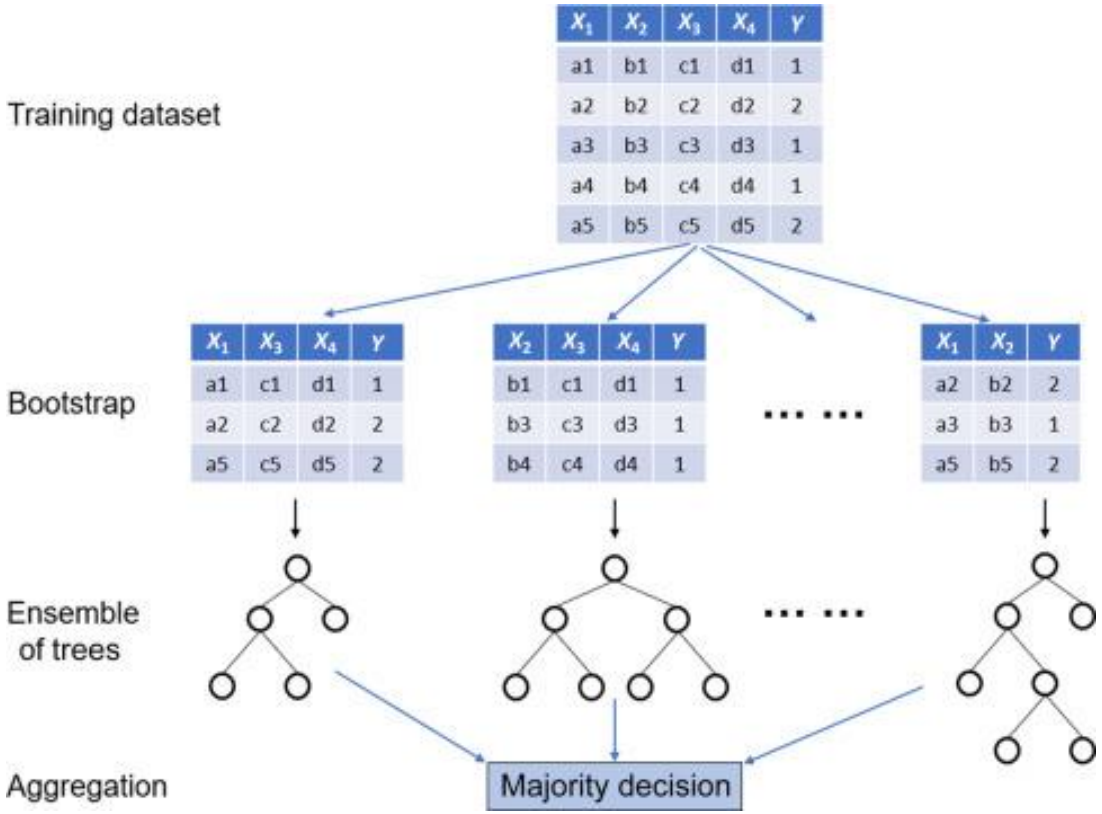
فرآیند داده کاوی

مدل های پیش بینانه – یادگیری گروهی

❑ رویکرد Bagging – (Bootstarp AGGregatING)

- **الگوریتم جنگل تصادفی (Random Forest)**

یک تفاوت مهم در الگوریتم جنگل تصادفی نسبت به رویکرد کلی Bagging استفاده از زیرمجموعه ای از ویژگی های ورودی در آموزش و ساخت مدل های پایه می باشد. این موضوع منجر به تفاوت بیشتر بین مدل های پایه و کاهش همبستگی نتایج آنها می شود.



معمولا در **مسائل رگرسیون** تعداد ویژگی های ورودی هر مدل پایه، در حدود یک سوم ویژگی های ورودی اصلی و در **مسائل رده بندی**، جذر تعداد ویژگی های ورودی در نظر گرفته می شود.

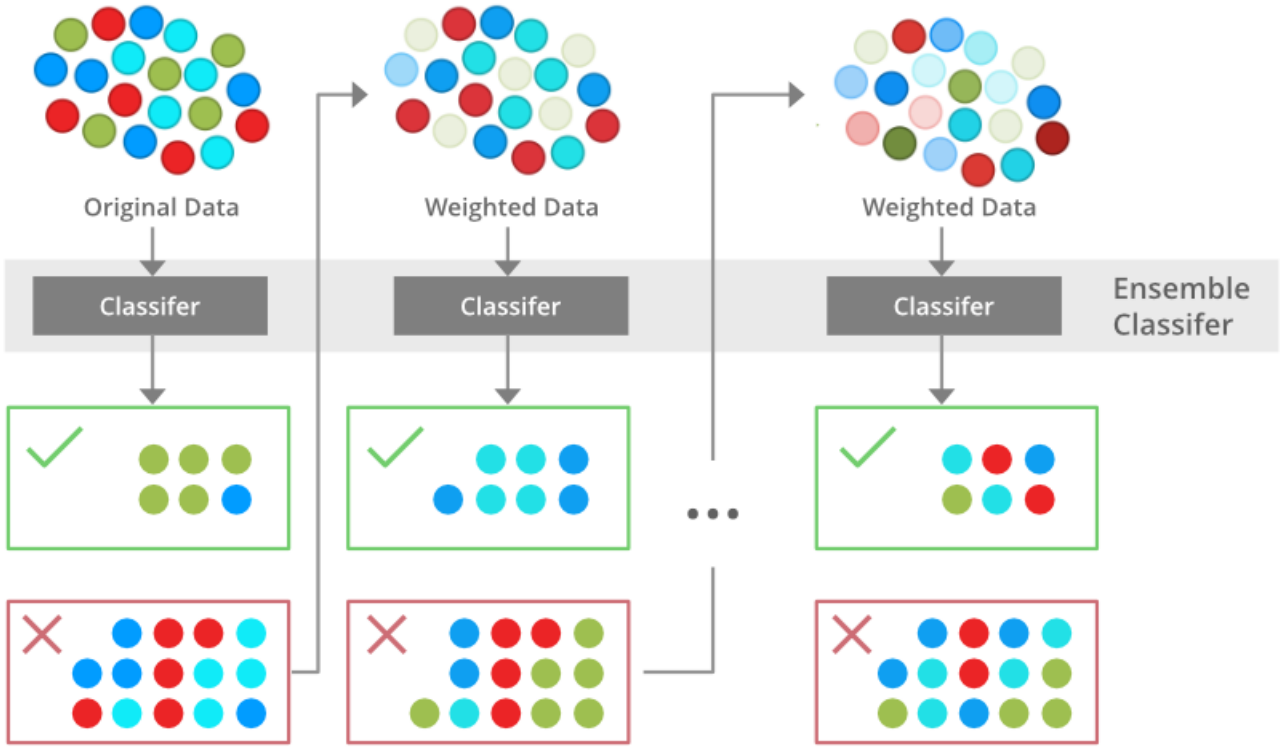
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – یادگیری گروهی

❑ رویکرد Boosting

رویکرد Boosting بر اساس آموزش مدل های پایه با الگوریتم های یکسان روی داده های آموزشی یکسان با وزن های متفاوت به مدلسازی و حل مسائل رده بندی و رگرسیون می پردازد.



در این رویکرد آموزش هر یک از مدل های پایه، با وزن دادن به رکوردهایی که در مدل قبلی دچار خطای پیش بینی بوده اند، تمرکز یادگیری مدل را روی خطاهای مدل قبلی قرار می دهد.

بنابراین تفاوت مدل های پایه در تمرکز یادگیری آنها بر روی زیرمجموعه هایی از داده هاست که یادگیری الگوی آنها سخت است.

رویکرد Boosting بهترین مثال برای ایده "**قدرت در یکپارچگی و وحدت ضعیف هاست**"، می باشد.

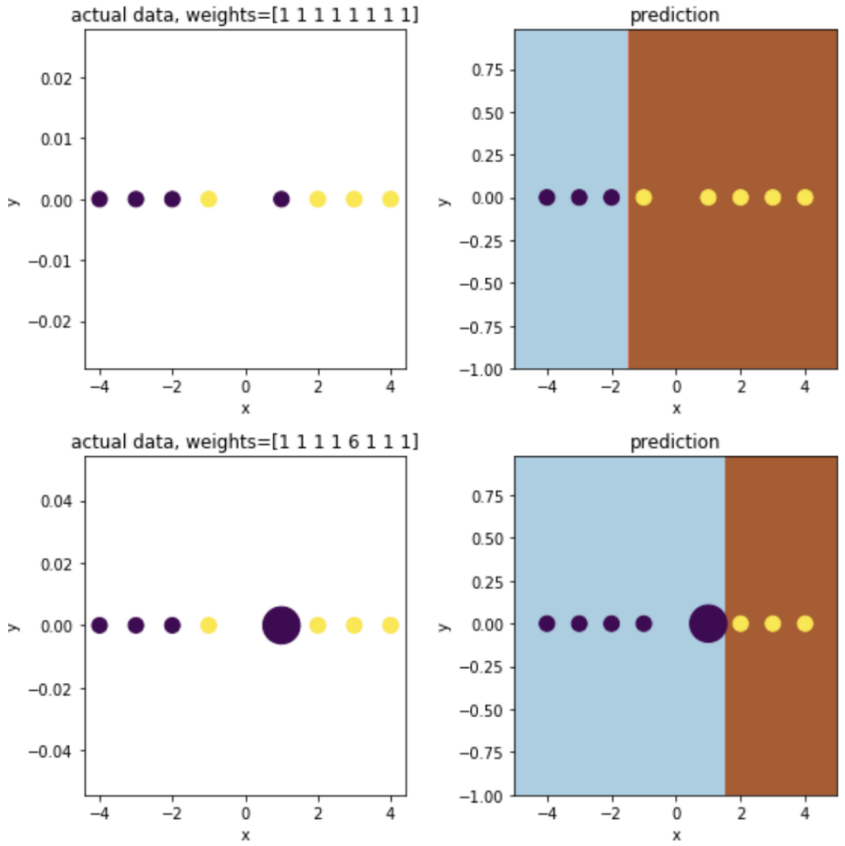
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – یادگیری گروهی

❑ رویکرد Boosting

کارکرد اصلی رویکرد Boosting در کاهش بایاس مدل های پیش بینانه است. بنابراین از این رویکرد عموما در مدل هایی که دارای واریانس کم بوده ولی به علت سادگی مدل، دقت آن پایین و دچار کم برازشی می باشد، استفاده می شود.



این رویکرد در چارچوب یادگیری گروهی وابسته قرار دارد و به علت وابستگی هر مدل به خطاهای مدل قبلی، پیاده سازی آن در قالب آموزش متوالی (Sequential) مدل های پایه انجام می شود. از این رو در ساخت مدل با این رویکرد، زمان آموزش با افزایش تعداد مدل های پایه نیز بصورت خطی افزایش می یابد.

نکته: هرچند رویکرد Boosting به علت ایده ترکیب مدل های پایه در یادگیری گروهی، کاهش واریانس را در پی خواهد داشت و نسبت به بیش برازشی مقاوم است، اما در صورتی که مدل های انفرادی دچار بیش برازشی باشد، توانایی حل آن را نخواهد داشت.

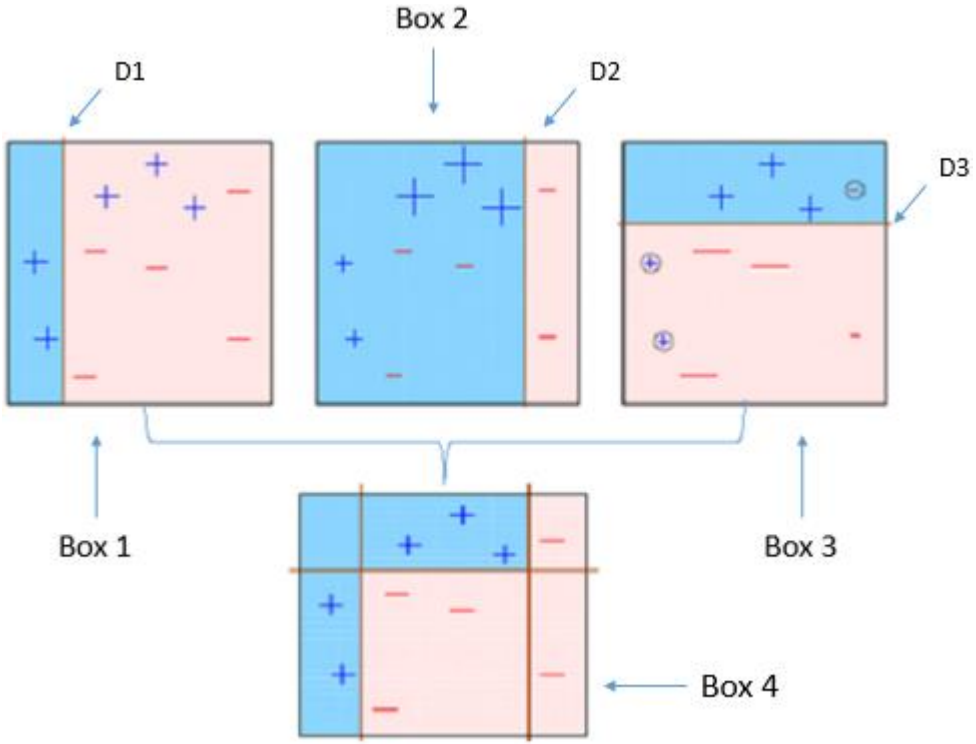
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مدل های پیش بینانه – یادگیری گروهی

❑ رویکرد Boosting

- الگوریتم AdaBoost (Adaptive Boosting)



این الگوریتم با رویکرد Boosting و با استفاده از مدل های پایه درخت تصمیم با عمق یک (یک انشعاب روی گره ریشه) توسعه می یابد.

مدل پایه اول روی همه داده های آموزشی با وزن یکسان ایجاد می شود و با مقایسه پیش بینی و مقادیر واقعی، وزن رکوردهای آموزشی تغییر می کند، بطوریکه رکوردهای دارای خطا در مدل اول، دارای وزن بیشتری نسبت به رکوردهایی که به درستی پیش بینی شده اند می گردد.

این روند تا رسیدن به قوانین توقف (تعداد مدل های پایه یا مقدار تعیین شده از شاخص ارزیابی) ادامه پیدا می کند.

الگوریتم های مطرح دیگری همچون Gradient Boosting و نسخه بهینه شده آن XGBoost نیز در همین چارچوب توسعه داده شده که جزو الگوریتم های قدرتمند و محبوب در یادگیری گروهی می باشند.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه