درخت تصمیم

Decision Tree

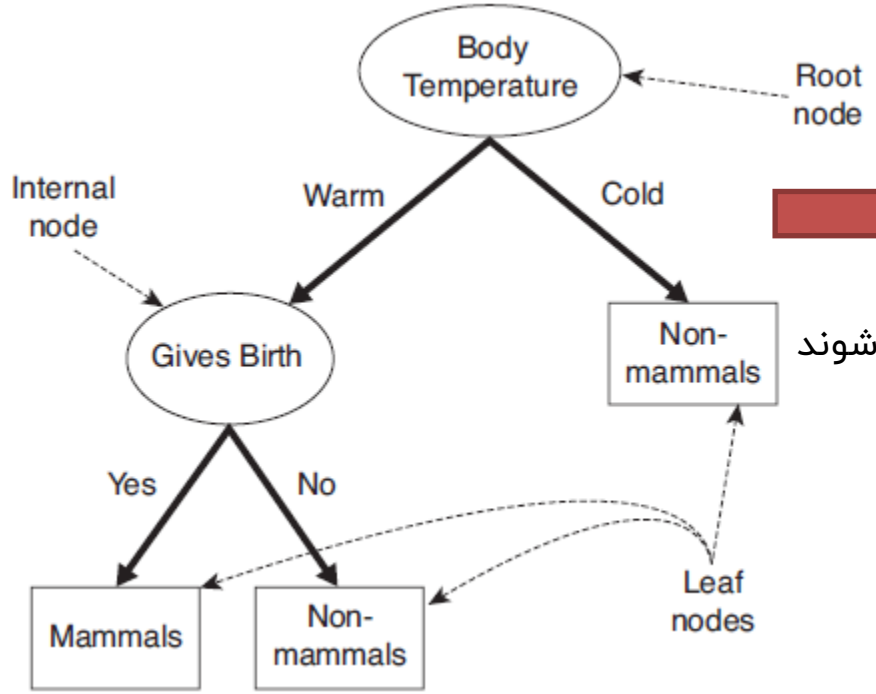
dayche.com . گروه دایکه

دای‌کِه DAYCHE

درخت تصمیم

- درخت تصمیم
  - تصمیم‌گیری بر اساس مجموعه‌ای قوانین – پرسش و پاسخ

مسئله: تمایز پستاندار از غیرپستاندار



اگر حیوان خونسرد است، پستاندار نیست

به ازای هر سوال بخشی از رکوردها تعیین تکلیف می‌شوند

مجموعه متغیرهای در دست
- دمای بدن
- نوع زاد و ولد

تولید محتوا: وحید محمدزاده ایوقی

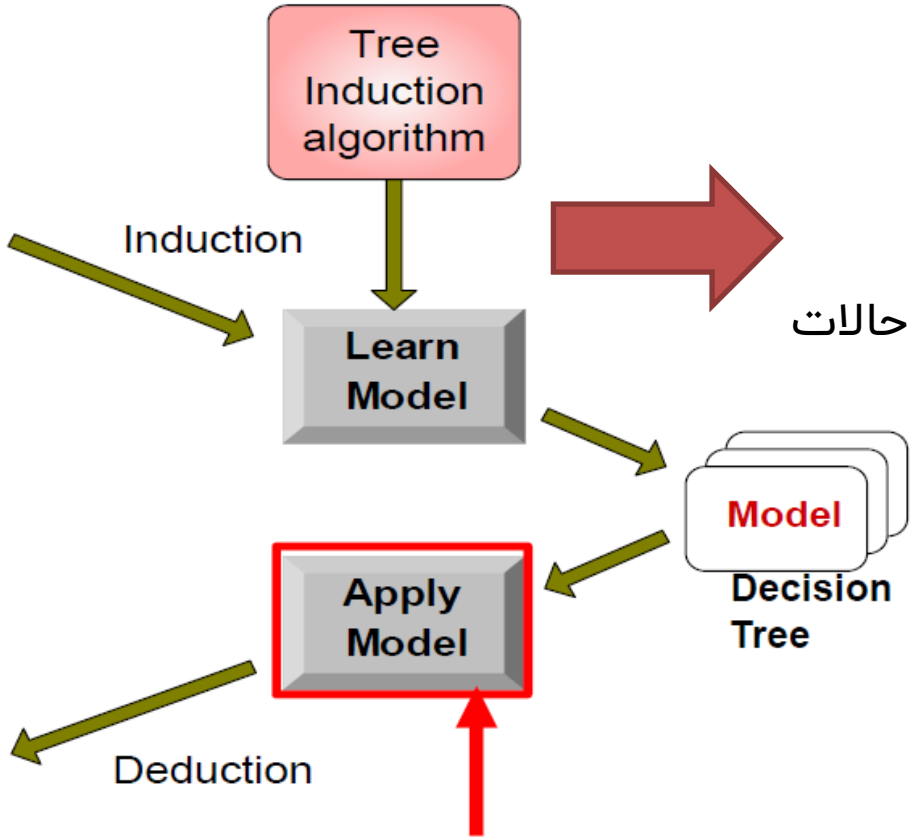
فرآیند توسعه درخت

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

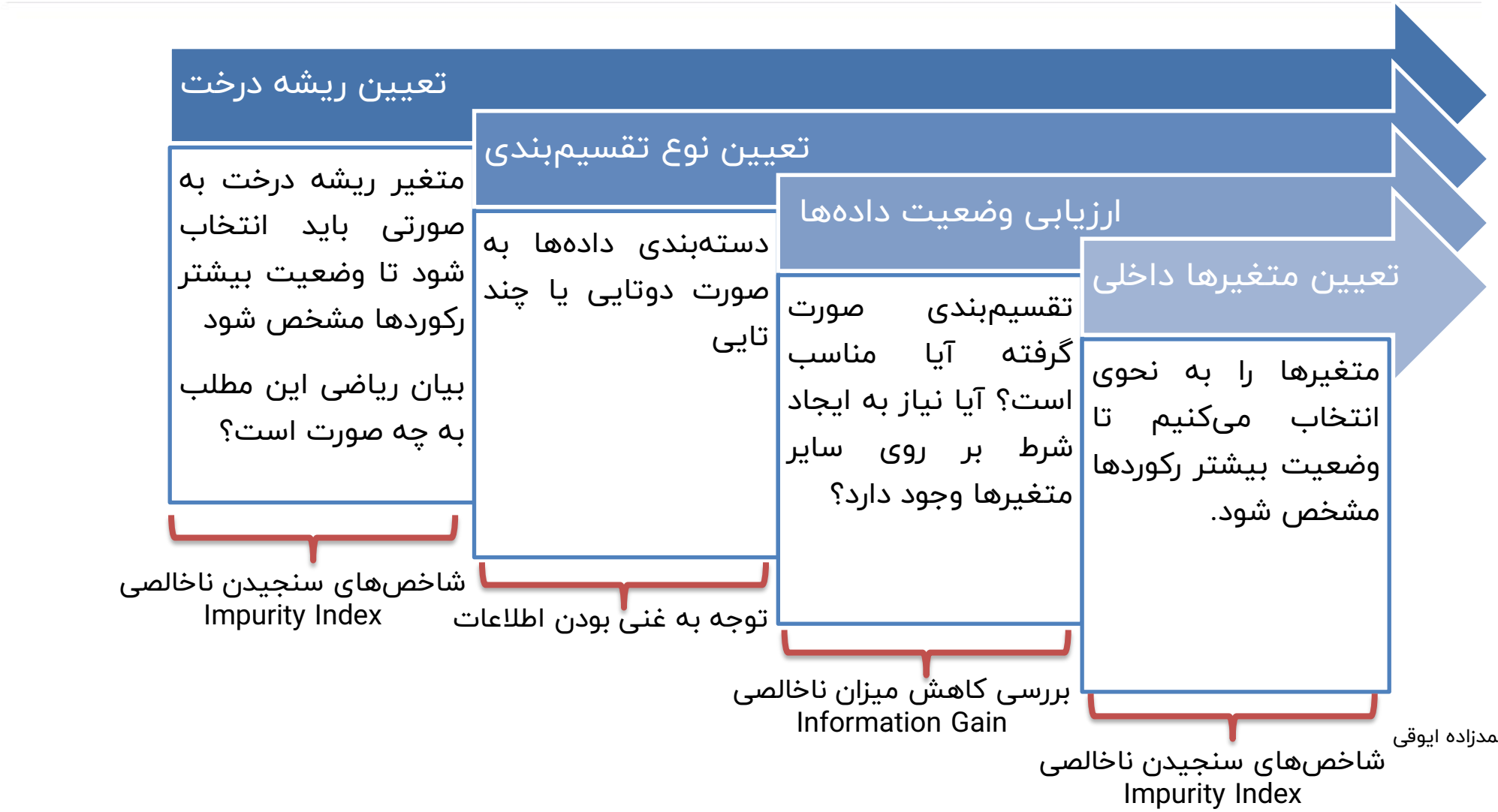
Test Set



- هدف – با حداقل سوال ممکن بتوان تصمیم‌گیری کرد
- معیار مناسب برای کاهش تعداد حالات بررسی چیست؟

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

فرآیند توسعه درخت

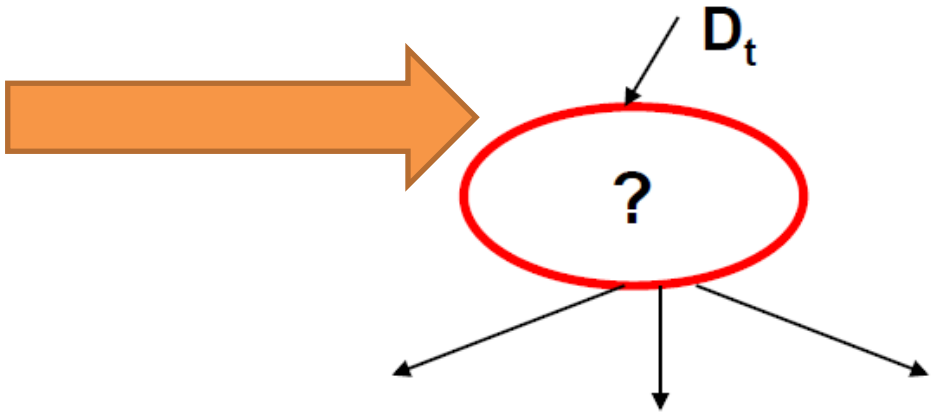


تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

استراتژی ساخت درخت تصمیم

- الگوریتم Hunt
- اگر D_t شامل رکوردهایی باشد که همگی دارای برچسب y_t باشد، آنگاه این گره، برگ درخت خواهد بود.
- اگر D_t شامل هیچ رکوردی نباشد، آنگاه این گره، برگ درخت خواهد بود. (برچسب آن چه خواهد بود؟)
- اگر D_t شامل رکوردهایی از کلاس‌هایی مختلف باشد، تقسیم‌بندی داده‌ها با شروط دیگری بر روی متغیرهای دیگر ادامه خواهد داشت.

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



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

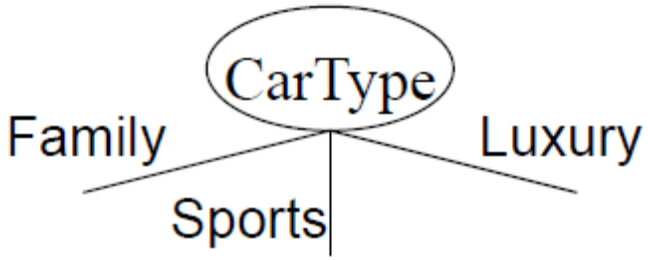
تعیین نوع تقسیم‌بندی

- به نوع متغیر بستگی دارد
  - اسمی Nominal
  - ترتیبی Ordinal
  - پیوسته Continues
- تقسیم‌بندی کردن – خروجی شرط
  - دوگانه
  - چندگانه

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

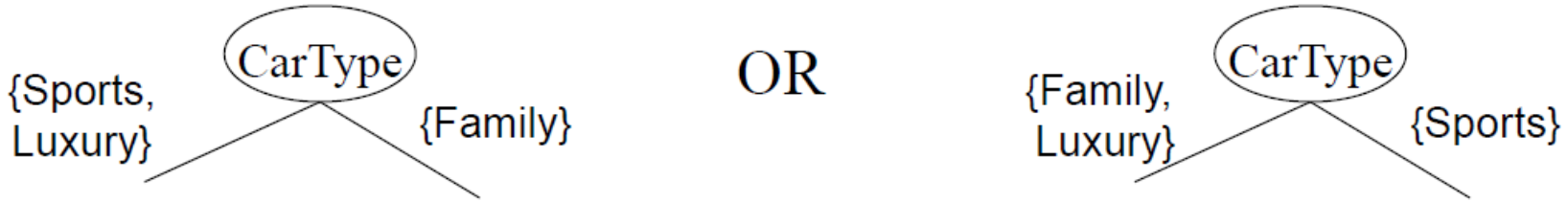
تقسیم‌بندی متغیرهای اسمی

- تقسیم‌بندی چندگانه



تقسیم براساس تعداد سطوح مختلف یکتا

- تقسیم‌بندی دوگانه

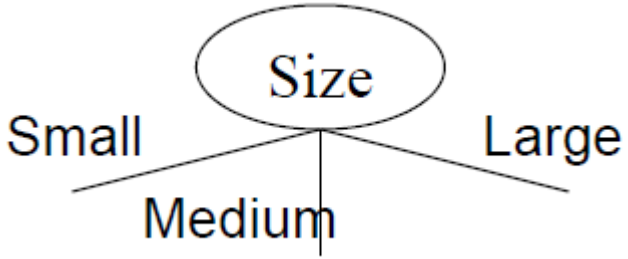


تقسیم براساس گروه‌های مختلف

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

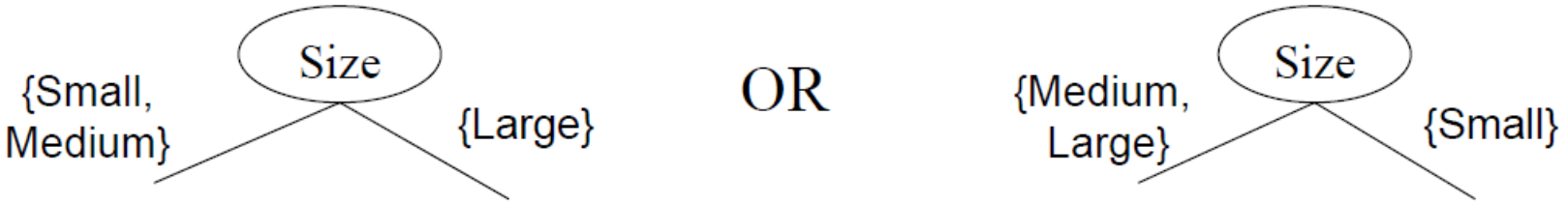
تقسیم‌بندی متغیرهای ترتیبی

- تقسیم‌بندی چندگانه



تقسیم براساس تعداد سطوح مختلف یکتا

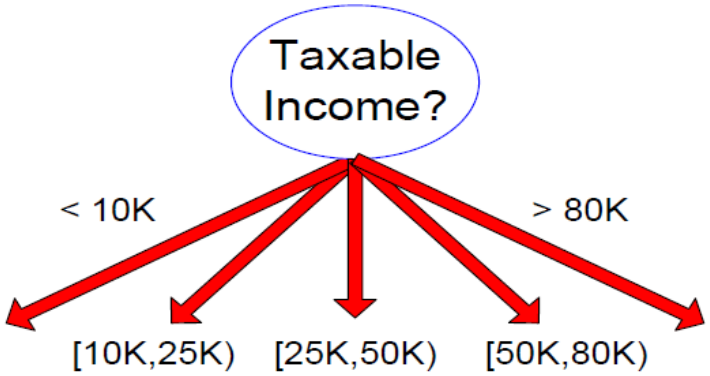
- تقسیم‌بندی دوگانه



تقسیم براساس گروه‌های مختلف به صورت **منطقی**

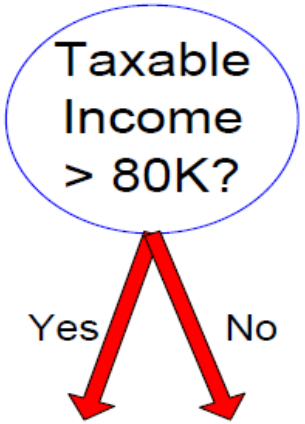
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

تقسیم‌بندی متغیرهای پیوسته



- تقسیم‌بندی چندگانه – گسسته سازی بازه‌ای

تقسیم براساس تعداد حالت مختلف



- تقسیم‌بندی دوگانه

تقسیم براساس یک مقدار آستانه

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

تعیین متغیرهای شرط

- هدف
  - قبل و بعد از تقسیم‌بندی کردن داده‌ها بر اساس یک متغیر، تصمیم‌گیری ساده‌تر شده باشد.
  - تصمیم‌گیری ساده هم ارز رسیدن به یک پراکندگی همگن است.
- معیارهای تقسیم‌بندی کردن
  - شاخص جینی Gini index
  - آنتروپی Entropy
  - خطای دسته‌بندی Classification error

بعد از تقسیم‌بندی توزیع داده‌های دو کلاس ناهمگن است

C0: 5
C1: 5

C0: 9
C1: 1

بعد از تقسیم‌بندی توزیع داده‌های دو کلاس همگن است

بهترین تقسیم‌بندی منجر به بیشتر همگنی بین داده‌ها خواهد شد

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

ضریب جینی

- تعریف ریاضی

$$\text{Gini(t)} = 1 - \sum_j [P(j|t)]^2$$

$P(j|t)$ بسامد نسبی کلاس j در گره t

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Gini = 1 – P(C1)² – P(C2)² = 1 – 0 – 1 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Gini = 1 – (1/6)² – (5/6)² = 0.278

C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

Gini = 1 – (2/6)² – (4/6)² = 0.444

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

ضریب جینی

- از این معیار در الگوریتم‌های CART, SLIQ, SPRINT استفاده می‌شود.
- وقتی که گره p به k دسته تقسیم‌بندی می‌شود، ضریب جینی بعد از تقسیم‌بندی به صورت زیر محاسبه می‌شود:

$$GINI_{split}(t) = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

- n_i تعداد نمونه‌های موجود در گره i
- n تعداد نمونه‌های قبل از تقسیم‌بندی در گره t

ضریب جینی

محاسبه ضریب جینی برای یک متغیر اسمی

قبل از تقسیم بندی کردن

	Parent
C1	6
C2	6
Gini = 0.500	

$$GINI_{befor}(B) = 1 - \frac{1}{4} - \frac{1}{4} = 0.5$$

بعد از تقسیم بندی کردن

	N1	N2
C1	5	1
C2	2	4
Gini=0.333		

$$GINI(N_1) = 1 - \left(\frac{5}{7}\right)^2 - \left(\frac{2}{7}\right)^2 = 0.4082$$

$$GINI(N_2) = 1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2 = 0.3200$$

$$GINI_{after}(B) = \frac{7}{12} GINI(N_1) + \frac{5}{12} GINI(N_2) = 0.3715$$

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

ضریب جینی

محاسبه ضریب جینی برای یک متغیر ترتیبی

	CarType		
	Family	Sports	Luxury
C1	1	2	1
C2	4	1	1
Gini	0.393		

تقسیم‌بندی چند گانه

	CarType	
	{Sports, Luxury}	{Family}
C1	3	1
C2	2	4
Gini	0.400	

	CarType	
	{Sports}	{Family, Luxury}
C1	2	2
C2	1	5
Gini	0.419	

تقسیم‌بندی دو گانه

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

ضریب جینی

محاسبه ضریب جینی برای یک متغیر پیوسته

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No
Taxable Income										
	60	70	75	85	90	95	100	120	125	220
	55	65	72	80	87	92	97	110	122	172	230
	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >
Yes	0 3	0 3	0 3	0 3	1 2	2 1	3 0	3 0	3 0	3 0	3 0
No	0 7	1 6	2 5	3 4	3 4	3 4	3 4	4 3	5 2	6 1	7 0
Gini	0.420	0.400	0.375	0.343	0.417	0.400	*0.300*	0.343	0.375	0.400	0.420

تقسیم‌بندی دوگانه

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

آنتروپی

- تعریف ریاضی

$$Entopy(t) = -\sum_j P(j|t) \log P(j|t)$$

$P(j|t)$ بسامد نسبی کلاس j در گره t

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Entropy = – 0 log 0 – 1 log 1 = – 0 – 0 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Entropy = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (1/6) = 0.65

C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

Entropy = – (2/6) $\log_2$ (2/6) – (4/6) $\log_2$ (4/6) = 0.92

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

آنتروپی

- از این معیار در الگوریتم‌های C4.5 و ID3
- وقتی که گره p به k دسته تقسیم‌بندی می‌شود، آنتروپی بعد از تقسیم‌بندی به صورت زیر محاسبه می‌شود:

$$Entropy(t) = \sum_{i=1}^{k} \frac{n_i}{n} Entopy(i)$$

- n_i تعداد نمونه‌های موجود در گره i
- n تعداد نمونه‌های قبل از تقسیم‌بندی در گره t

$$\Delta_{info} = Entopy(t) - \sum_{i=1}^{k} \frac{n_i}{n} Entopy(i)$$

بهره اطلاعات

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

آنتروپی

- ایرادات
  - آنتروپی و شاخص جینی تمایل دارد ناخالصی را کاهش دهد حتی اگر تعداد split ها زیاد باشد. (توجه به اصل غنی بودن اطلاعات)
  - راه حل
    - محدود کردن splitها به دسته‌بندی دوگانه - الگوریتم CART
    - تعیین نسبت بهره – الگوریتم C4.5

$$GainRatio_{split} = \frac{Gain_{split}}{Gain_{info}}, \qquad Gain_{info} = -\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}$$

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

خطای دسته‌بندی

- تعریف ریاضی

$P(j|t)$ بسامد نسبی کلاس j در گره t

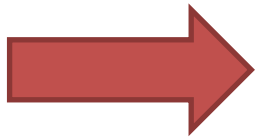
$$Error(t) = 1 - \max P(j|t)$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Error = 1 – max (0, 1) = 1 – 1 = 0

C1	**1**
C2	**5**



P(C1) = 1/6 P(C2) = 5/6

Error = 1 – max (1/6, 5/6) = 1 – 5/6 = 1/6

C1	**2**
C2	**4**

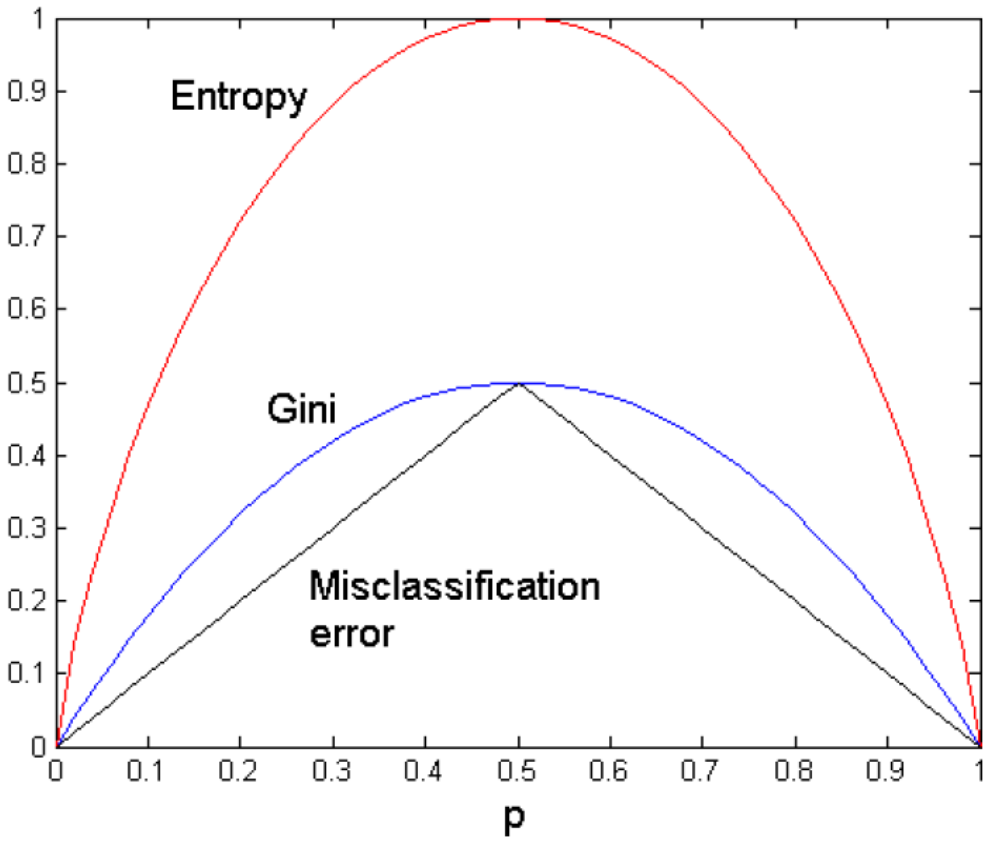
P(C1) = 2/6 P(C2) = 4/6

Error = 1 – max (2/6, 4/6) = 1 – 4/6 = 1/3

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

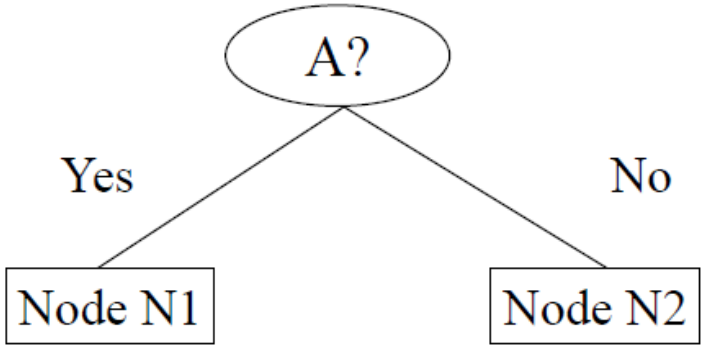
مقایسه شاخص‌های ناخالصی

برای یک مسئله دو کلاسه



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

مقایسه شاخص‌های ناخالصی



	Parent
C1	7
C2	3
Gini = 0.42	

Gini(N1)
= 1 – (3/3)² – (0/3)²
= 0

Gini(N2)
= 1 – (4/7)² – (3/7)²
= 0.489

	N1	N2
C1	3	4
C2	0	3

Gini(Children)
= 3/10 * 0
+ 7/10 * 0.489
= 0.342

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

شرایط توقف توسعه یک درخت

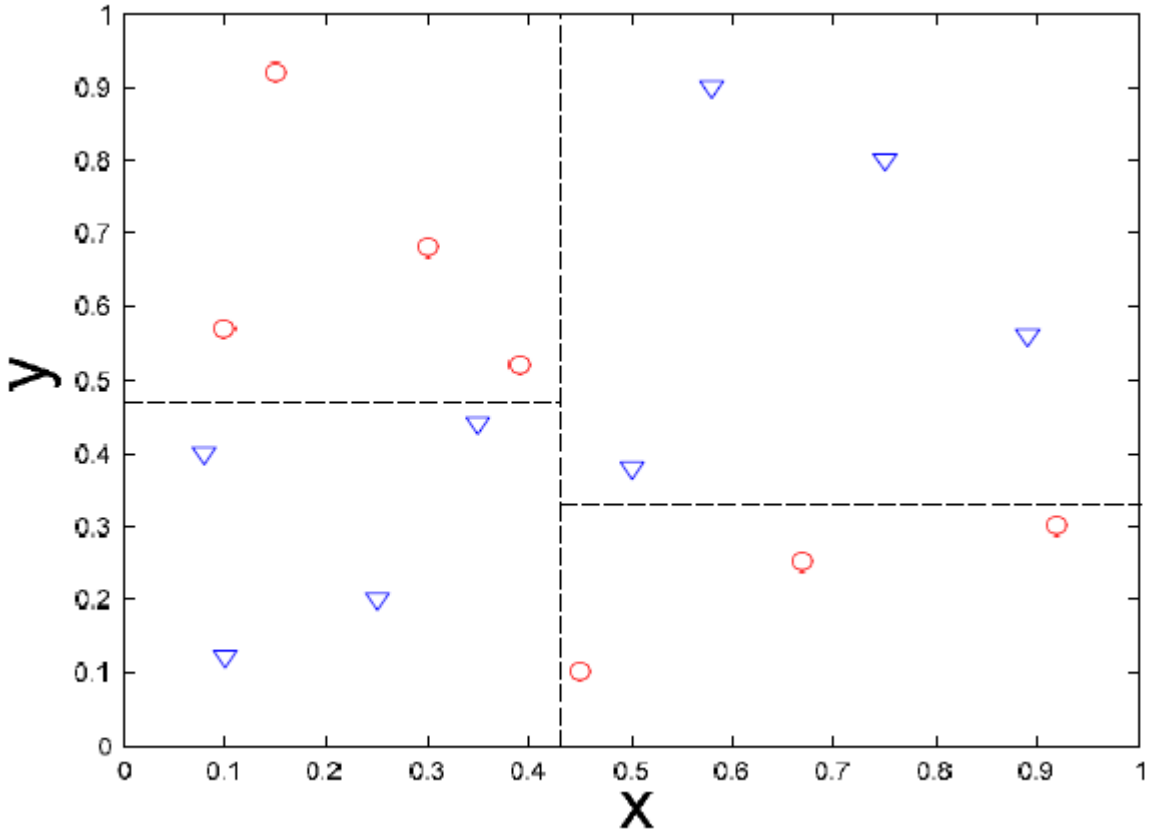
- شرایط

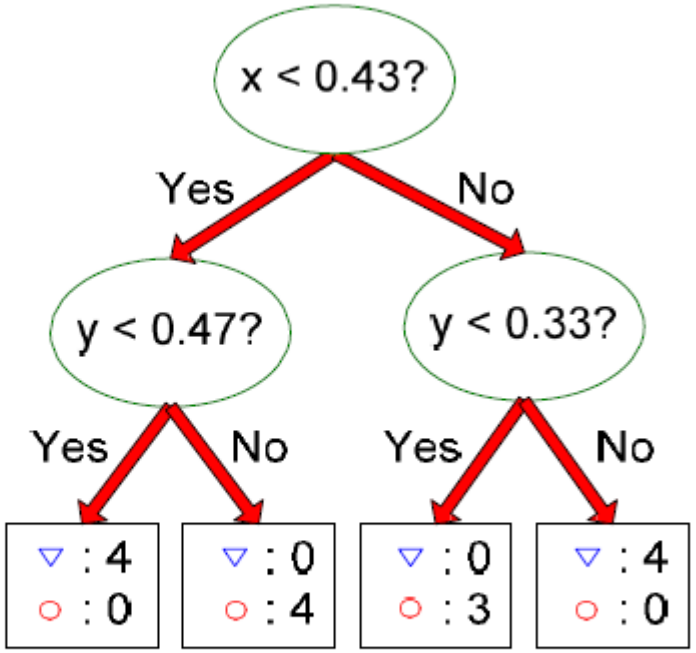
- توسعه درخت و اعمال شرایط تقسیم کردن متغیرها تا جایی ادامه می‌یابد که به جمعیت همگنی از داده‌ها برسیم و یا مقادیر ویژگی‌های مختلف برای یک متغیر یکسان باشد
- توقف زودهنگام – جلوگیری از رشد بی‌رویه درخت

مزایا و معایب

- مزایا
  - سرعت تصمیم‌گیری
  - سادگی و تفسیرپذیری
- معایب
  - جزئی شدن داده‌ها – محدود کردن تعداد تقسیم‌بندی‌ها
  - نوع تقسیم‌بندی کردن فضای داده‌ها – موازی با محورهای مختصات
  - نامناسب برای تعداد متغیرهای بالا

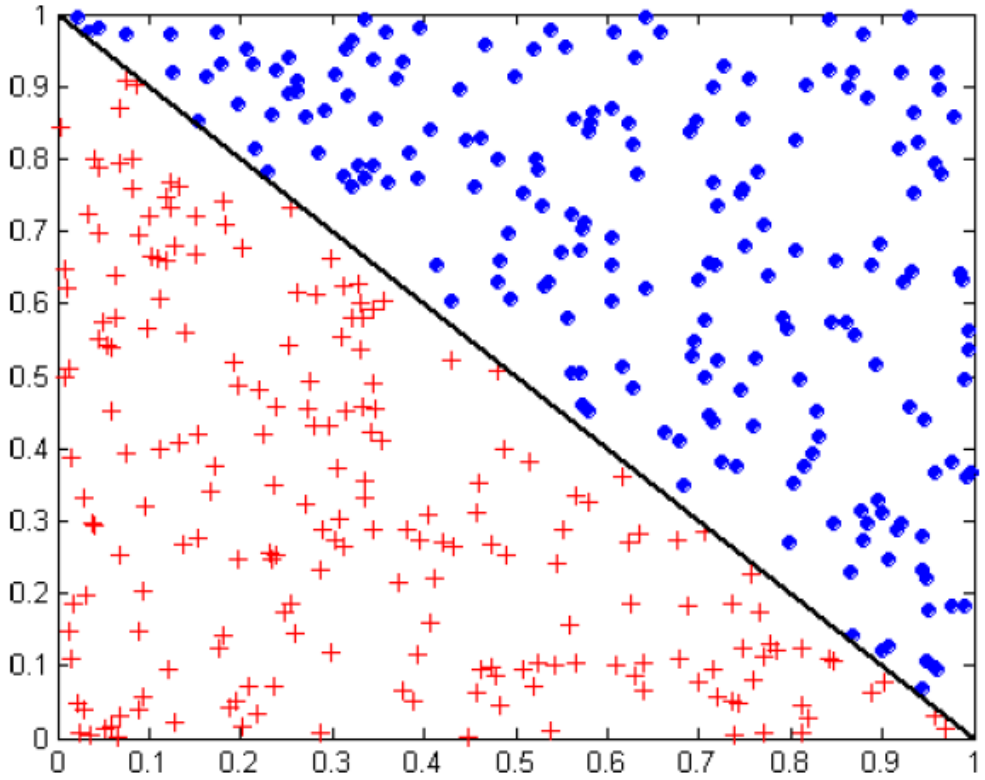
تقسیم‌بندی مورب

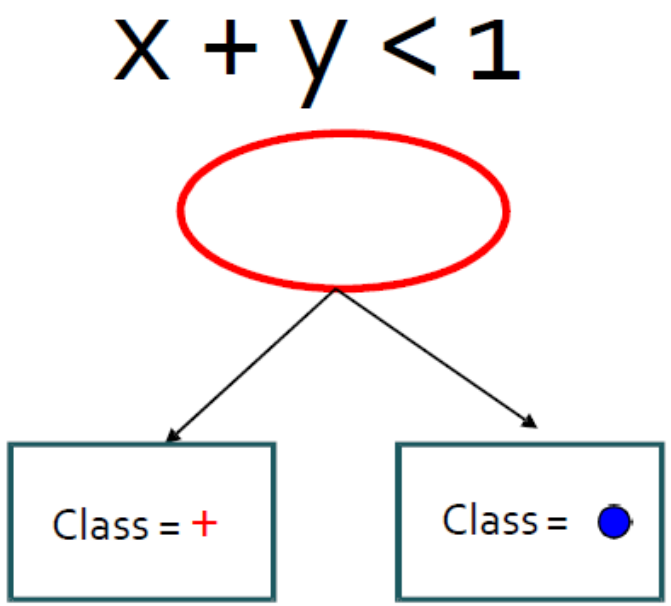




تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

تقسیم‌بندی مورب





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

مشکلات توسعه درخت

- چرا رشد بی رویه درخت نامطلوب است؟
  - پیچیدگی مدل
  - مثال - پیش‌بینی روند شاخص بورس در یک بازه زمانی 10 روزه

احتمال تخمین درست روند در حداقل 8 روز آینده

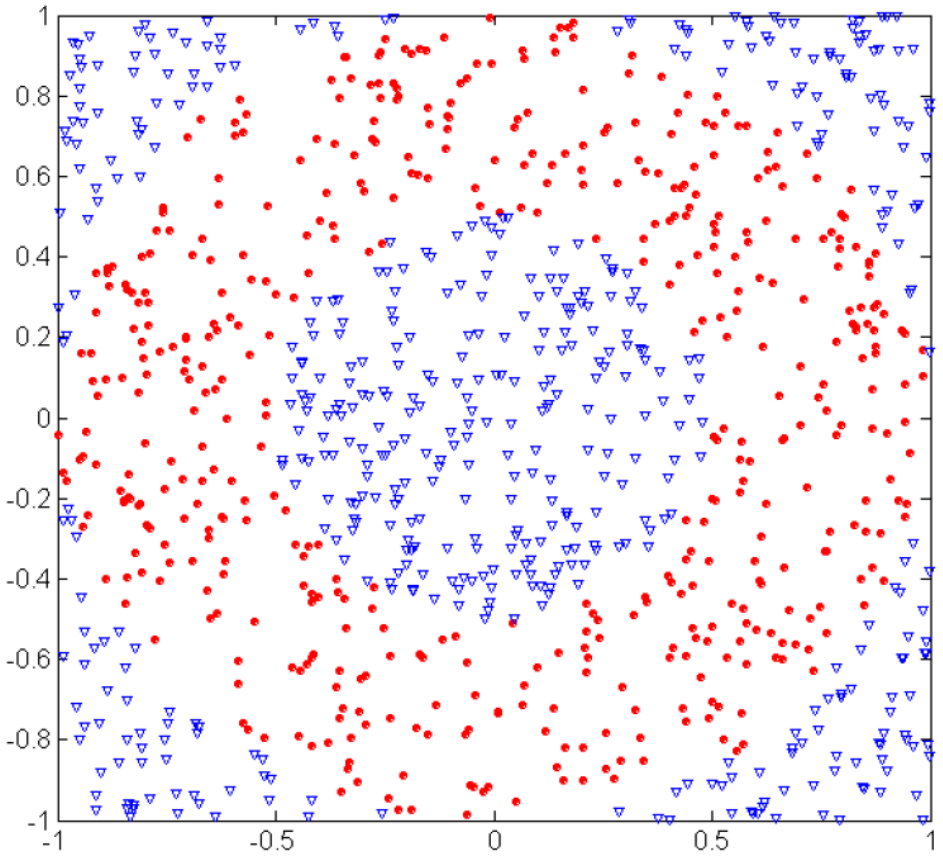
$$P = \frac{\binom{10}{8} + \binom{10}{9} + \binom{10}{10}}{2^{10}} = 0.05$$

  - اگر تعداد متخصصین را 50 در نظر بگیریم:

$$P = 1 - (1 - 0.05)^2 = 0.9$$

تولید محتوا: وحید محمدزاده ایوقی

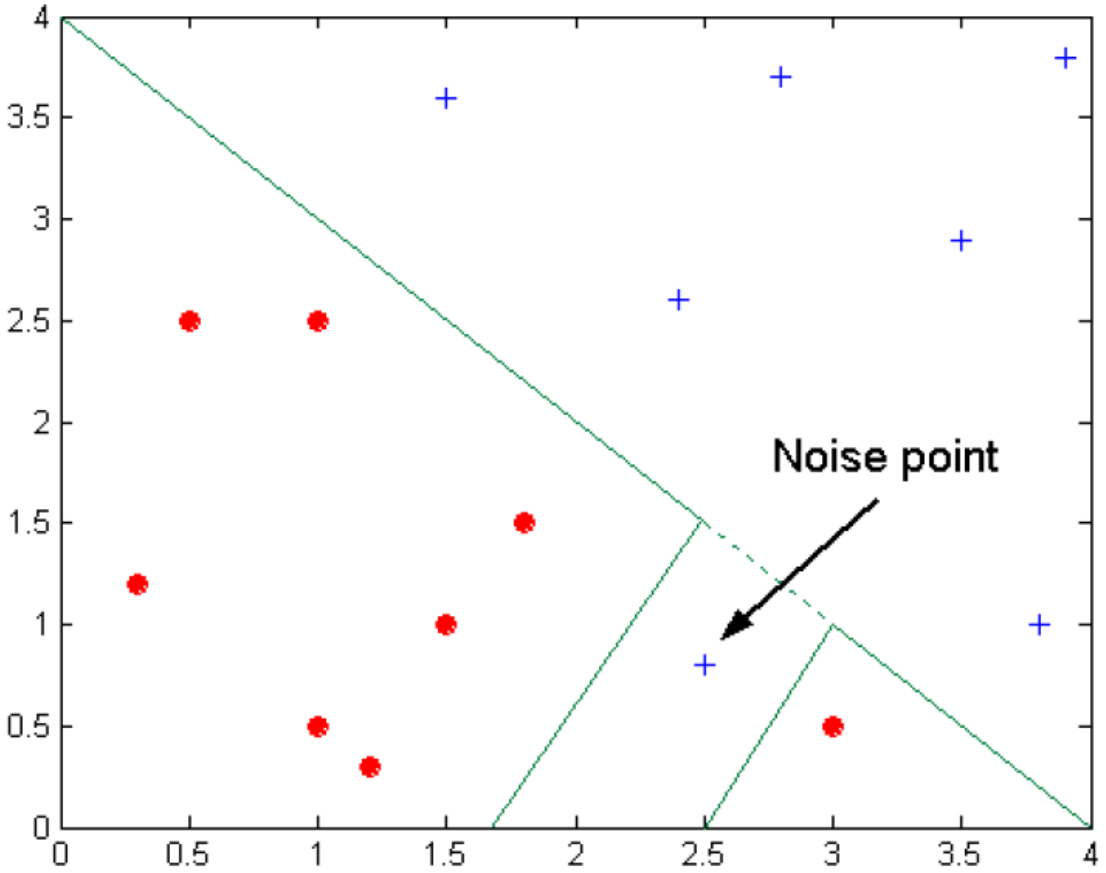
مشکلات توسعه درخت



- شاخص‌های ناخالصی تعریف شده برای توسعه درخت تمایل دارد تا حد امکان به سمت خالص شدن متغیرها برگ شود.
- برای این مسئله، ابعاد درخت توسعه داده شده بسیار بزرگ خواهد بود – پیچیدگی مدل بالا خواهد بود.
- از طرفی ساده بودن مدل نیز باعث می‌شود مدل به خوبی توسعه نیابد.

بیش برازش شدن درخت

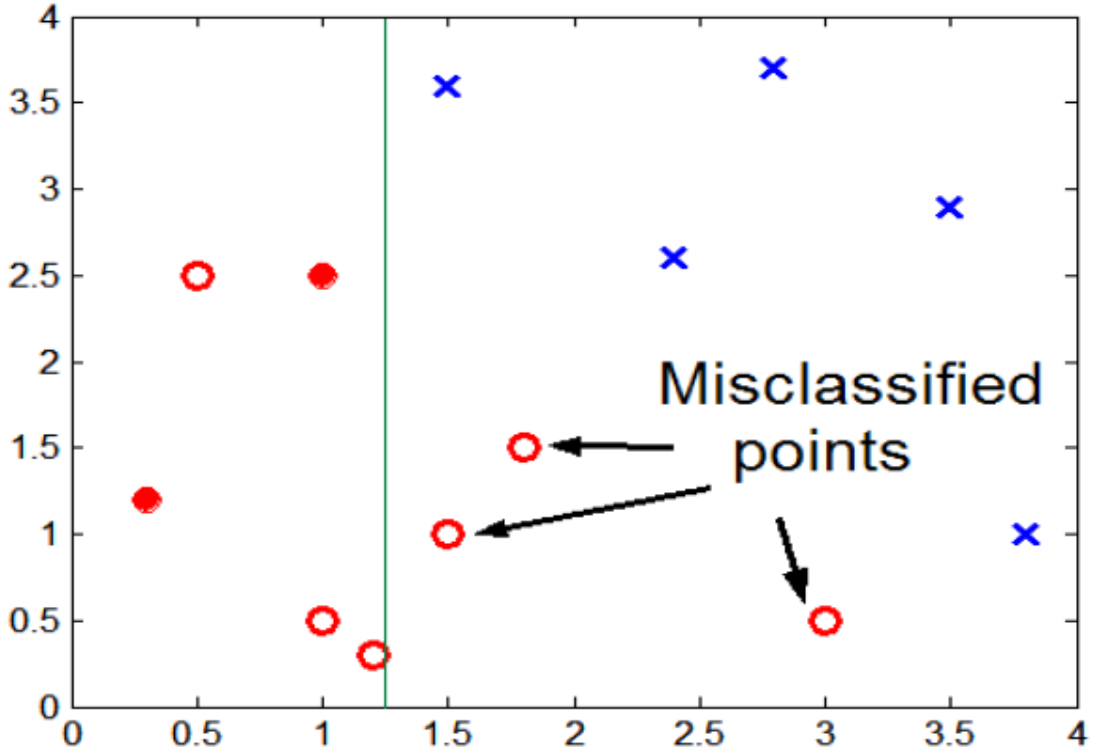
- نویز



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

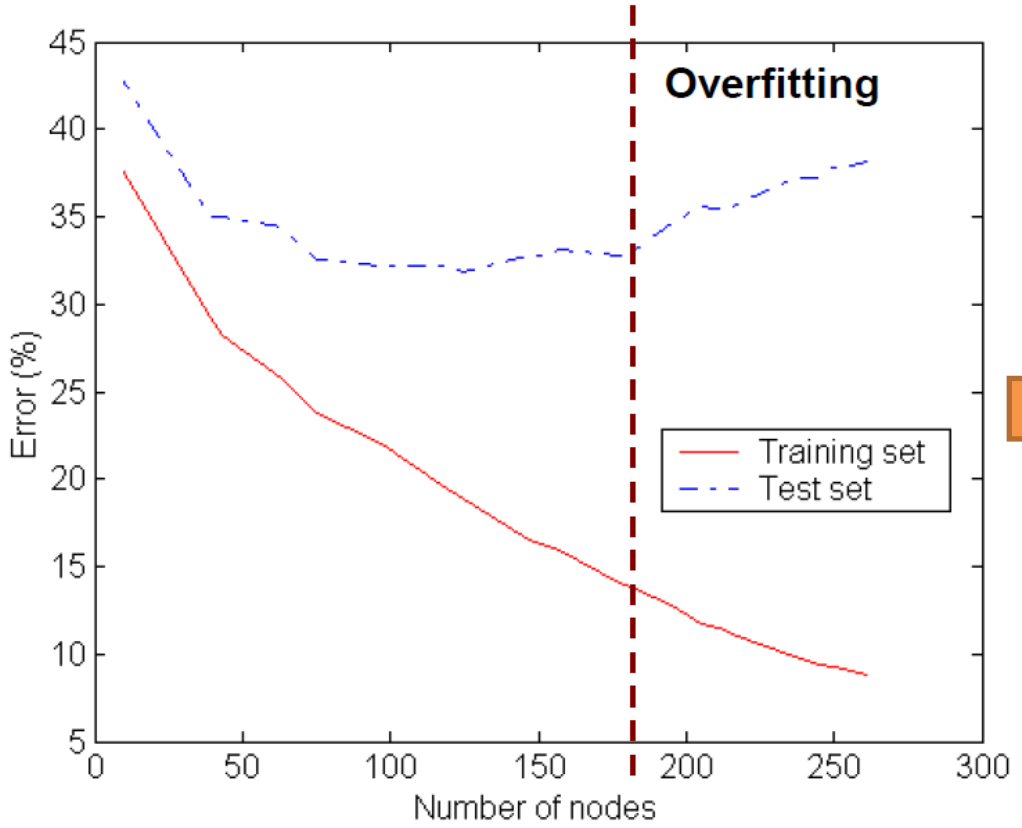
بیش برازش شدن درخت

- تعداد داده ناکافی



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

بیش برازش شدن درخت



- بیش برازش شدن درخت
- پیچیده شدن مدل بیش از حد نیاز

نیازمند تخمین زدن خطای تست هستیم.

Optimistic: $e(T') = e(T)$

Pessimistic: $e(T') = e(T) + 0.5 \times N$

Reduced Error Pruning (REP): نیازمند دیتاست اعتبارسنجی

جلوگیری از رشد درخت

- جلوگیری از رشد درخت – Pre pruning
  - همگن شدن متغیرهای برگ
  - یکسان شدن مقادیر ویژگی‌های مختلف برای یک متغیر
  - توزیع داده‌های کلاس بعد از تقسیم‌بندی کردن متفاوت از توزیع داده‌های اصلی باشد (آزمون آماری)
  - تعداد نمونه‌های موجود در هر برگ از یک حد آستانه کم‌تر باشد
  - عدم بهبود در شاخص ناخالصی

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

ساده‌سازی درخت

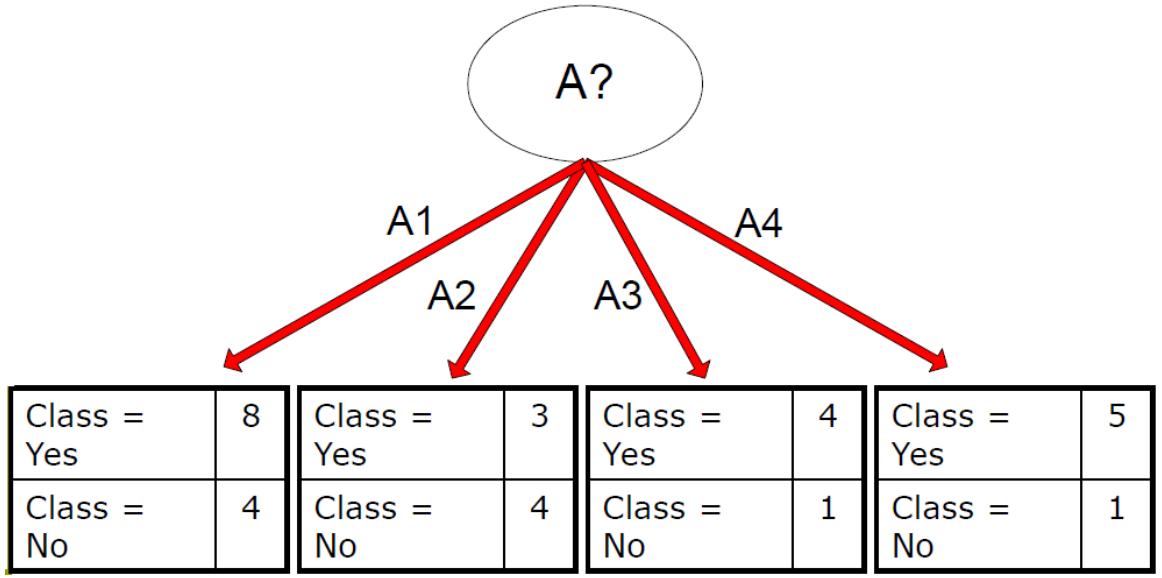
- ساده‌سازی درخت – Post pruning

- قوانین از پایین به بالا شروع به حذف شدن می‌کنند.
- اگر بهبود در خطای دسته‌بندی رخ داد، قانون متناظر را با یک متغیر برگ تغییر می‌دهیم.
  - بهبود در تخمین خطای تست باید در نظر گرفته شود.

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

ساده‌سازی درخت

Class = Yes	20
Class = No	10
Error = 10/30	



Training Error (Before splitting) = 10/30

Pessimistic error = (10 + 0.5)/30 = 10.5/30

Training Error (After splitting) = 9/30

Pessimistic error (After splitting)

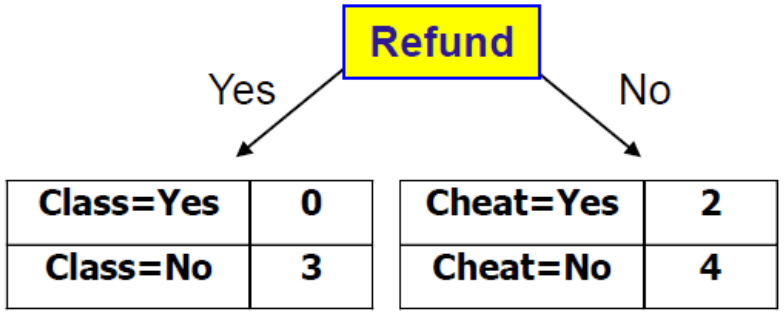
= (9 + 4 × 0.5)/30 = 11/30

PRUNE!

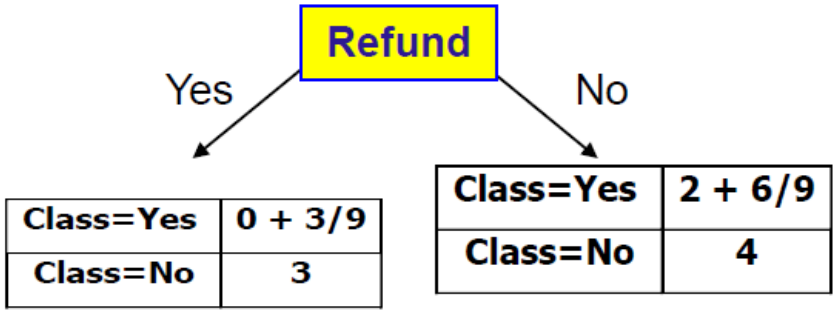
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
گروه دایکه | dayche.com

تاثیر مقادیر گمشده بر هرس کردن درخت

Tid	Refund	Marital Status	Taxable Income	Class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No



Tid	Refund	Marital Status	Taxable Income	Class
10	?	Single	90K	Yes



Probability that Refund=Yes is 3/9

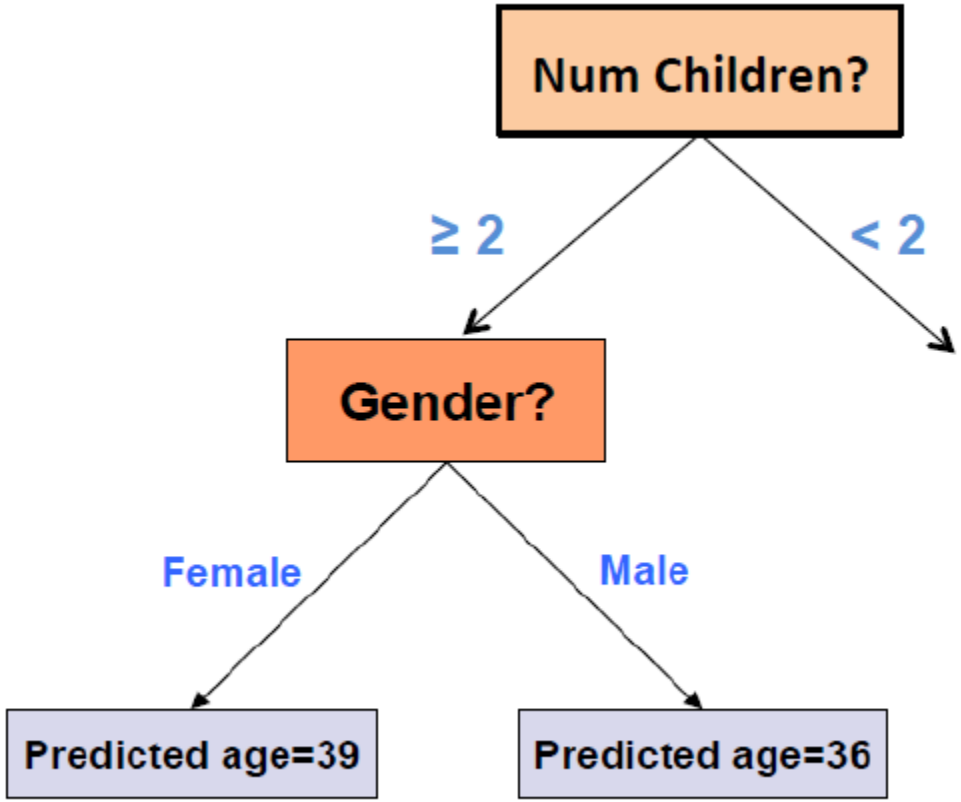
Probability that Refund=No is 6/9

Assign record to the left child with weight = 3/9 and to the right child with weight = 6/9

کاربرد درخت تصمیم در رگرسیون

$X^{(1)}$ .... $X^{(p)}$ Y

Gender	Rich?	Num. Children	# travel per yr.	Age
F	No	2	5	38
M	No	0	2	25
M	Yes	1	0	72
:	:	:	:	:



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