مباحث تکمیلی

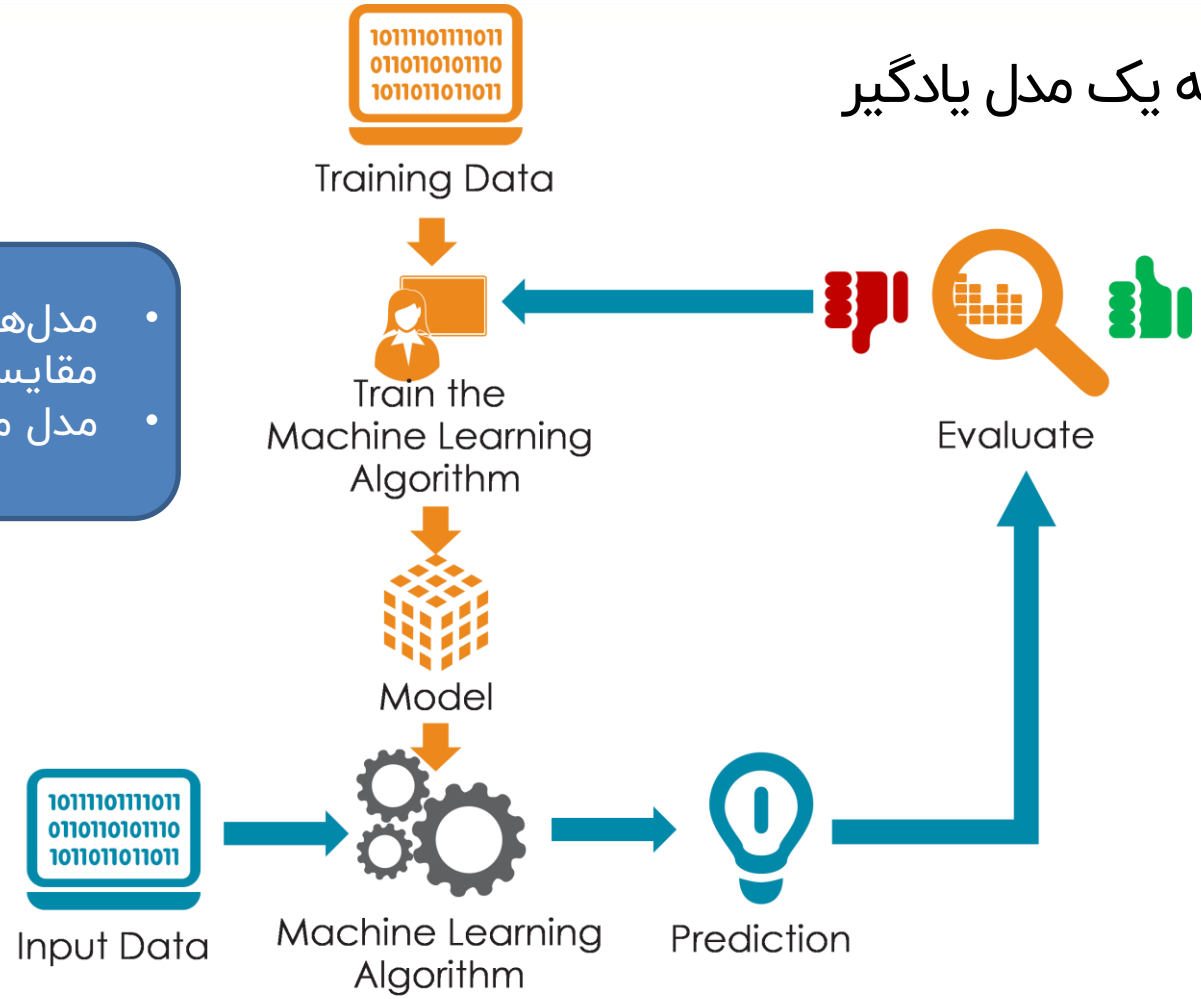
یادگیری ماشین

dayche.com . گروه دایکه

دای‌کِه DAYCHE

توسعه مدل یادگیری ماشین

- مسیر توسعه یک مدل یادگیر



- مدل‌های متفاوت یادگیری ماشین چگونه با هم مقایسه می‌شوند؟
- مدل مناسب چگونه انتخاب می‌شود؟

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

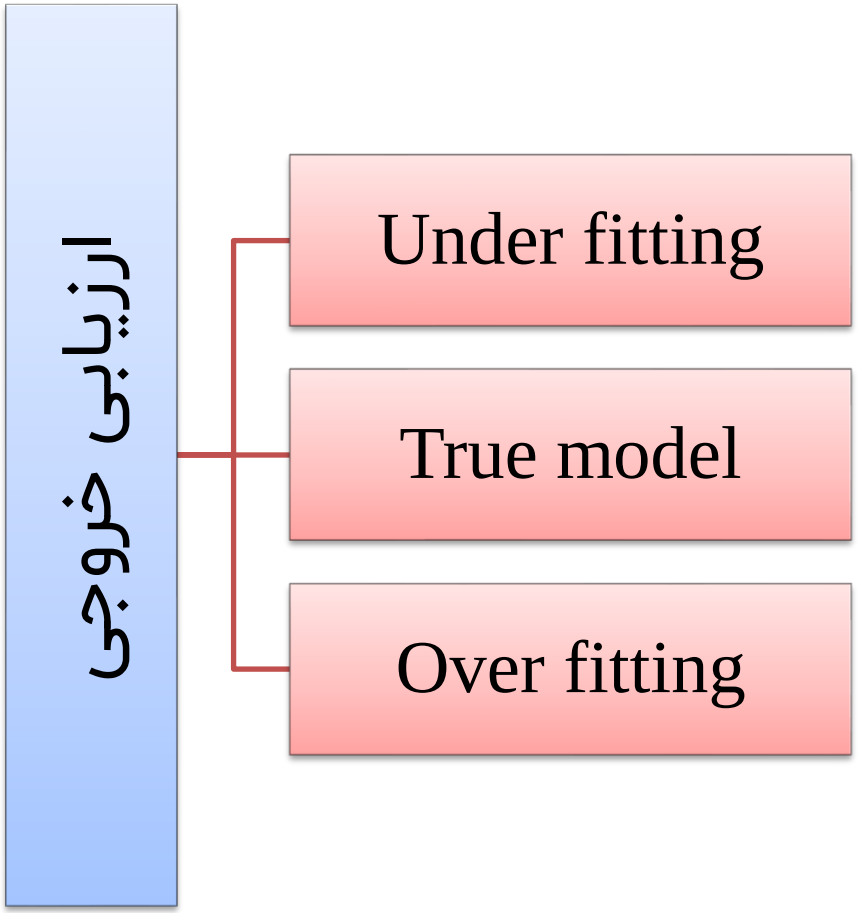
بهبود عملکرد مدل

- چگونه می‌توان یک مدل پیش‌بین و یا دسته‌بند طراحی کرد؟
- بهبود عملکرد
  - چگونه میتوان یک الگوریتم را توسعه داد طوری که نه تنها بر روی داده‌های آموزش عملکرد مناسبی داشته باشد، بلکه بر روی داده‌های تست نیز به خوبی عمل کند.
  - کاهش خطای تست – یک مسئله مهم و اساسی در یادگیری ماشین

***Regularization**: Any modification that makes a learning algorithm to reduce the generalization error but not training error*

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

تنظیم‌سازی مدل



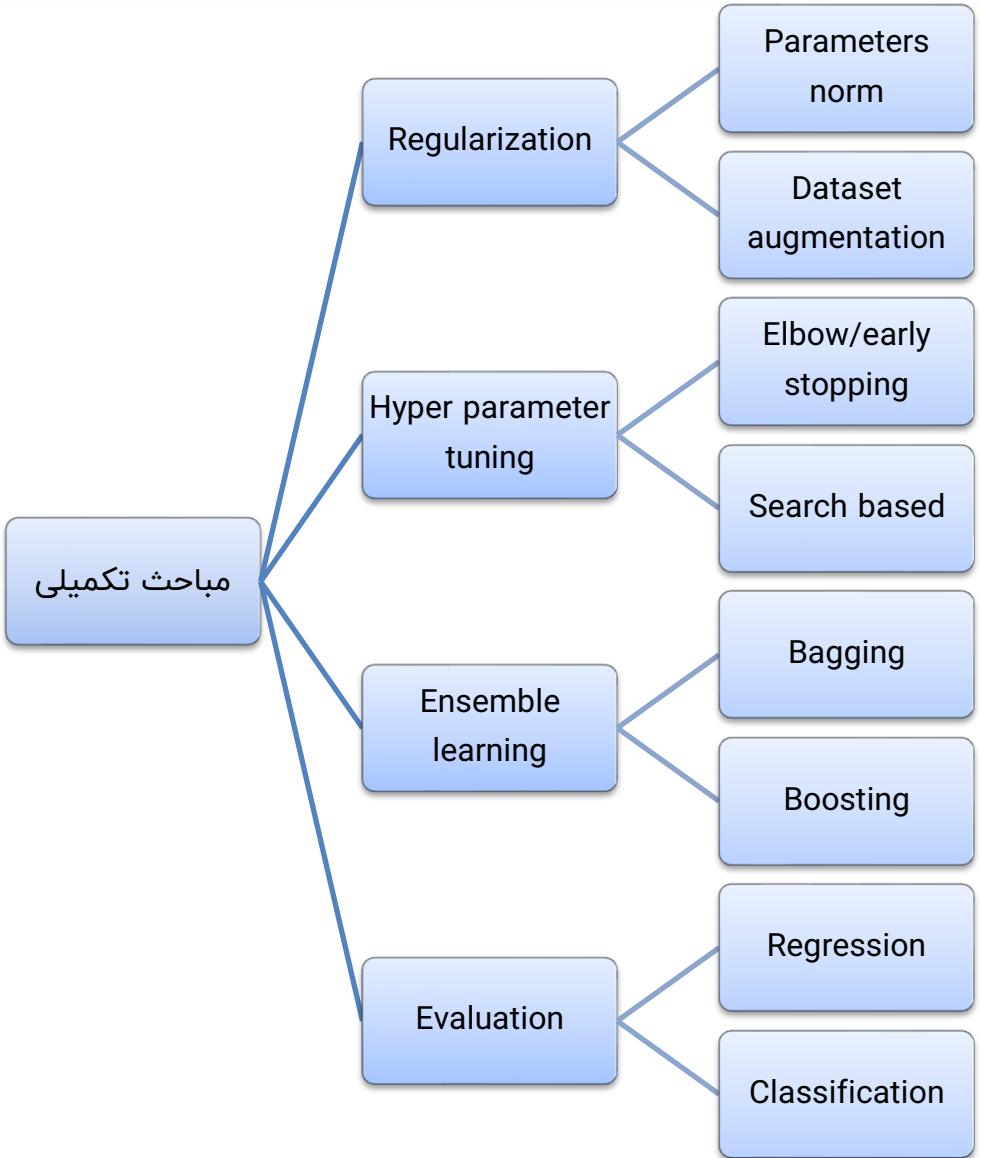
مدل قادر به شناسایی فرآیند تولید داده نیست.
- نیازمند تعیین کردن ابرپارامترها به صورت بهینه است

مدل تنها قادر به شناسایی فرآیند تولید داده است.
- ابر پارامترها به درستی انتخاب شده‌اند

مدل قادر به شناسایی فرآیند پیچیده‌تری از تولید داده است.
- ابرپارامترها بیش از حد پیچیده انتخاب شده‌اند

تولید محتوا: وحید محمدزاده ایوقی

مباحث تکمیلی



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

تنظیم‌سازی بر اساس نرم پارامترها

- محدود کردن ظرفیت مدل‌ها (مدل‌های رگرسیون خطی، لوجستیک، شبکه‌های عصبی و ...) با افزودن نرم پارامترها

Regularized cost function $\tilde{J}(x, y; \theta) = J(x, y; \theta) + \alpha\underbrace{\phi(\theta)}_{\text{Norm function}}$

- تعبیر حداقل کردن این تابع هزینه
- نرم‌های متفاوت اثر متفاوتی بر روی مدل می‌گذارند.
- برای مدل‌های شبکه عصبی تنظیم‌کنندگی تنها بر روی وزن‌ها صورت می‌گیرد. چرا؟

تنظیم‌سازی L2

- تنظیم‌سازی Ridge regression (Tikhonov regularization) - L2

$$\tilde{J}(x, y; \theta) = J(x, y; \theta) + \alpha\phi(\theta)\ , \phi(\theta) = \frac{1}{2}|w|_2^2$$

$$w^+ = w^- - \eta(\alpha w^- + \nabla J) = (1 - \eta\alpha)w^- - \eta\nabla J$$

- قبل از به‌روزرسانی، وزن‌ها کوچک می‌شوند. آیا می‌توان گفت کدام وزن‌ها بیشتر کاهش خواهند شد؟
- اگر ترم تنظیم‌کنندگی وجود نداشته باشد:

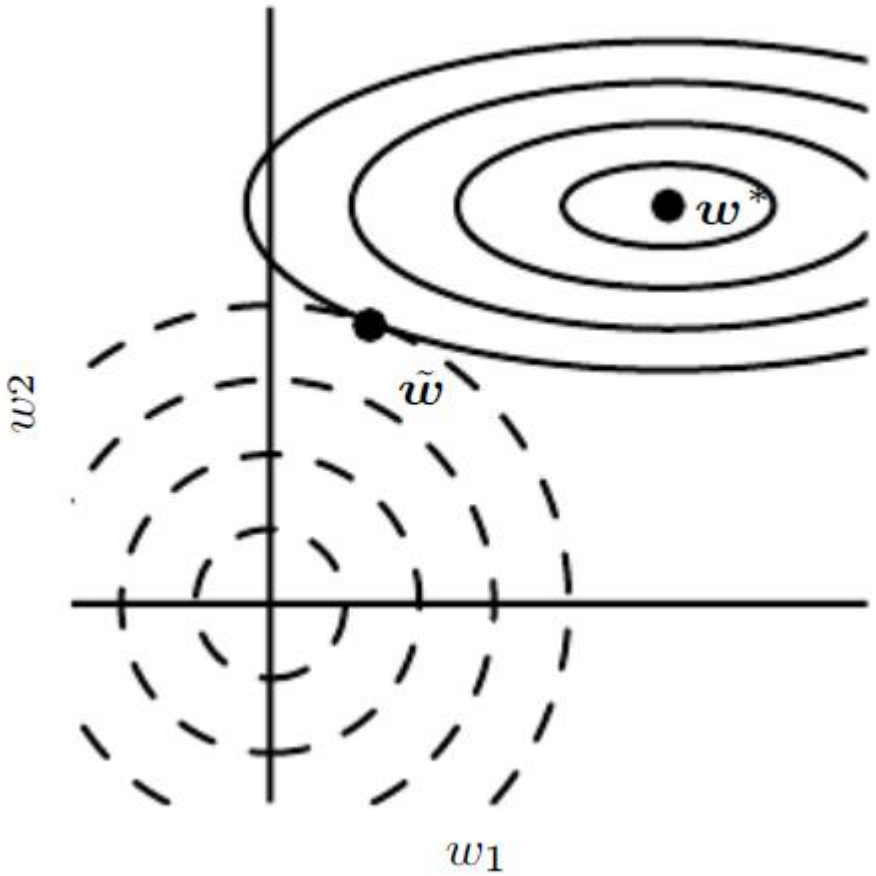
$$w^* = \arg\min J(w) \rightarrow J(w) \approx J(w^*) + (w - w^*)H(w - w^*)$$

$$\alpha w + H(w - w^*) = 0 \rightarrow w = (\alpha I + H)^{-1} H w^*$$

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

تنظیم‌سازی L2

- تنظیم‌سازی L2 - تعبیر هندسی



$$w = (\alpha I + H)^{-1} H w^*$$

پارامترها در راستای کوچکترین بردار ویژه ماتریس هسیان به صفر متمایل می‌شوند

تقریب مرتبه دو از تابع هزینه حول نقطه بهینه تنظیم نشده

- اگر تابع هزینه مرتبه دو باشد، این تحلیل دقیق است.

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

تنظیم‌سازی L2

- تنظیم‌سازی L2 - مسئله رگرسیون خطی
- تنظیم‌سازی غیر صفر

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

تنظیم‌سازی L1

- تنظیم‌سازی L1

$$\tilde{J}(x, y; \theta) = J(x, y; \theta) + \alpha\phi(\theta)\ , \phi(\theta) = |w|_1 = \Sigma|w_i|$$

$$w^+ = w^- - \eta(\alpha sign(w^-) + \nabla J)$$

- چه تغییری بر روی وزن‌ها ایجاد خواهد شد؟

$$w_i = sign(w_i^*)\max\left(|w_i^*| - \frac{\alpha}{H_{i,i}}, 0\right)$$

- ماتریس هسیان یک ماتریس قطری فرض شده است.
- چرا خاصیت تنک‌زدایی در L2 regularization وجود ندارد؟

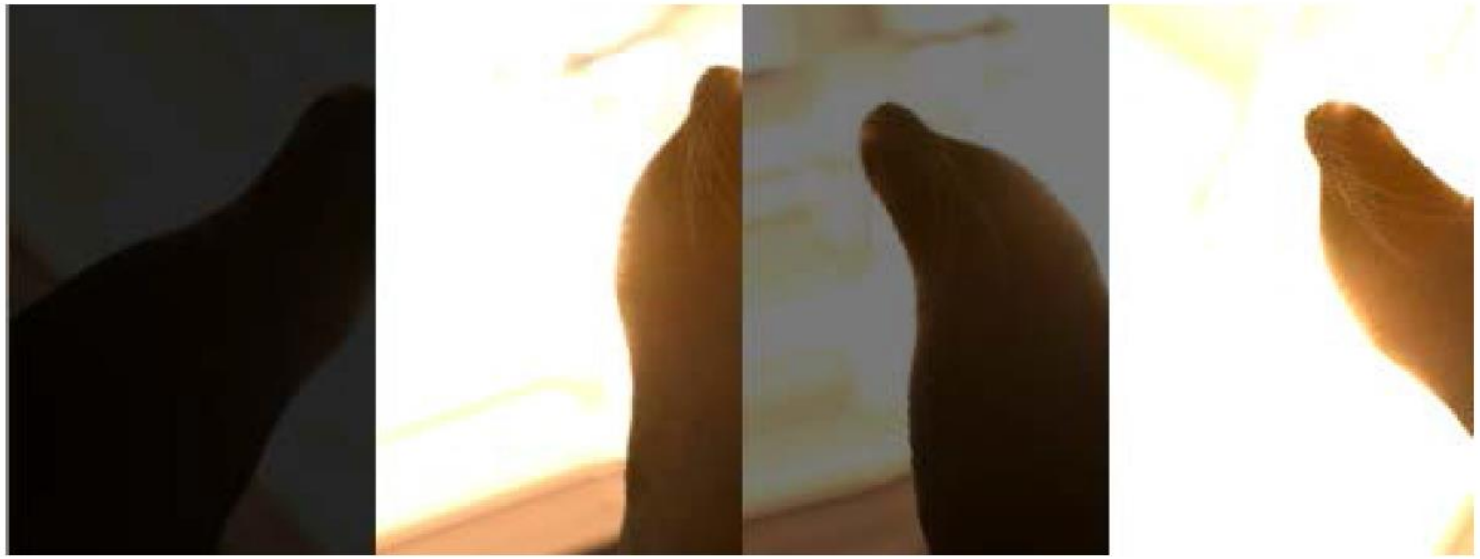
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

افزایش داده Dataset augmentation

- بیش برازش شدن

در کاربردهای عملیاتی تعداد داده‌ها محدود است. برای آموزش بر اساس داده‌های زیاد چه باید کرد؟

Image recognition tasks

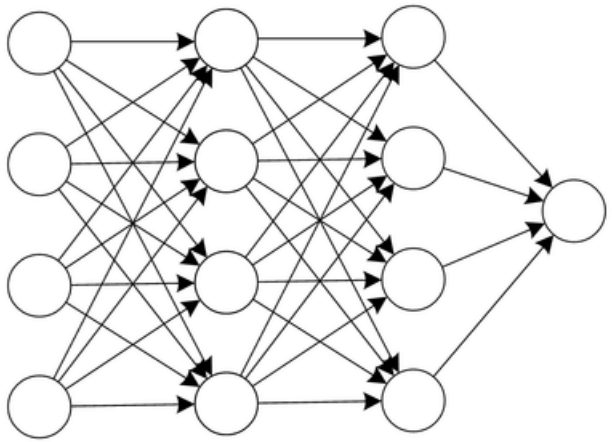


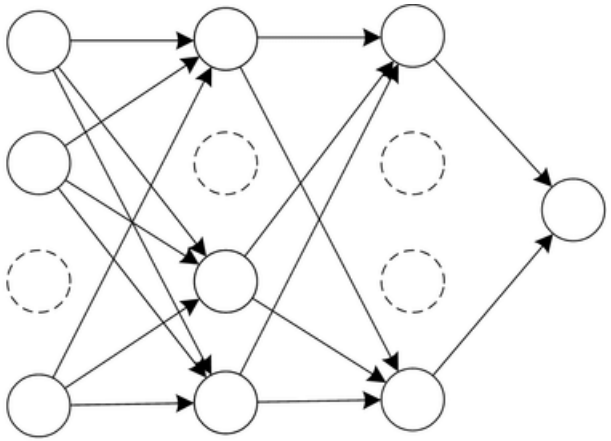
آیا این رویکرد برای تمام تسک‌ها مناسب است؟

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Dataset augmentation افزایش داده

- افزودن نویز ورودی Noise injection
  - خصوصا برای مدلهایی که به صورت گام به گام آموزش می‌بینند مانند شبکه‌های عصبی
- Dropout
  - می‌توان به صورت استراتژی افزایش داده نیز به آن گاه کرد





تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

افزودن نویز Noise injection

- افزودن نویز ورودی
  - افزودن نویز غالبا بر روی ورودی صورت می‌گیرد و نوعی از رویکردهای افزایش داده است.
  - می‌توان نویز را به سطوح دیگری از مدل اضافه کرد.
    - افزودن نویز به وزن‌های شبکه عصبی ,افزودن نویز به لایه فعال‌ساز ,افزودن نویز به بردار گرادیان
- افزودن نویز خروجی – احتمال اشتباه بودن برچسب خروجی

$$w^* = \arg\max P(y|x)$$

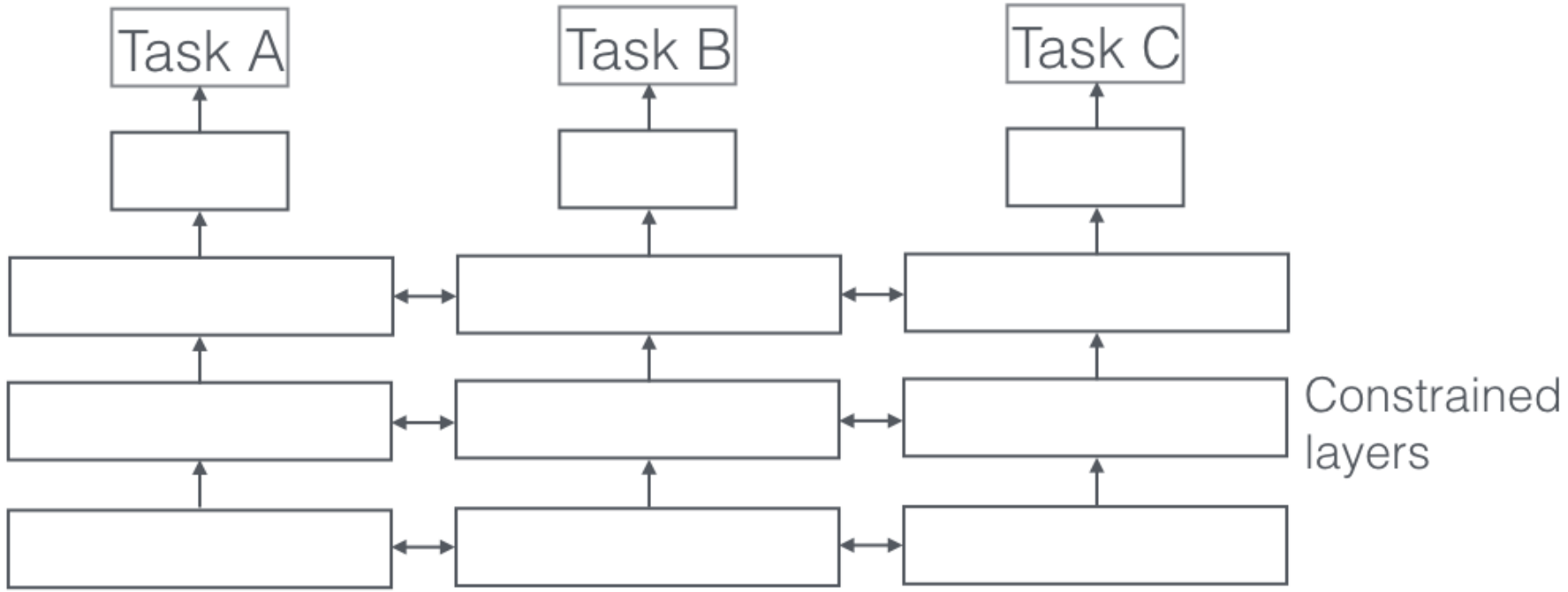
بهینه‌سازی این کمیت منجر به خطای زیادی در مدل‌سازی خواهد شد، چاره چیست؟

Label smoothing

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

رویکردهای خاص

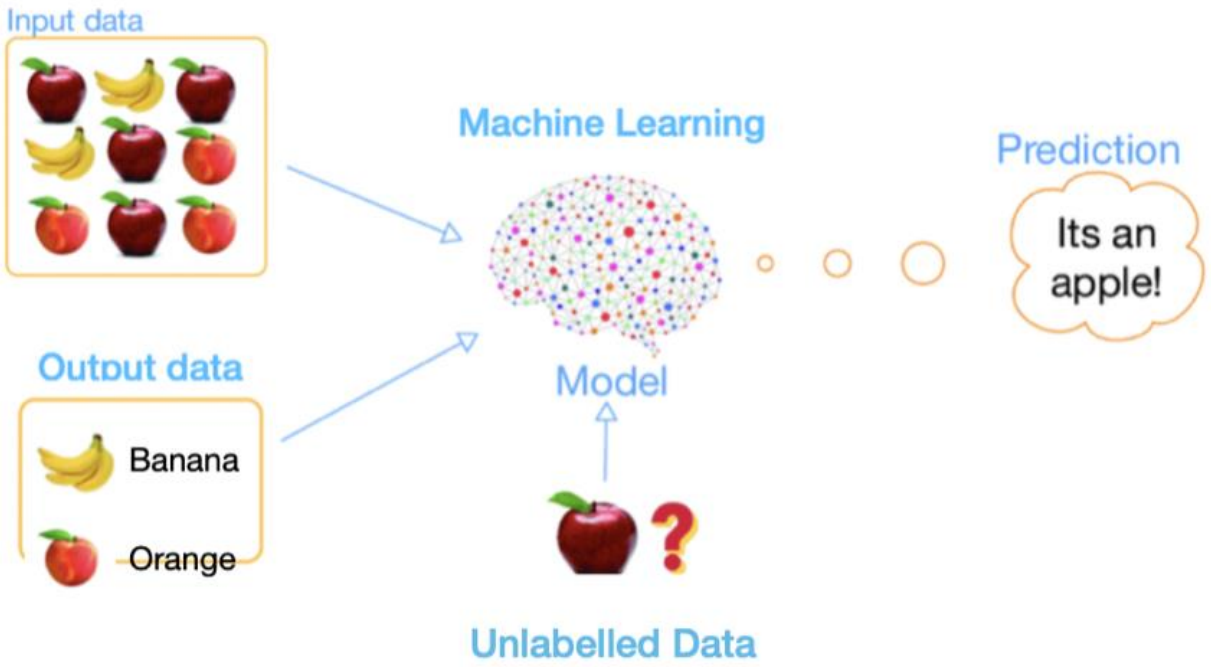
یادگیری چند وظیفه‌گی



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
گروه دایکه | dayche.com

رویکردهای خاص

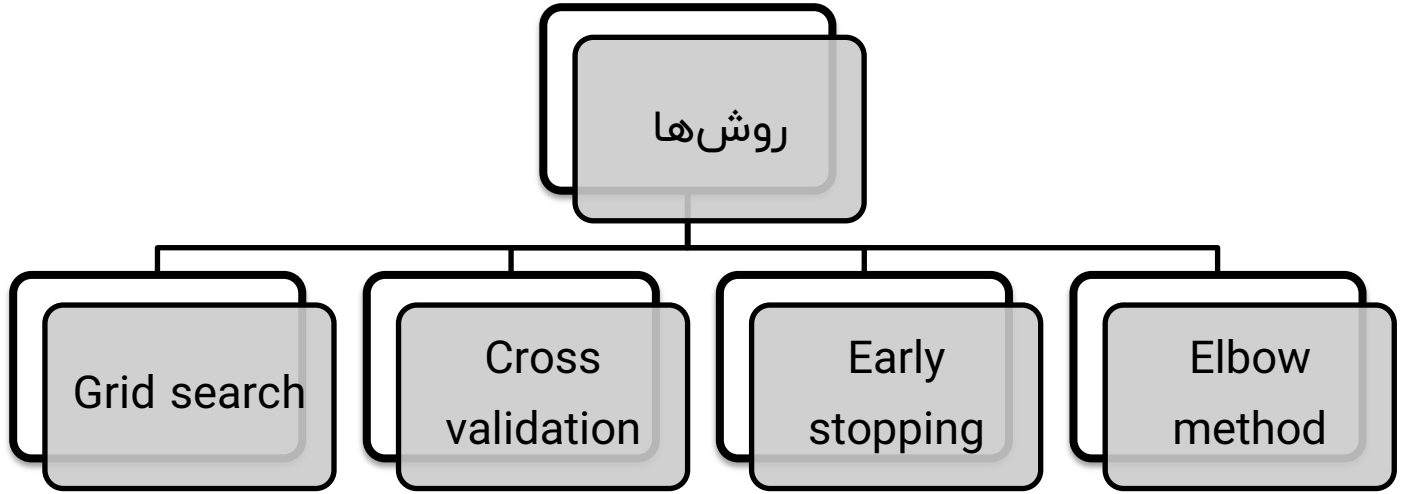
یادگیری نیمه نظارتی



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

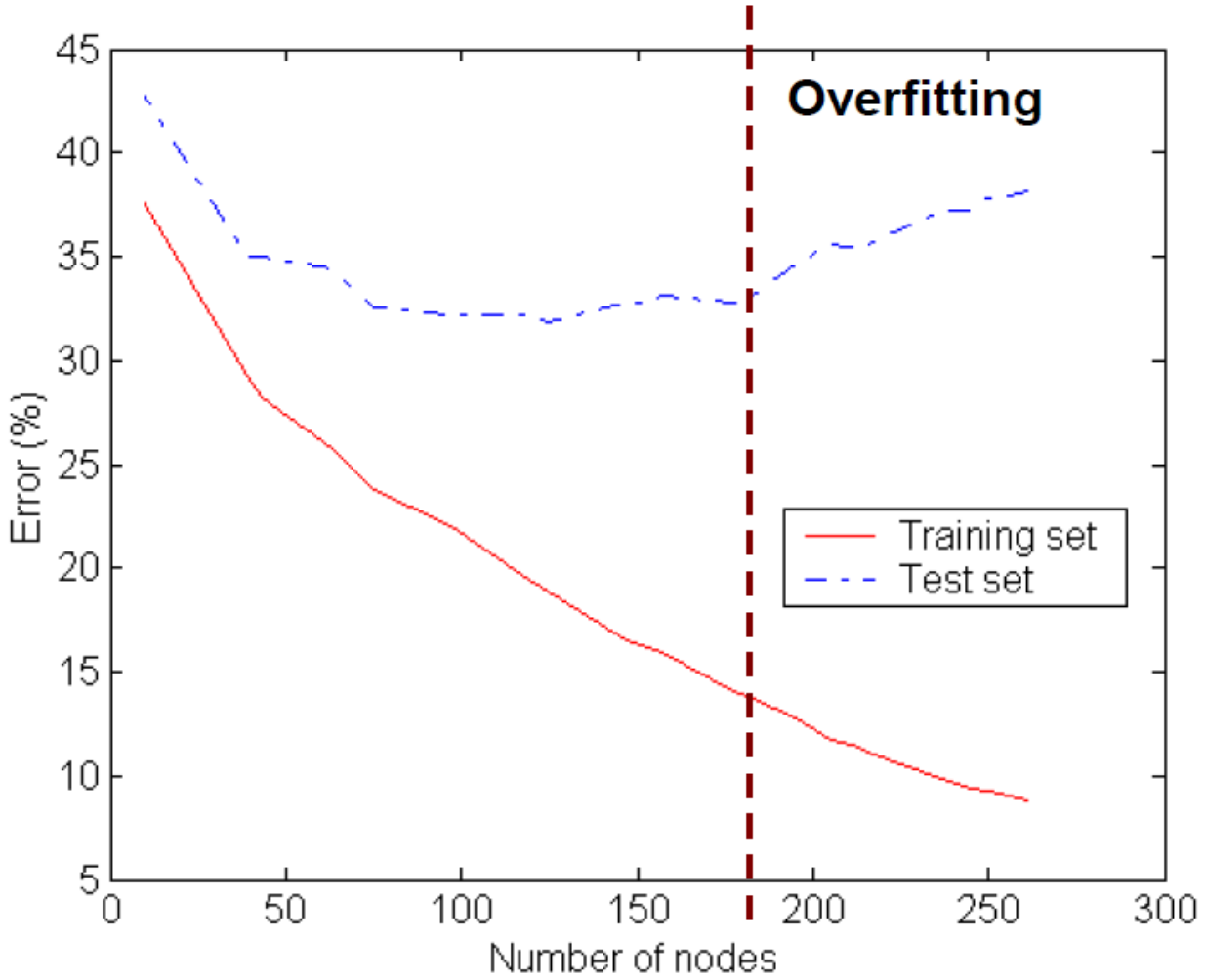
تنظیم ابرپارامترها

- ابرپارامترها

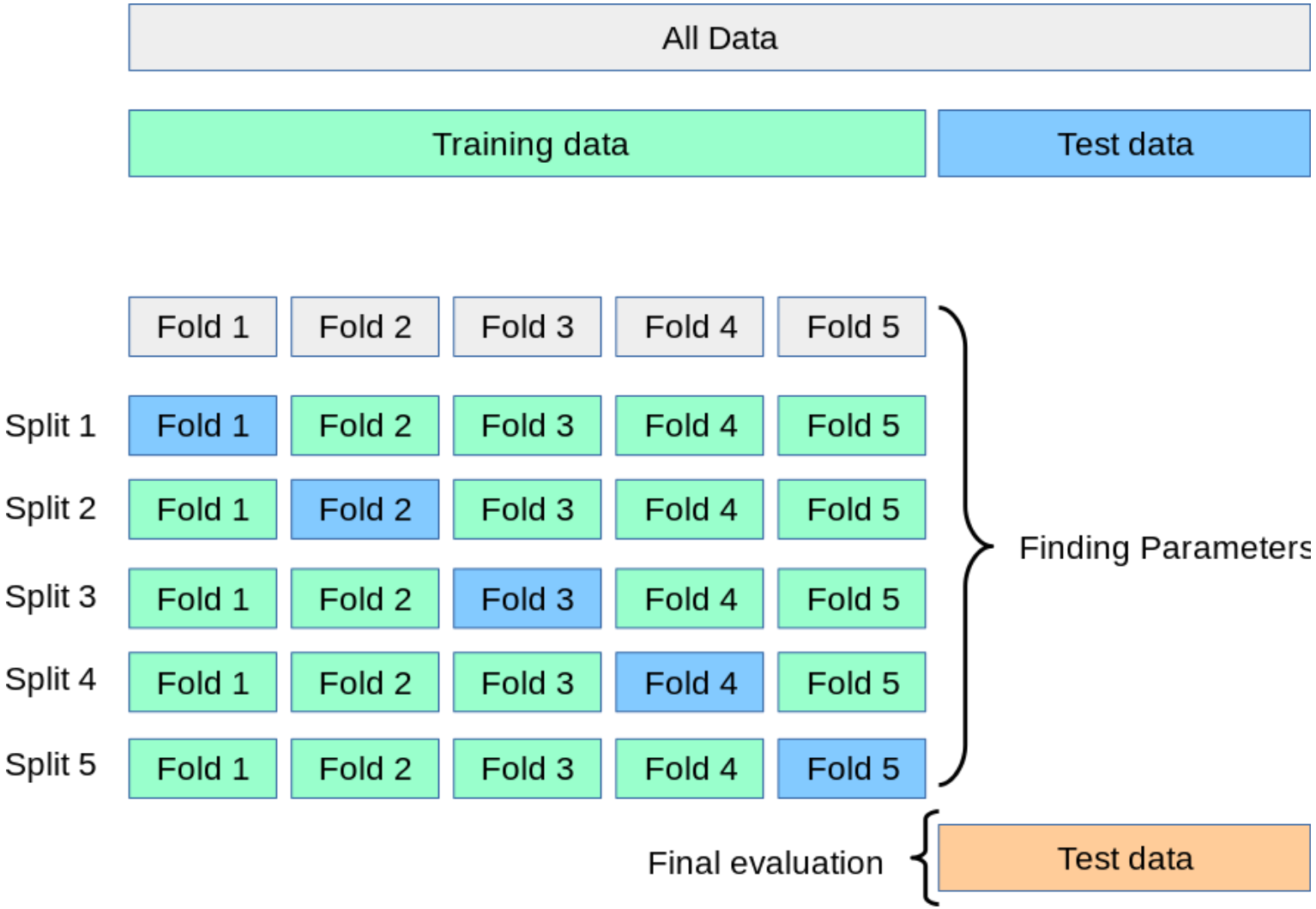


تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

روش Elbow

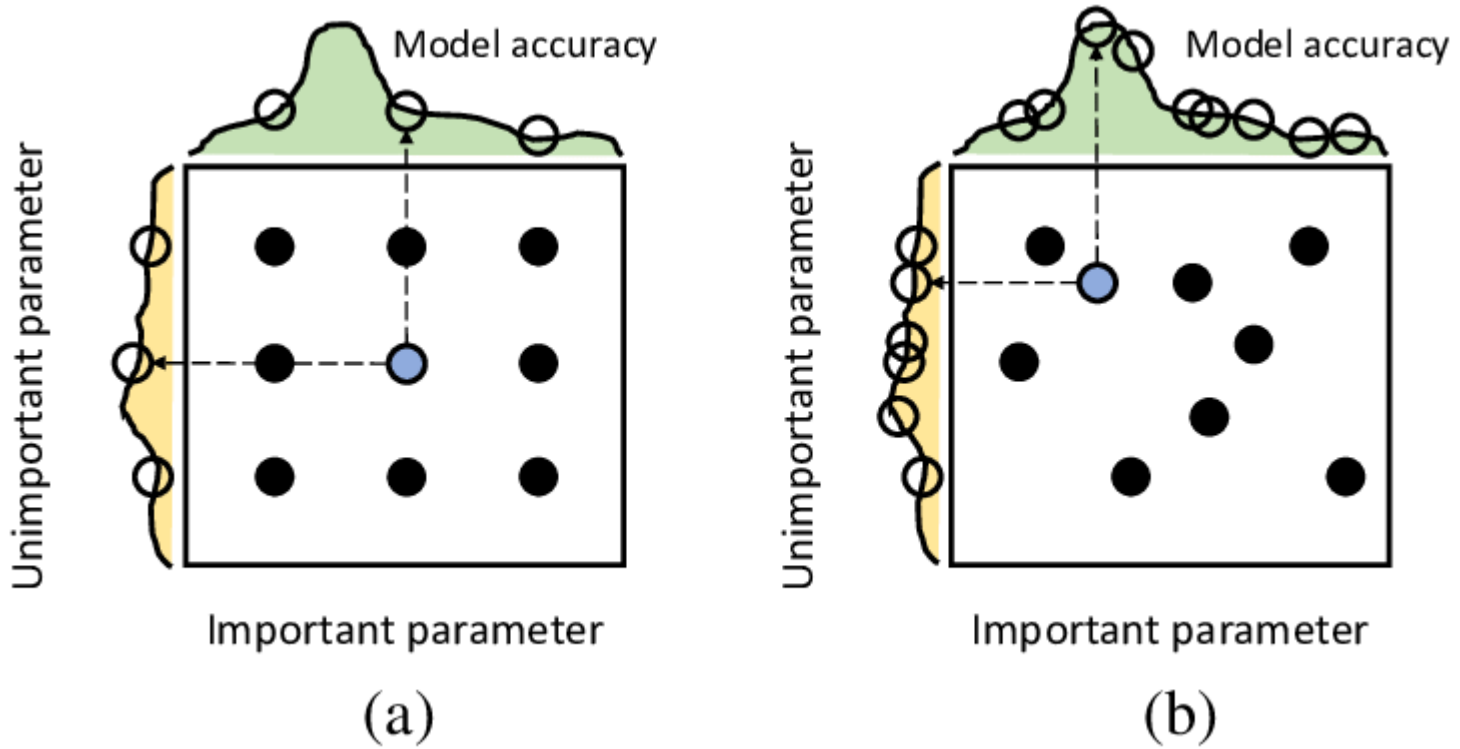


روش Cross validation



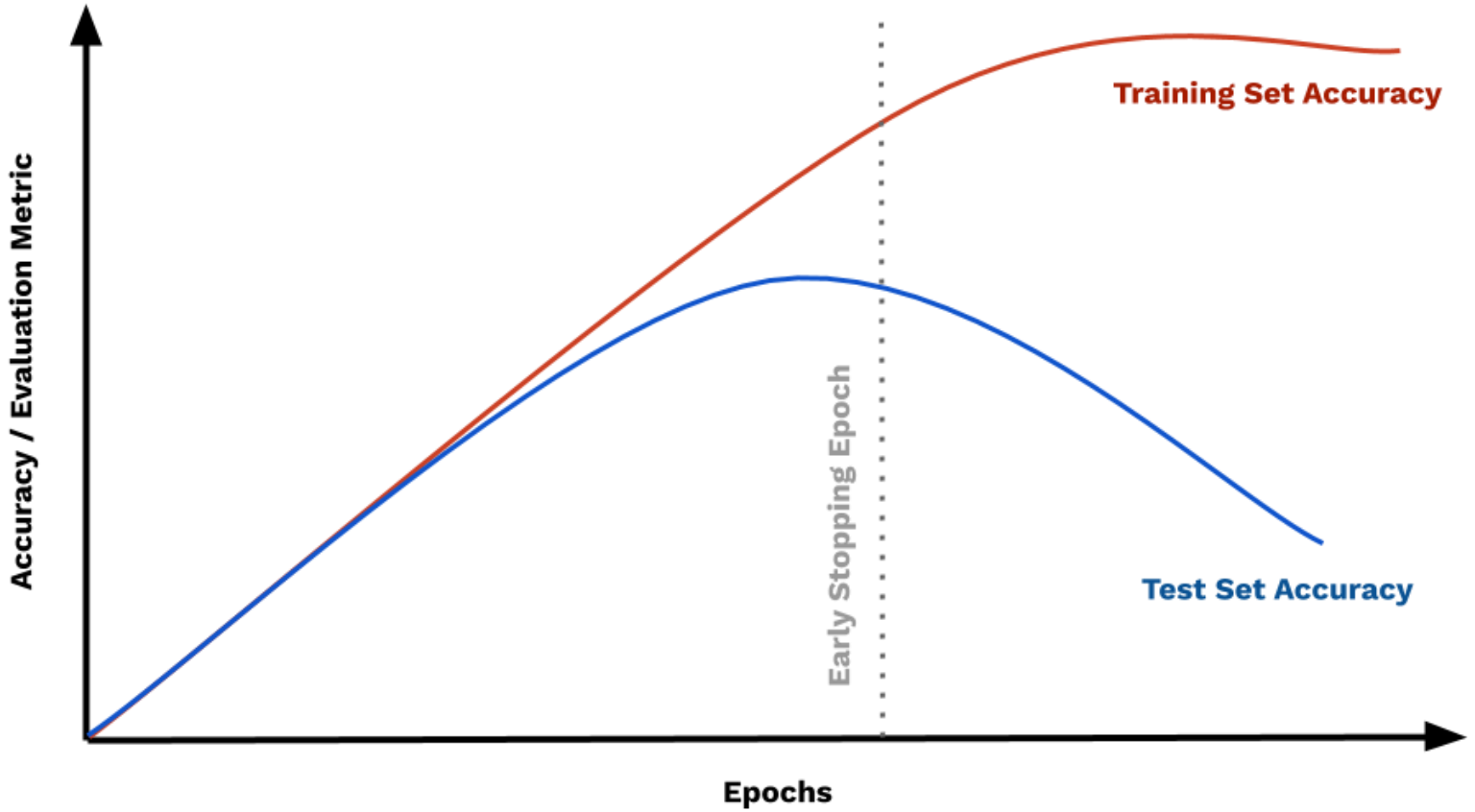
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
گروه دایکه | dayche.com

روش Grid search



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

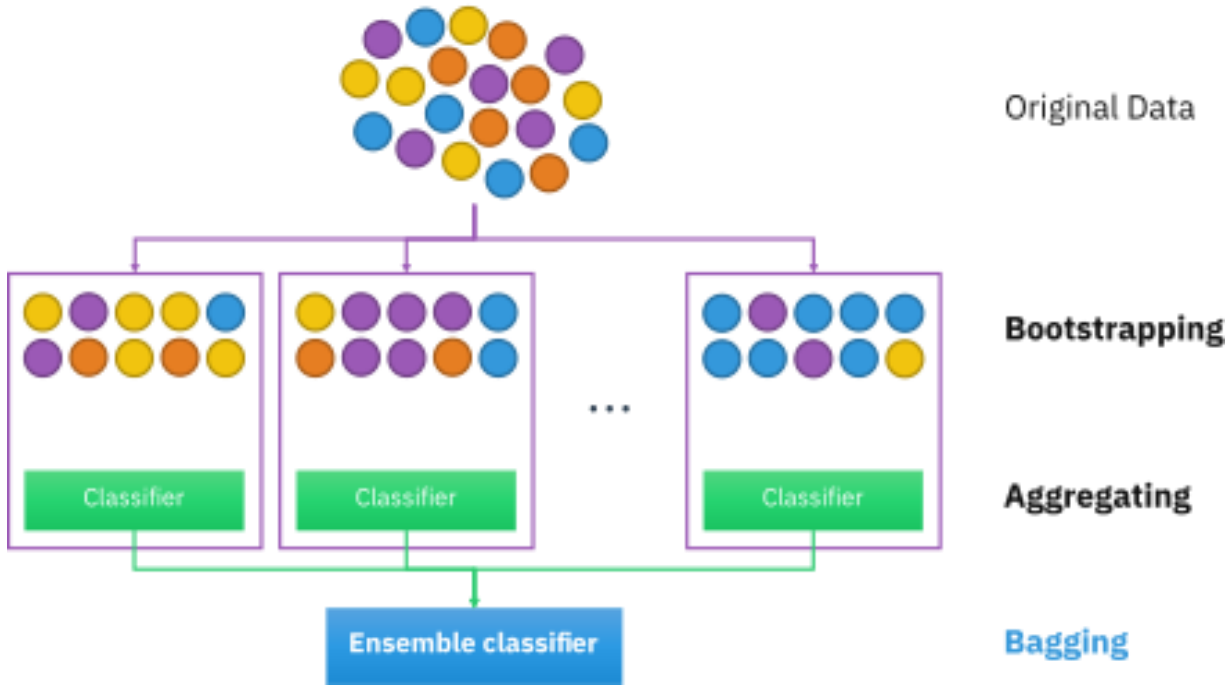
روش Early stopping



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

تجمیع مدل‌ها

- Bagging – Bootstrap aggregation



- مدل‌های مختلف دارای خطای تست یکسان نیستند
- کاربردی از مفهوم امید ریاضی
- در تنظیم‌سازی مدل، توجه بر روی کاهش واریانس بود، در حالی که در رویکرد ترکیب مدل‌ها تمرکز بر روی کاهش بایاس و واریانس به صورت توامان است.

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

تجمیع مدل‌ها

- مدل رگرسیون

$$y_m(x) = h(x) + \epsilon_m(x) \qquad \hat{y}_{com}(x) = \frac{1}{m}\sum_{k=1}^{m} \hat{y}_k(x)\,, \qquad e_m = y_m(x) - h(x)$$

$$E_{av} = \frac{1}{m}\sum_{k=1}^{m} E_x(\epsilon_m(x)^2)$$

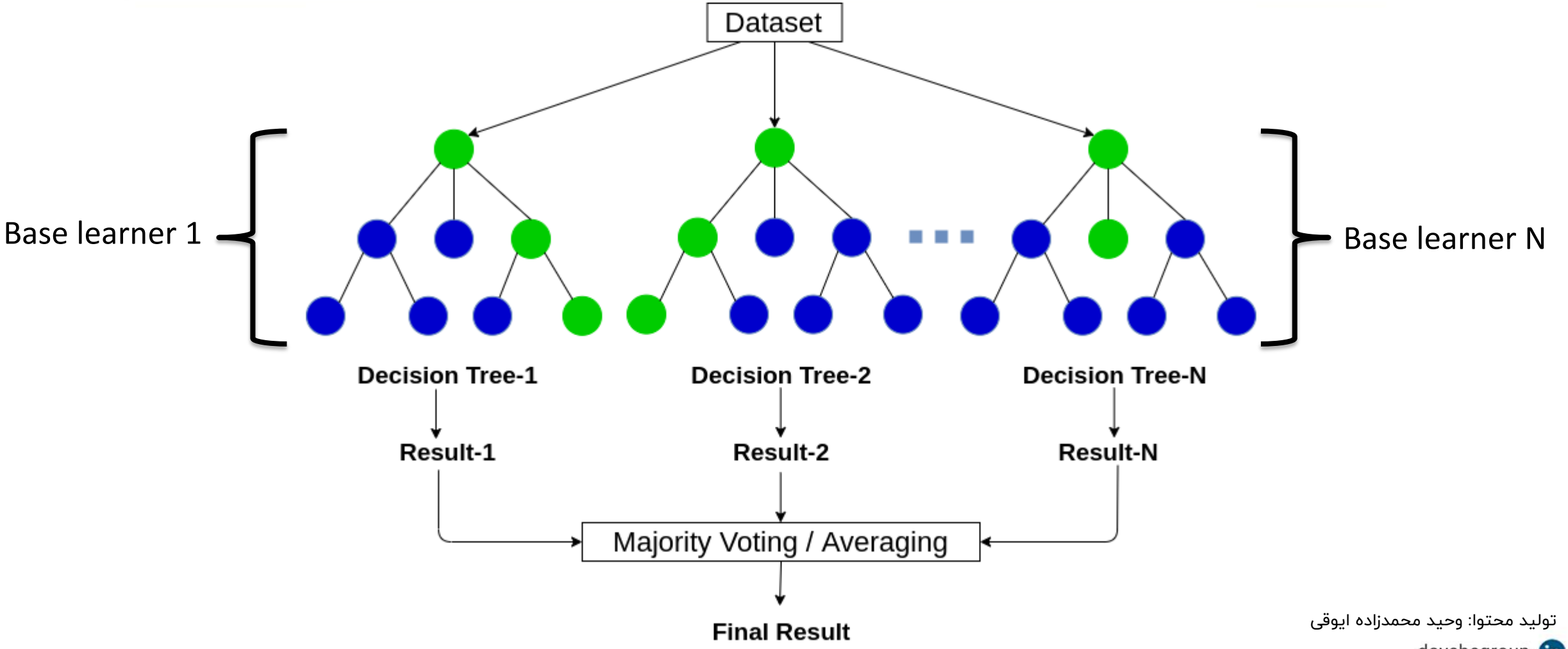
$$E_{com} = E_x\left(\frac{1}{m}\sum_{k=1}^{m} \hat{y}_k(x) - h(x)\right)^2 = E_x\left(\frac{1}{m}\sum_{k=1}^{m} \epsilon_m(x)\right)^2$$

- فرض ناهمبسته بودن بین مدل‌ها مطرح است که ممکن است در واقعیت درست نباشد (در بیشتر موارد درست نیست)، چرا؟
- آموزش مدل‌ها از هم مستقل نباشد.

$$E_{com} = \frac{1}{m} E_{av}$$

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

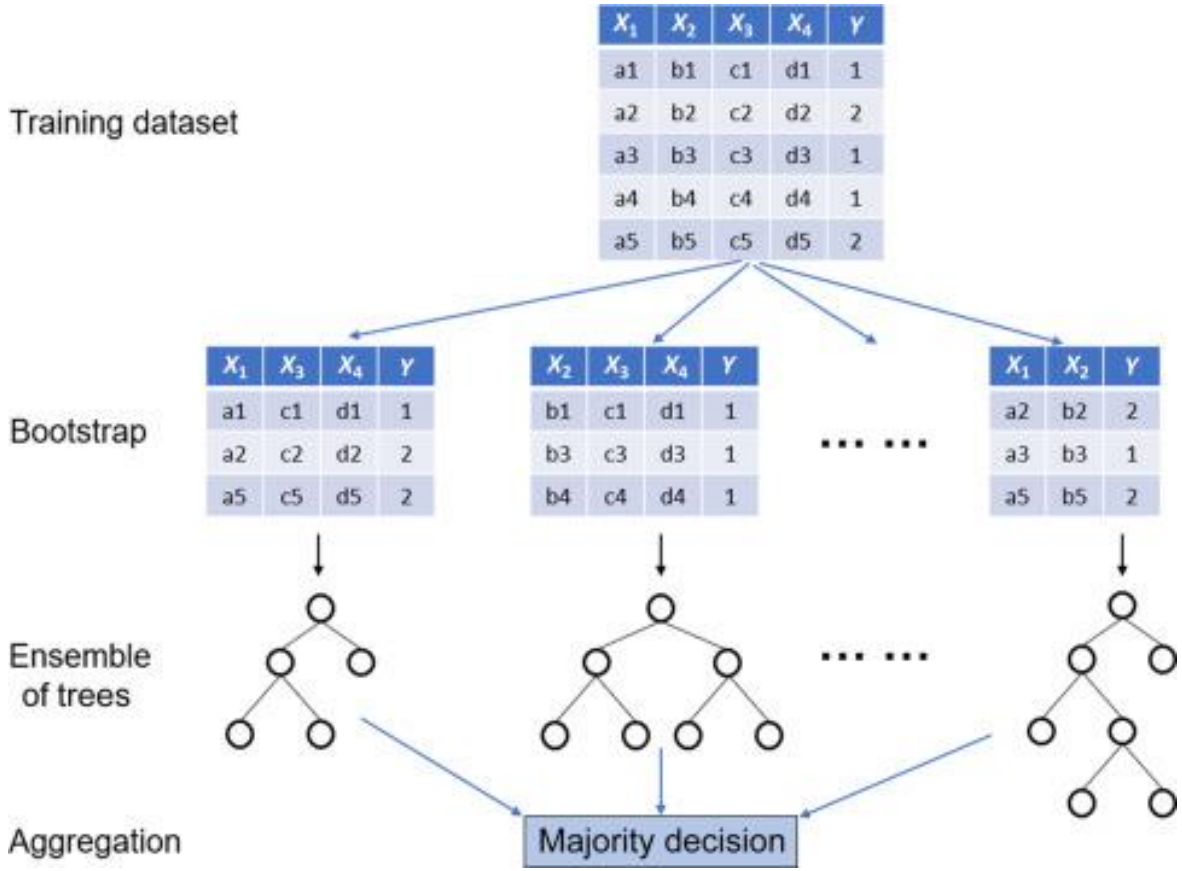
Random forest



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

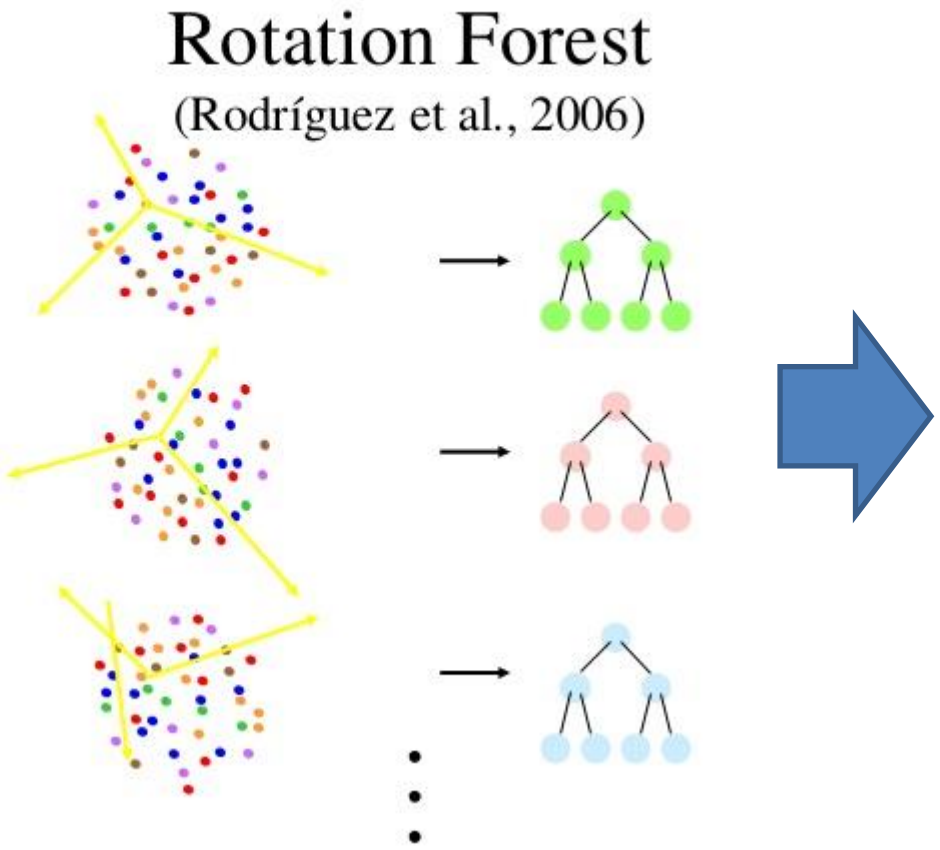
Random forest

انتخاب این مجموعه‌ها تصادفی است و می‌تواند دارای همپوشانی باشد



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Rotation forest

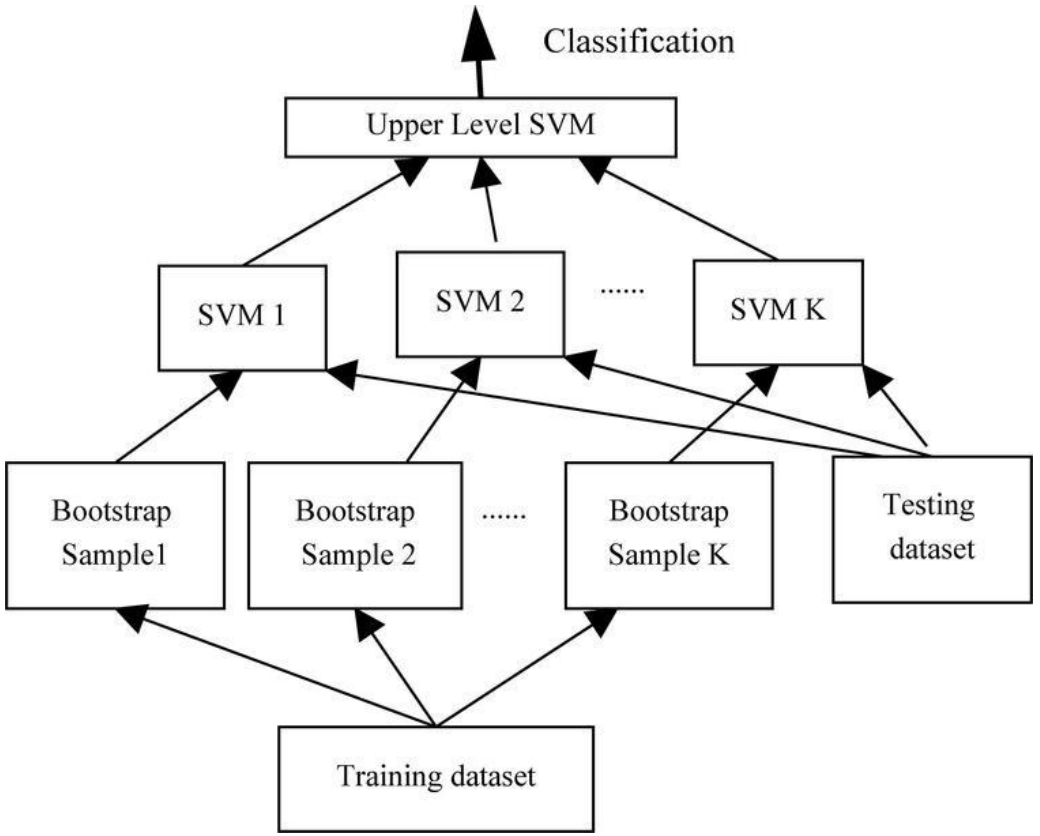


- تقسیم‌بندی تصادفی فضای ویژگی‌ها به چند زیرمجموعه با طول یکسان
- کاهش بعد هر زیر مجموعه
- ذخیره‌سازی ماتریس تبدیل
- مدل‌سازی

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

تجمیع مدل‌های غیر درخت

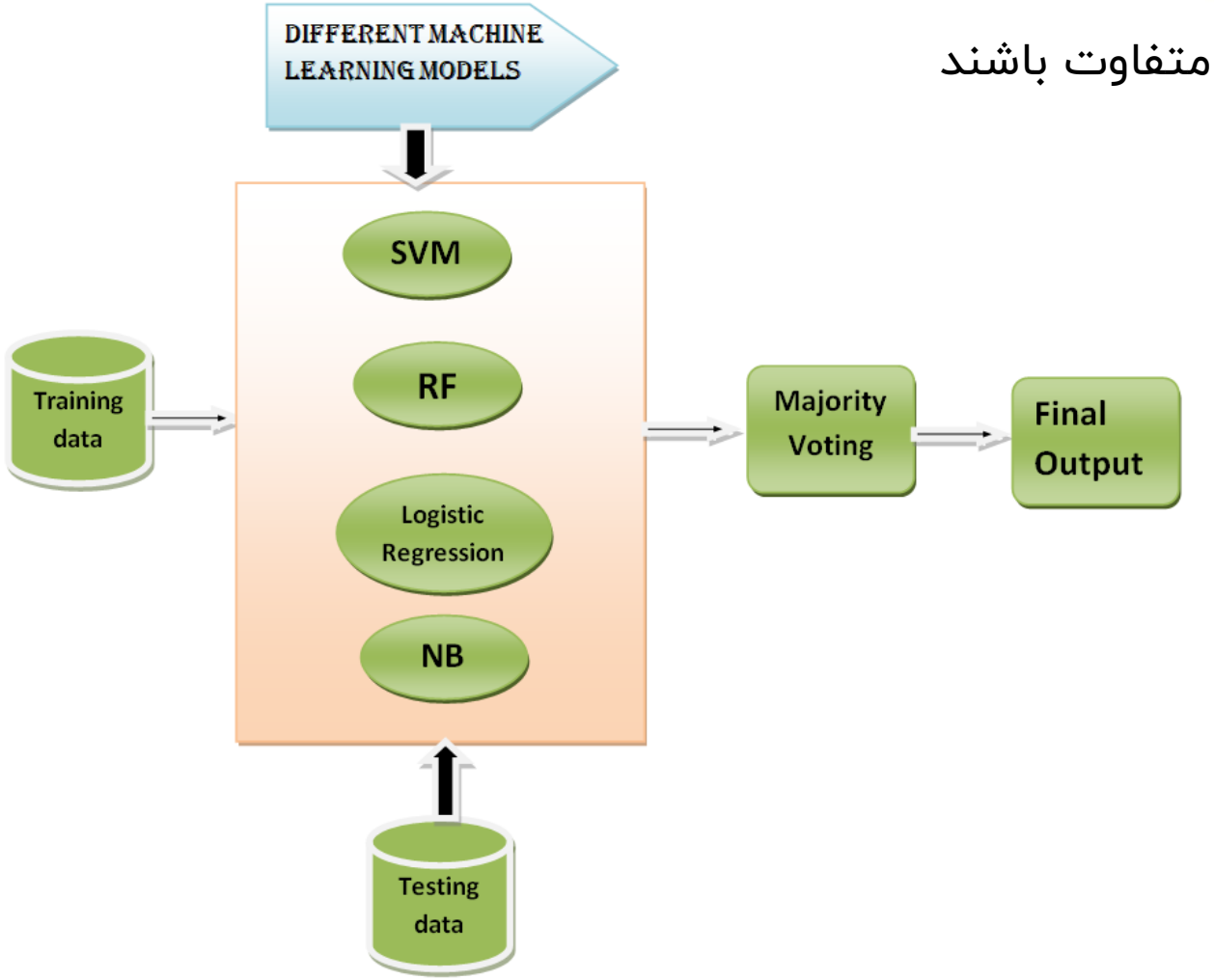
مدل پایه برای دو روش بیان شده درخت تصمیم بود. آیا می‌توان از مدل‌های دیگر نظیر مدل‌های خطی، مدل‌های احتمالاتی، ماشین بردار پشتیبان و یا شبکه‌های عصبی استفاده کرد؟



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

تجمیع مدل‌های متفاوت

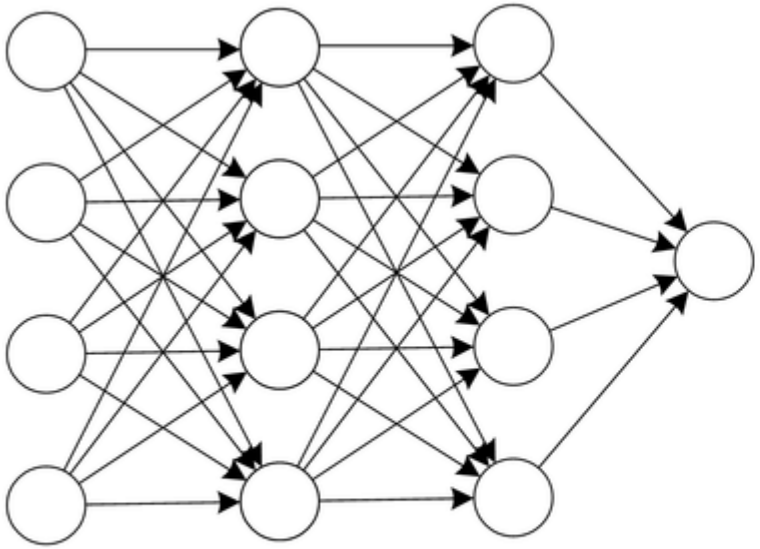
مدل‌های پایه می‌توانند متفاوت باشند



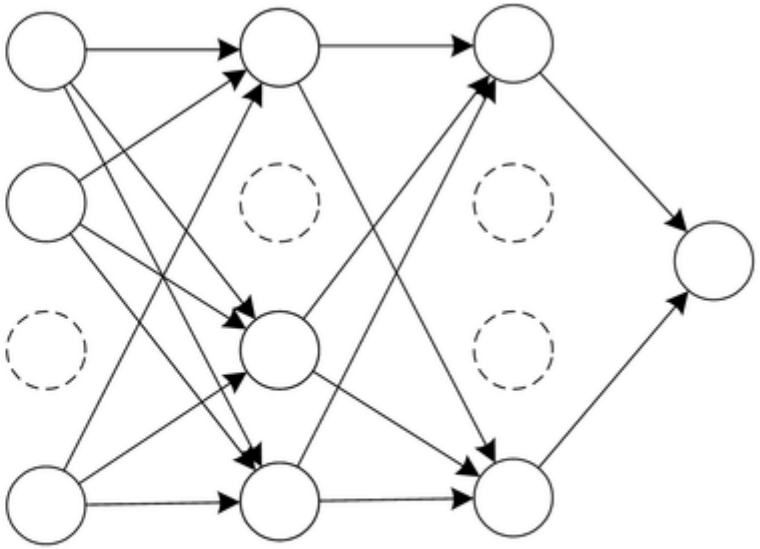
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

رویکرد حذف تصادفی Dropout

- تجمیع مدل – آموزش چندین مدل مختلف
- آموزش چندین شبکه عصبی به طور مجزا هزینه محاسباتی و اجرایی بالایی داردو چه باید کرد؟



(a) Standard Neural Network

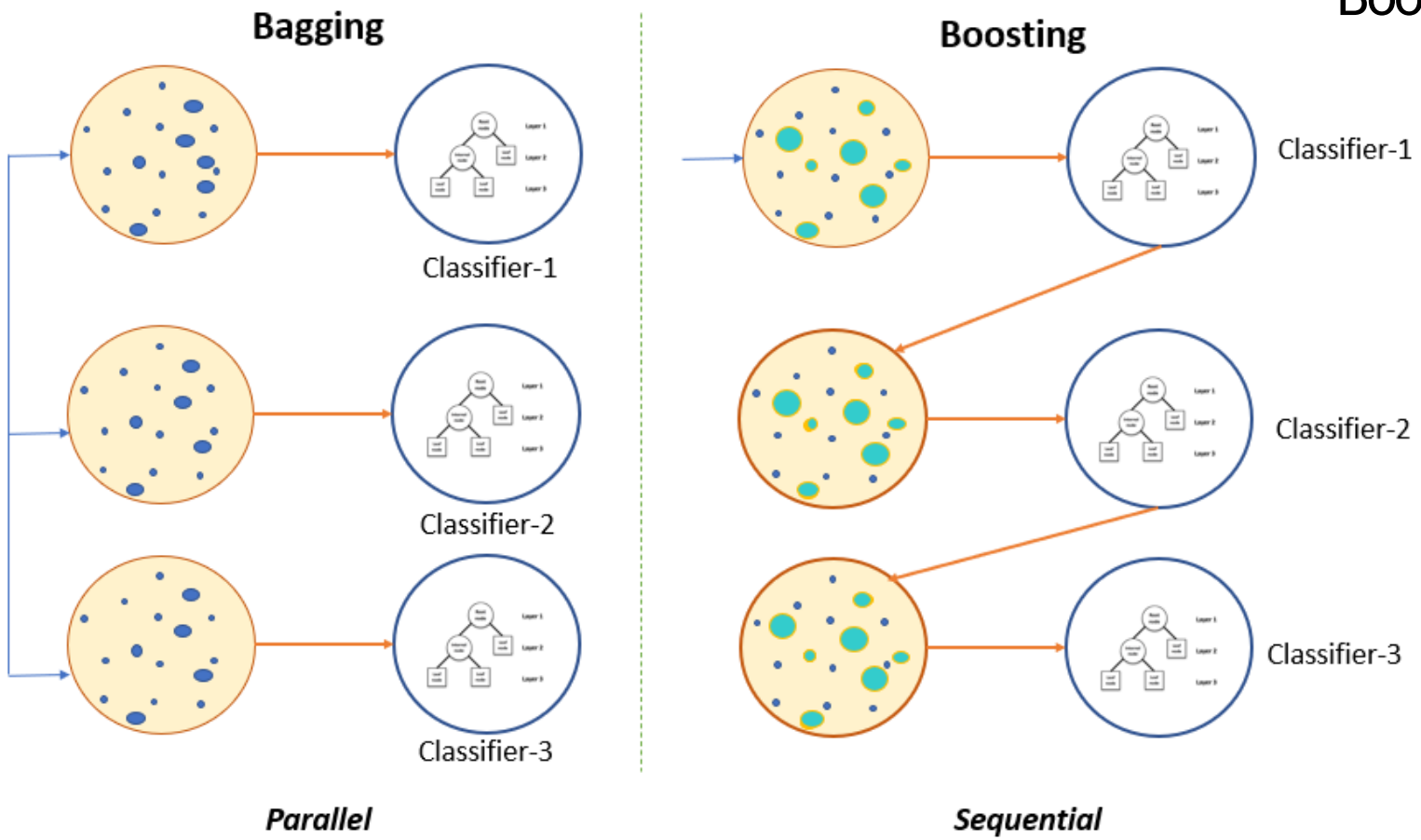


(b) Network after Dropout

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

تجمیع مدل‌ها

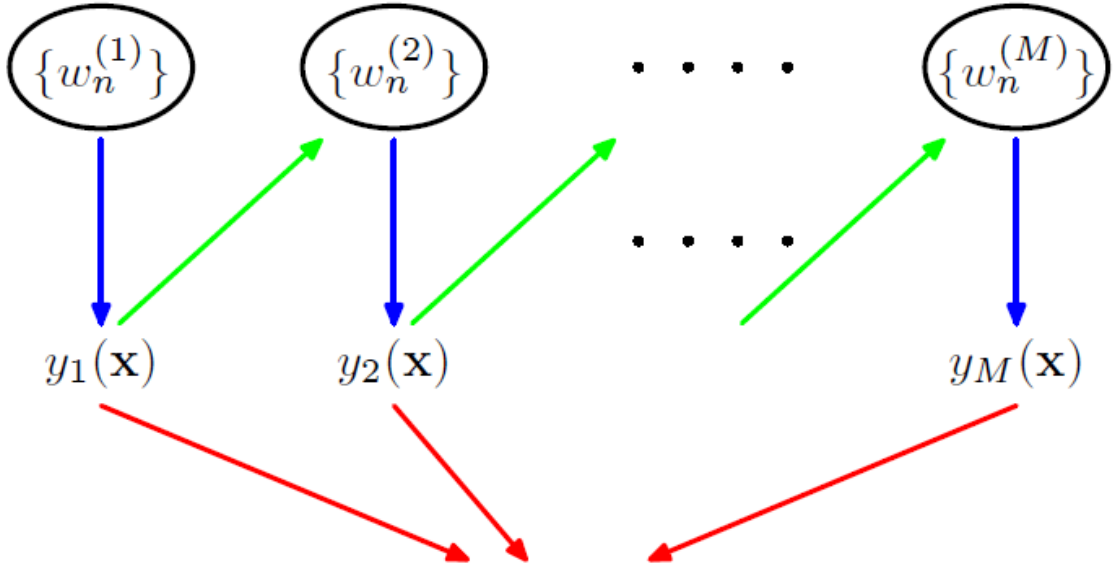
- روش Boosting



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

روش Boosting

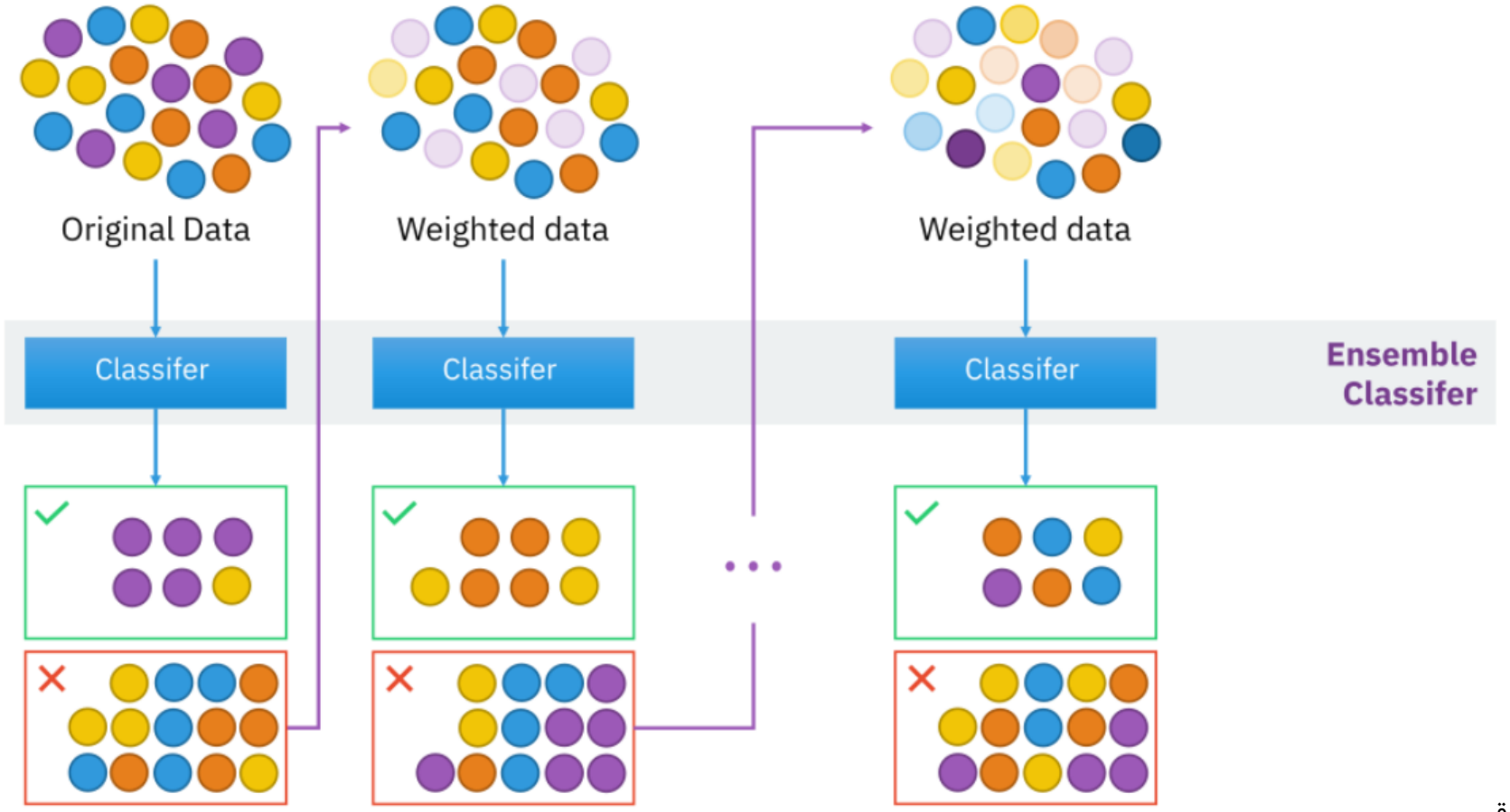
- مجموعه‌ای از مدل‌ها به صورت متوالی و وابسته بهم بر اساس یک وزن مشخصی برای داده‌ها، آموزش می‌بینند.
- نتیجه نهایی به صورت رای‌گیری وزن‌دار بیان می‌شود.



$$Y_M(\mathbf{x}) = \text{sign}\left(\sum_m^M \alpha_m y_m(\mathbf{x})\right)$$

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

روش Adaptive Boosting (AdaBoost)



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Adaptive Boosting (AdaBoost) روش

AdaBoost Training

Input: I (a weak inducer), T (the number of iterations), and S (a training set).

Output: $M_t, \alpha_t; t = 1, \ldots, T$

1: $t \leftarrow 1$
2: $D_1(i) \leftarrow 1/m; i = 1, \ldots, m$
3: **repeat**
4: Build Classifier M_t using I and distribution D_t
5: $\varepsilon_t \leftarrow \sum_{i:M_t(x_i) \neq y_i} D_t(i)$
6: **if** $\varepsilon_t > 0.5$ **then**
7: $T \leftarrow t - 1$
8: exit Loop.
9: **end if**
10: $\alpha_t \leftarrow \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$
11: $D_{t+1}(i) = D_t(i) \cdot e^{-\alpha_t y_t M_t(x_i)}$
12: Normalize D_{t+1} to be a proper distribution.
13: $t \leftarrow t + 1$
14: **until** $t > T$

روش Gradient Boosting Machine (GBM)

ترکیب مدل: $\hat{F}(x) = \sum_m \alpha_m h_m(x)$

$$F_0(x) = \arg\min L(x, y; \theta)$$

$$F_m(x) = \underbrace{F_{m-1}(x)}_{\text{Weak learner}} + \underbrace{h_m(x)}_{\text{Variation of weak learner}} = y$$

قصد داریم پیش‌بینی کننده ضعیف را طوری تغییر دهیم تا وضعیت پیش‌بینی مدل بهبود پیاده کند

$$L\big(y, F_m(x)\big) = L\big(y, F_{m-1}(x) + h_m(x)\big) = ?$$

$$L\big(y, F_{m-1}(x) + h_m(x)\big) = L\big(y, F_m(x)\big) + \nabla L_{F_{m-1}}\big(y, F_{m-1}(x)\big) h_m(x)$$

هدف

$$L\big(y, F_{m-1}(x) + h_m(x)\big) < L\big(y, F_m(x)\big) \rightarrow h_m(x) = -\alpha_m \nabla L_{F_{m-1}}\big(y, F_{m-1}(x)\big)$$

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
گروه دایکه | dayche.com

Gradient Boosting Machine (GBM) روش

GBM Training

Input: A base regression algorithm - I, number of iterations T, the training set, $S = \{(x_i, y_i)\}_{i=1}^{m}$, and a differentiable loss function $L(y, F(x))$

1: Initialize model with a constant value: $F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$

2: $j \leftarrow 1$

3: **repeat**

4: For $i = 1, \ldots, m$, compute pseudo-residuals:
$r_{ij} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{j-1}(x)}$

5: Construct regression model h_j using I using the training set $\{(x_i, r_{ij})\}_{i=1}^{m}$.

6: Find multiplier γ_j by performing line search on the following one-dimensional optimization problem:
$\gamma_j = \arg\min_{\gamma} \sum_{i=1}^{m} L\left(y_i, F_{j-1}(x_i) + \gamma h_j(x_i)\right)$.

7: Update the model: $F_j(x) = F_{j-1}(x) + \gamma_j h_j(x)$.

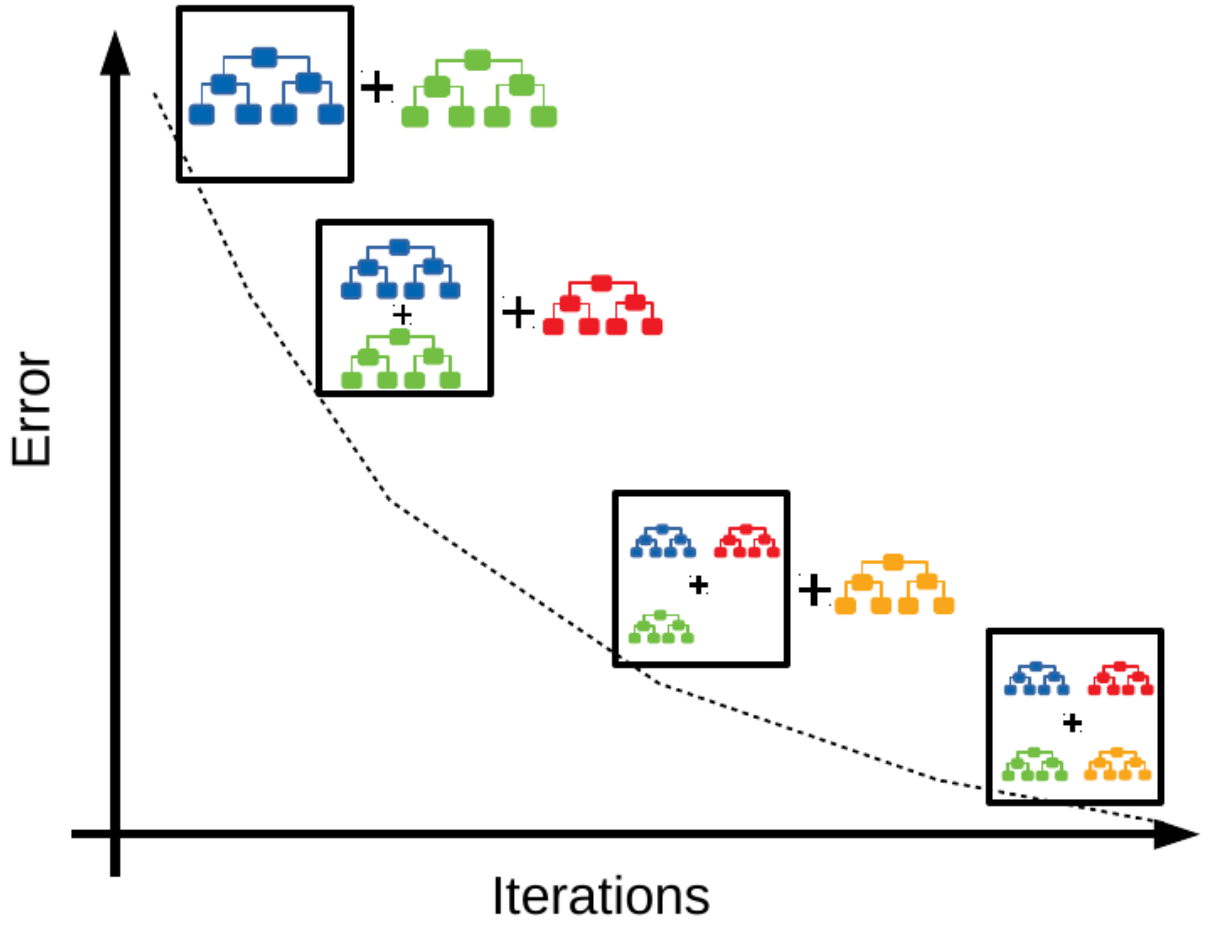
8: $j \leftarrow j + 1$

9: **until** $j > T$

10: Output $F_M(x)$

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Gradient Boosting Machine (GBM) روش



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
گروه دایکه | dayche.com

ارزیابی مدل

- تعریف مسئله
  - چگونه می‌توان فهمید که مدل‌های یادگیری ماشین عملکرد مناسبی دارند؟
  - هدف چیست؟

همواره می‌تواند یک معیار ارزیابی باشد $y = -\log P(y|x)$

- دسته‌بندی

		Target Value	
		Pos	Neg
Predicted Value	Pos	**TP** (True Positive)	**FP** (False Positive)
	Neg	**FN** (False Negative)	**TN** (True Negative)

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

ارزیابی مدل‌های دسته‌بند

- صحت
  - چند درصد از برچسب‌ها به درستی تشخیص داده شدند؟

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

		Target Value	
		Pos	**Neg**
Predicted Value	**Pos**	**TP** (True Positive)	**FP** (False Positive)
	Neg	**FN** (False Negative)	**TN** (True Negative)

- اگر مسئله دارای تعادل باشد، این معیار برای ارزیابی مناسب است
  - تعداد داده‌های هر کلاس تفاوت معنادار نداشته باشد

ارزیابی مدل‌های دسته‌بند

- صحت
  - چند درصد از برچسب‌ها به درستی تشخیص داده شدند؟
  - اگر مسئله دارای تعادل باشد، این معیار برای ارزیابی مناسب است
  - تعداد داده‌های هر کلاس تفاوت معنادار نداشته باشد

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

- دقت
  - چند درصد از داده‌های مثبت واقعا مثبت هستند.

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

		Target Value: Pos	Target Value: Neg
Predicted Value	Pos	**TP** (True Positive)	**FP** (False Positive)
Predicted Value	Neg	**FN** (False Negative)	**TN** (True Negative)

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

ارزیابی مدل‌های دسته‌بند

- Recall
  - چند درصد داده‌های مثبت به درستی مثبت تشخیص داده شده‌اند.

$$Recall = \frac{TP}{TP + FN}$$

- F1 score
  - میانگین هارمونیک – زمانی که داده‌ها دارای تعادل نیستند این معیار از اعتبار بیشتری نسبت به صحت برخوردار است.

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

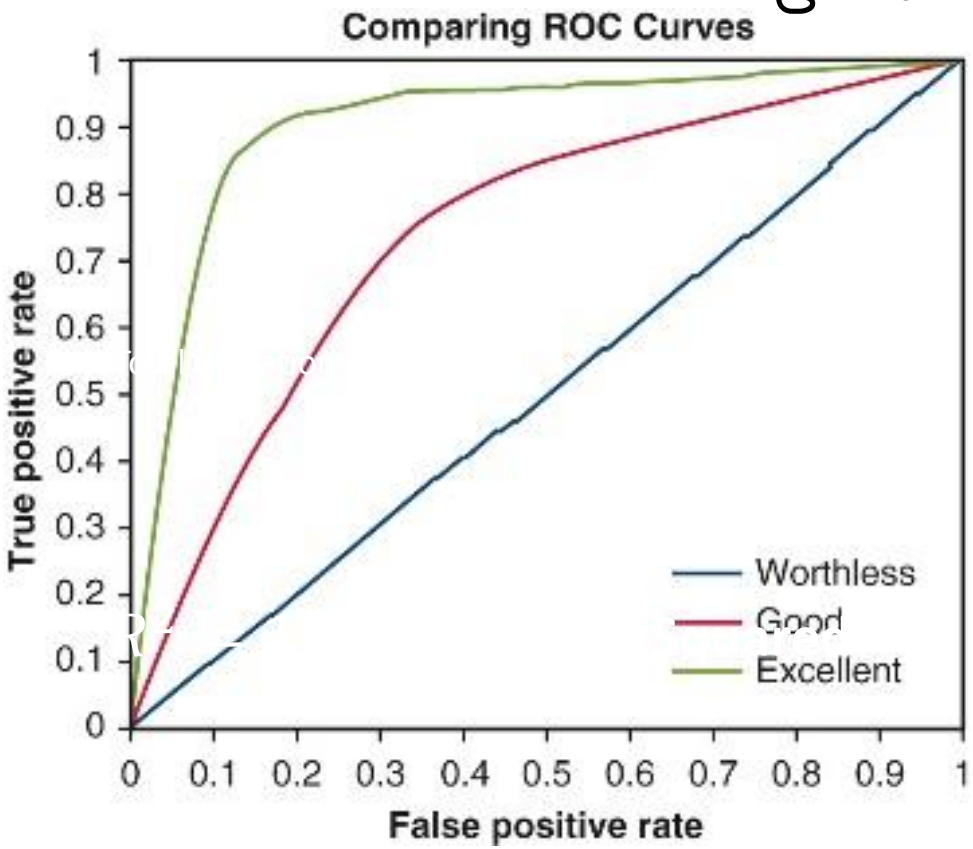
$$F_\beta = (1 + \beta^2) * \frac{Precision * Recall}{(\beta^2 * Precision) + Recall}$$

ارزیابی مدل‌های دسته‌بند

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

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

- منحنی ROC



سطح زیر این منحنی باید بیشینه باشد

- کاربرد در تشخیص نابهنجاری

ارزیابی مدل‌های رگرسیونی

- تعریف مسئله

$$y = f(x, w) + n = \phi(x)w + n$$ True model

$$\hat{y} = g(x, \hat{w})$$ Prediction model

$$SST = |y - \bar{y}|^2$$

$$SSR = |\hat{y} - \bar{y}|^2$$

$$SSE = |y - \hat{y}|^2$$

$$\underbrace{SST}_{\text{Total deviation}} = \underbrace{SSR}_{\text{Regression deviation}} + \underbrace{SSE}_{\text{Model deviation}}$$

$$R^2 = 1 - \frac{SSE}{SST}$$ **R Squared**

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
گروه دایکه | dayche.com