

درخت تصمیم

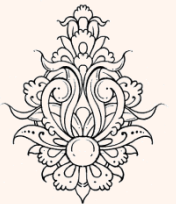
# یادگیری ماشین



دانشگاه شهید بهشتی  
پژوهشکده فضای مجازی  
پاییز ۱۳۹۸  
احمد محمودی ازناوه

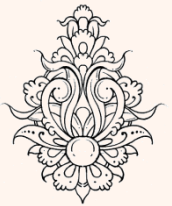
# فهرست مطالب

- درخت تصدیق
- درخت تصدیق تک‌متخیره
  - درخت‌های دسته‌بندی
  - درخت‌های رگرسیون
- هرس کردن
- استخراج قانون با کمک درخت
- استنتاج قانون
- درخت‌های چند متخیره



# یادگیری درخت تصمیم

- در روش‌های پارامتری برای همه‌ی داده‌ها یک مدل به دست می‌آید، در حالی که در روش‌های ناپارامتری داده‌ها بر اساس یک معیار فاصله به «نواحی محلی» تقسیم می‌شوند، سپس برای هر ناحیه یک مدل استخراج می‌شود.
  - در این حالت یافتن ناحیه‌ی مربوط هزینه‌بر است؛ به ازای هر ورودی همه‌ی داده‌های آموزشی جستجو شده تا ناحیه‌ی مورد نظر پیدا شود.
- «درخت تصمیم» یک ساختمان داده‌ی سلسله‌مراتبی است که برای یادگیری «بانظارت» مورد استفاده قرار می‌گیرد.
  - با تعداد مراحل کم‌تری ناحیه‌ی محلی شناسایی می‌شود.
  - بر پایه‌ی سیاست «تفرقه‌بینداز و حکومت‌کن» عمل می‌کند.
  - این شیوه نیز «ناپارامتری» است، از این حیث که یک مدل ثابت برای داده‌ها در نظر نمی‌گیرد، درخت با توجه به داده‌ها منشعب می‌شود و رشد می‌کند.
  - بیشتر در «داده‌کاوی» مورد استفاده قرار می‌گیرد.



# ساختار درخت تصمیم

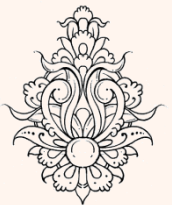
- درخت تصمیم از یک سری گرهی «تصمیم (داخلی)» و «برگ» تشکیل شده است.

**internal decision node**

- گره‌های تصمیم، معادل یک «تابع آزمون»  $(f_m(x))$  هستند که بر اساس نتیجه‌ی آن‌ها خروجی‌ها برچسب می‌خورند. به ازای هر ورودی جستجو از ریشه شروع شده و در هر گرهی تصمیم یکی از انشعاب‌ها انتخاب می‌شود. این فرآیند تا رسیدن به برگ ادامه می‌یابد.

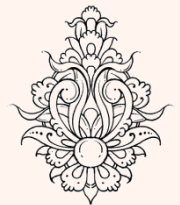
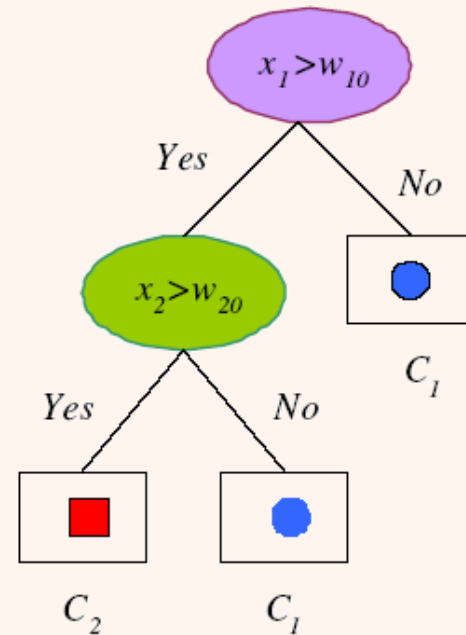
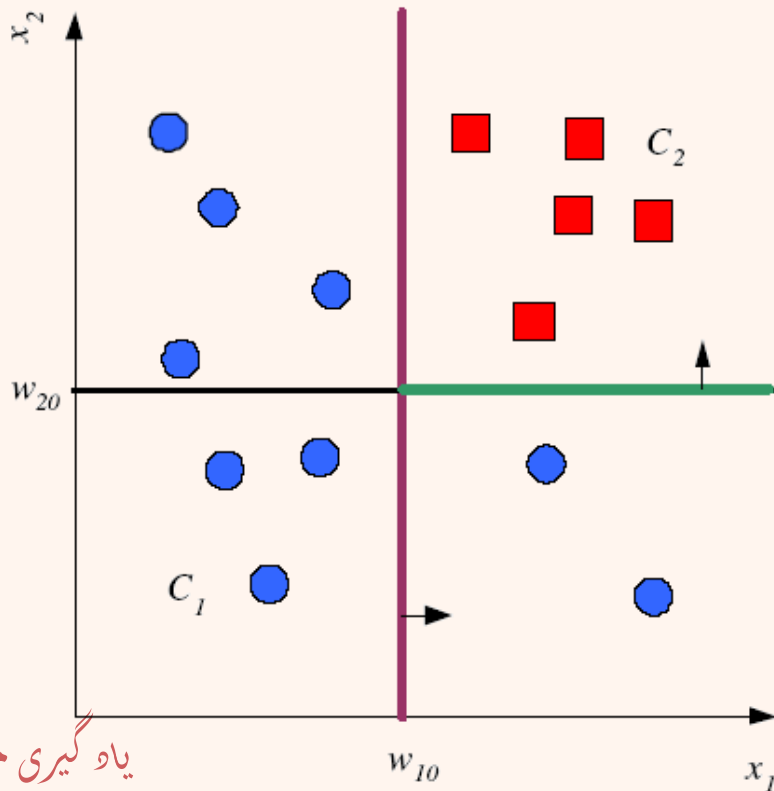
**Leaves**

- مقدار گرهی برگ خروجی به ازای آن ورودی را نشان می‌دهد. در دسته‌بندی شماره‌ی کلاس و در رگرسیون مقدار میانگین



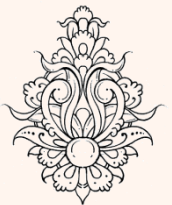
# ساختار درخت تصمیم (ادامه...)

- هر «تابع آزمون» ( $f_m(x)$ ) یک جداساز در فضای  $d$ -بعدی است، که فضای ورودی را به نواحی کوچکتر تقسیم می‌کند. هر  $f$  یک تابع ساده است که با ترکیب با یکدیگر توابعی پیچیده را خواهند ساخت.



# ساختار درخت تصمیم (ادامه...)

- بدین ترتیب درخت تصمیم، به یافتن نمونه‌های مشابه سرعت می‌بخشد. به عنوان مثال در صورت داشتن  $b$  نامیه‌ی مختلف و در صورتی که تصمیم‌ها دودویی باشند، در بهترین حالت  $\log_2 b$  گام تا رسیدن به نامیه‌ی محلی نیاز است.
- از دیگر مزایای درخت تصمیم «قابلیت تفسیرپذیری» آن است.
- می‌توان از درخت تشکیل شده یک سری قانون به صورت **if-then** استخراج کرد.



# Univariate Trees

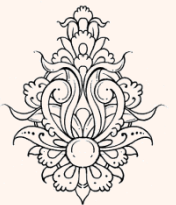
- در «درخت‌های تک‌متغیره»، در هر گرهی داخلی تنها یکی از ابعاد ورودی مورد بررسی قرار می‌گیرد.
  - در صورتی که متغیرها گسسته باشند، مقدار آن‌ها (با یکی از  $n$  مقدار ممکن) بررسی می‌شود.
  - و گرنه مقایسه‌ای صورت می‌گیرد که نتیجه‌ی گسسته داشته باشد:

$$- f_m(\mathbf{x}) : x_j \geq w_{m0}$$

- بدین ترتیب فضای ورودی را به دو قسمت تقسیم می‌شود:

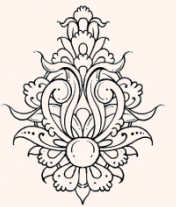
$$L_m = \{ \mathbf{x} \mid x_j \geq w_{m0} \} \quad R_m = \{ \mathbf{x} \mid x_j < w_{m0} \}$$

Binary Split



# آموزش درخت تصمیم

- به فرآیند ساخت درخت بر اساس داده‌های آموزشی «tree induction» می‌گویند.
- برای یک مجموعه‌ی آموزشی، درخت‌های زیادی می‌توان در نظر گرفت که دارای خطای صفر باشند.
- مطلوب‌ترین درخت، کوچک‌ترین درخت از نظر تعداد گره‌ها و سادگی عملیات مقایسه‌ی هر گره است.
- بیشتر الگوریتم‌های یادگیری درخت تصمیم مریضانه است.
  - در هر گام در پی بهترین جداسازی است.



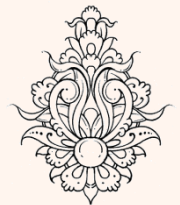


- «**ناخالصی**» معیاری برای تشخیص میزان مطلوبیت درخت دسته‌بندی است.
- یک انشعاب (split) «**خالص**» است، در صورتی که هر بخش آن شامل نمونه‌های یک کلاس باشد.
- $N_m$  تعداد نمونه‌هایی است که به گرهی  $m$ -ا می‌رسیده‌اند (کل گره‌ها  $N$  است).
- از بین آن‌ها  $N_m^i$  متعلق به کلاس  $i$ -ا ( $C_i$ ) است.

$$\sum_i N_m^i = N_m$$

- احتمال این که نمونه‌ای که به گرهی  $m$  رسیده است، به دسته‌ی  $i$ -ا تعلق داشته باشد:

$$\hat{P}(C_i | \mathbf{x}, m) \equiv p_m^i = \frac{N_m^i}{N_m}$$

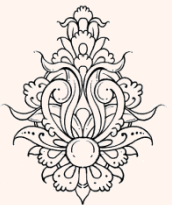


# درخت‌های دسته‌بندی (ادامه...)

- گرهی  $m$  خالص است، اگر

$$p_m^i = (0 \text{ or } 1) \quad \text{for all } i$$

- در صورتی که تمامی جداسازی‌ها خالص باشند، لزومی به انشعاب مجدد درخت نیست و می‌توان برگ‌ها با برچسب دسته را به درخت افزود.
- فرآیند آموزش درخت باید به گونه‌ای صورت پذیرد که در هر گام میزان خلوص افزایش یابد.
- از این جهت لازم است معیاری برای سنجش خلوص در نظر گرفته شود.

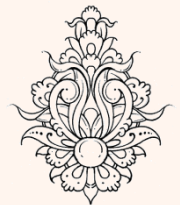
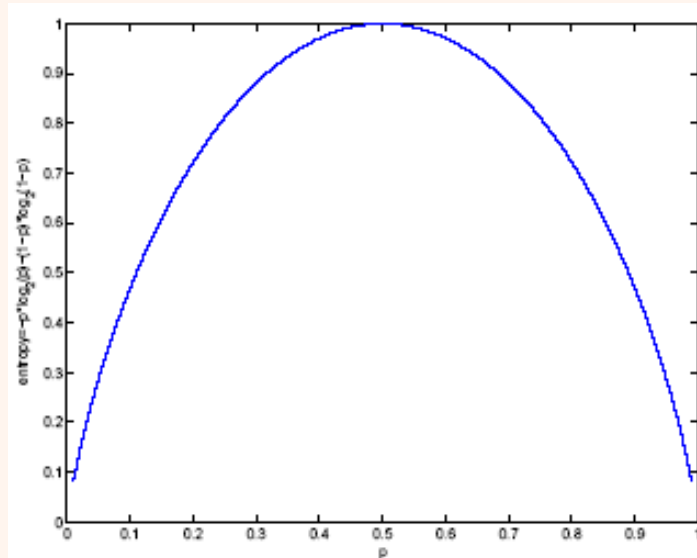


# خلوص

• یک انتخاب برای سنجش میزان خلوص، «آنتروپی اطلاعات» است.

– آنتروپی اطلاعات: رخداد اتفاقی تا چه حد تصادفی است.

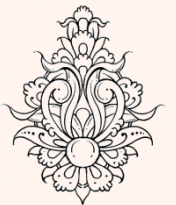
$$I_m = -\sum_{i=1}^K p_m^i \log_2 p_m^i$$



# سایر معیارها

• آنتروپی تنها معیار مناسب نیست؛ برای دسته‌بندی دو گروه، هر تابع نامنفی  $\phi(p, 1-p)$  که دارای خصوصیات زیر باشد، به عنوان معیار خلوص قابل استفاده است.

- $\phi(1/2, 1/2) \geq \phi(p, 1-p)$ , for any  $p \in [0, 1]$
- $\phi(0, 1) = \phi(1, 0) = 0$
- $\phi(p, 1-p)$  is increasing in  $p$  on  $[0, 1/2]$   
and decreasing in  $p$  on  $[1/2, 1]$



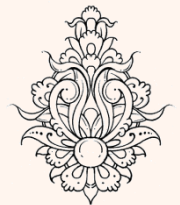
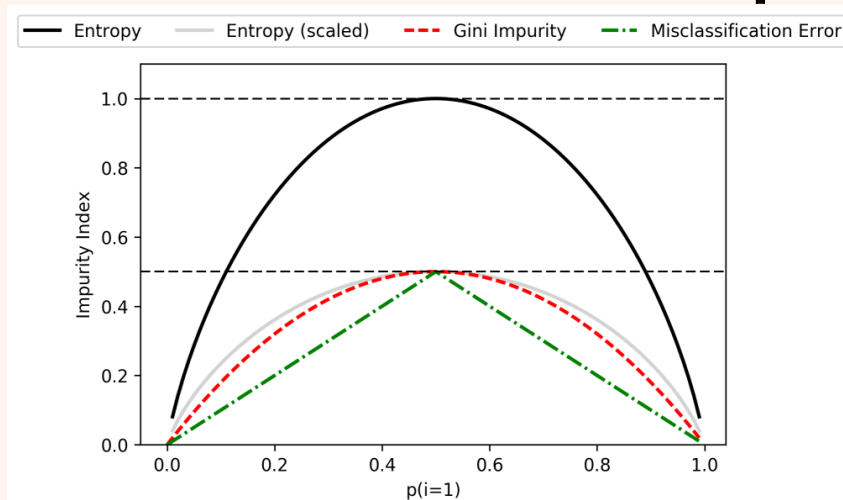
# سایر معیارها (ادامه...)

$$\text{entropy} : \phi(p, 1-p) = -p \log_2(p) - (1-p) \log_2(1-p)$$

$$\text{Gini index} : \phi(p, 1-p) = 2p(1-p)$$

$$\text{Missclassification error} : \phi(p, 1-p) = 1 - \max(p, 1-p)$$

- این معیارها قابل تعمیم به  $K > 2$  هستند.
- پژوهش‌ها نشان داده است که این سه معیار در عمل تفاوت چندانی با هم ندارند.



# تشکیل گرهی تصمیم

- انشعابی انتخاب می‌شود که باعث کاهش ناخالصی شود.

- در صورتی که از  $N_m$  ورودی که به گرهی  $m$ -اوم رسیده‌اند،  $N_{mj}$  ورودی انشعاب  $j$ -اوم این گره را انتخاب کنند، خواهیم داشت:

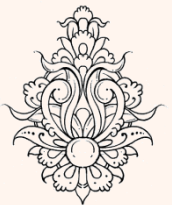
$$\sum_{j=1}^n N_{mj} = N_m \quad \sum_{i=1}^K N_{mj}^i = N_{mj} \quad \sum_{j=1}^n N_{mj}^i = N_m^i$$

- بعد از انشعاب و در صورت انتخاب انشعاب  $j$ -اوم احتمال تعلق به کلاس  $i$ -اوم:

$$\hat{P}(C_i | \mathbf{x}, m, j) \equiv p_{mj}^i = \frac{N_{mj}^i}{N_{mj}}$$

- و «ناخالصی کل» به صورت زیر به دست می‌آید:

$$I'_m = - \sum_{j=1}^n \frac{N_{mj}}{N_m} \sum_{i=1}^K p_{mj}^i \log_2 p_{mj}^i$$



# آموزش درخت‌های دسته‌بندی

- برای همه‌ی خصیصه‌ها، ناخالصی به ازای تمام حالات محاسبه می‌شود:

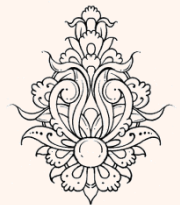
- در حالتی که خصیصه‌ها گسسته باشند (در صورتی که خصیصه‌ی مورد نظر دارای  $n$  مقدار مختلف باشد)، هر گره‌ی تصمیم به  $n$  گره‌ی دیگر، منشعب می‌شود.

- در مورد خصیصه‌های پیوسته، هر گره به دو گره‌ی دیگر منشعب می‌شود.

$$f_m(x) : x_j \geq w_{m0}$$

- در این حالت باید مقدار آستانه‌ی  $w_{m0}$  نیز مشخص شود،  $N_m - 1$  مقدار ممکن برای این حد آستانه وجود دارد، به ازای همه‌ی خصیصه‌ها و همه‌ی حد آستانه‌های ممکن ناخالصی محاسبه می‌شود.

- به جای ذخیره‌ی برچسب گروه در برگ‌ها، بهتر است احتمال تعلق به کلاس (posterior probability) ذخیره شود، چرا که ممکن است در مراحل بعدی، مانند محاسبه‌ی ریسک به کار آید.



## GenerateTree( $\mathcal{X}$ )

If NodeEntropy( $\mathcal{X}$ ) <  $\theta_I$  /\* eq. 9.3

$$I_m = -\sum_{i=1}^K p_m^i \log_2 p_m^i$$

Create leaf labelled by majority class in  $\mathcal{X}$

Return

$i \leftarrow \text{SplitAttribute}(\mathcal{X})$

For each branch of  $\mathbf{x}_i$

Find  $\mathcal{X}_i$  falling in branch

GenerateTree( $\mathcal{X}_i$ )

SplitAttribute( $\mathcal{X}$ )

MinEnt  $\leftarrow$  MAX

For all attributes  $i = 1, \dots, d$

If  $\mathbf{x}_i$  is discrete with  $n$  values

Split  $\mathcal{X}$  into  $\mathcal{X}_1, \dots, \mathcal{X}_n$  by  $\mathbf{x}_i$

$e \leftarrow \text{SplitEntropy}(\mathcal{X}_1, \dots, \mathcal{X}_n)$  /\* eq. 9.8 \*/

If  $e < \text{MinEnt}$  MinEnt  $\leftarrow e$ ; bestf  $\leftarrow i$

Else /\*  $\mathbf{x}_i$  is numeric \*/

For all possible splits

Split  $\mathcal{X}$  into  $\mathcal{X}_1, \mathcal{X}_2$  on  $\mathbf{x}_i$

$e \leftarrow \text{SplitEntropy}(\mathcal{X}_1, \mathcal{X}_2)$

If  $e < \text{MinEnt}$  MinEnt  $\leftarrow e$ ; bestf  $\leftarrow i$

Return bestf

در صورتی که تشکیل درخت تا حالتی که نافالسی به صفر برسد ادامه یابد ممکن است منجر به overfitting شود. به ویژه با وجود نویز، در این حالت یک مدآستانه برای نافالسی در نظر گرفته می شود. این معیار ( $\theta_I$ ) پیچیدگی مدل را مشخص می کند.

$$I'_m = -\sum_{j=1}^n \frac{N_{mj}}{N_m} \sum_{i=1}^K p_{mj}^i \log_2 p_{mj}^i$$





- درخت‌های رگرسیون، مانند درخت‌های دسته‌بندی هستند؛ با یک معیار ناخالصی مناسب:

«میانگین مربعات خطا»

$$b_m(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in \mathcal{X}_m : \mathbf{x} \text{ reaches node } m \\ 0 & \text{otherwise} \end{cases}$$

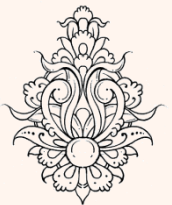
مقدار تخمین زده شده در گرهی  $m$ -ام

$$E_m = \frac{1}{N_m} \sum_t (r^t - g_m)^2 b_m(\mathbf{x}^t)$$

- می‌توان از میانگین (میان) برای داده‌های نویزی برای تخمین خروجی استفاده کرد.

$$g_m = \frac{\sum_t b_m(\mathbf{x}^t) r^t}{\sum_t b_m(\mathbf{x}^t)}$$

در صورت در نظر گرفتن میانگین داده‌های هر برگ، معیار خطا واریانس داده‌ها خواهد بود



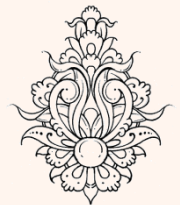
# درخت‌های رگرسیون (ادامه...)

- در صورتی که میزان خطا قابل پذیرش باشد، گرهی برگ ایجاد می‌شود، وگرنه روند تقسیم ادامه می‌یابد.
- علاوه بر میانگین مربعات خطا، می‌توان از معیار «بدترین خطای ممکن» نیز استفاده کرد، بدین ترتیب می‌توان حداکثر میزان خطای ممکن را محدود نمود:

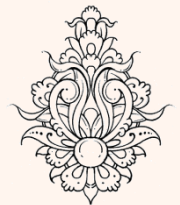
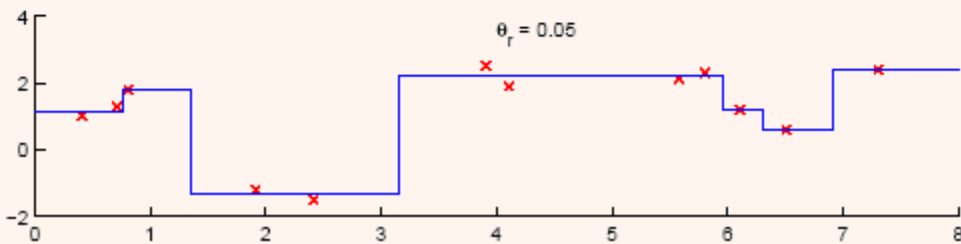
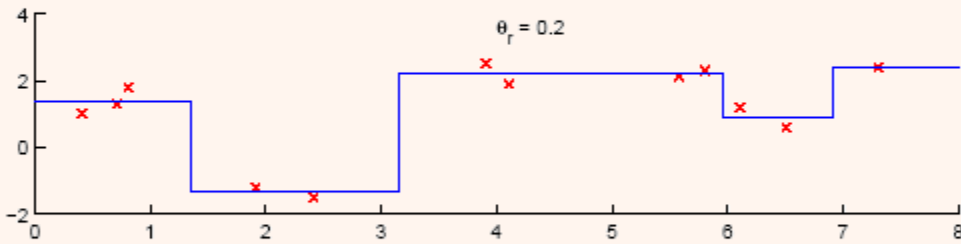
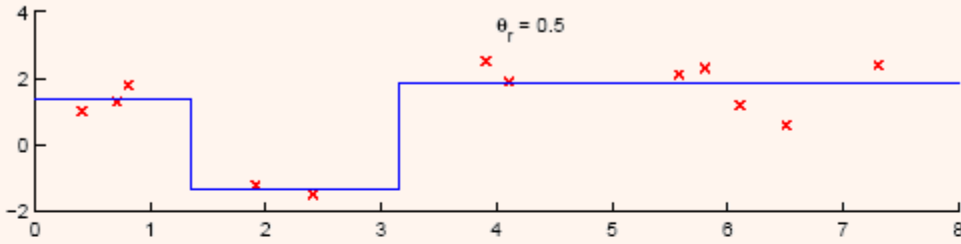
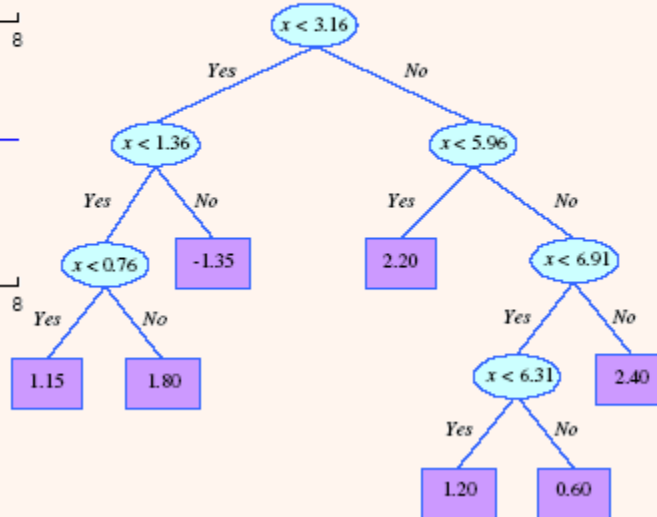
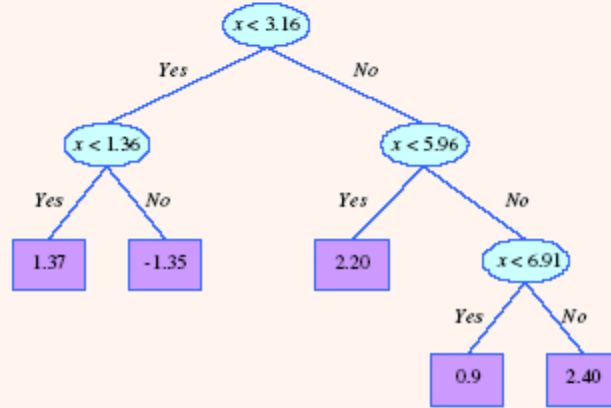
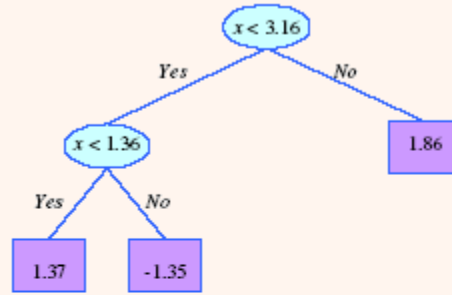
$$E_m = \max_j \max_t |r^t - g_{mj}| b_m(\mathbf{x}^t)$$

- میزان پیچیدگی مدل با حداکثر خطای قابل پذیرش ( $\theta_r$ ) مشخص می‌شود.
- به جای میانگین‌گیری، می‌توان در هر برگ برای داده‌ها یک مدل خطی نیز در نظر گرفت (رگرسیون خطی):

$$g_m(\mathbf{x}) = \mathbf{w}_m^T \mathbf{x} + w_{m0}$$



# مثال



# هرس کردن درخت

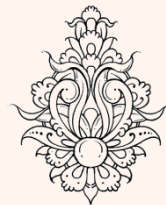
• «هرس کردن درخت»: حذف زیردرختها برای افزایش قابلیت تعمیم پذیری

– Prepruning: در صورتی که تعداد داده‌هایی که به یک گره می‌رسند از یک حد آستانه کمتر باشد (به عنوان مثال پنج درصد داده‌های آموزشی)، آن گره به گره‌های بیشتر منتسب نخواهد شد.

– Postpruning: بعد از رشد کامل درخت، زیر درخت‌هایی که موجب بیش‌برازش (Overfitting) می‌شوند، حذف خواهند شد.

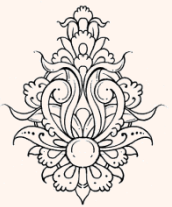
• نیاز به یک مجموعه هرس (pruning set) جدا از مجموعه آموزشی

• Prepruning سریع‌تر است، در حالی که postpruning دقیق‌تر است.



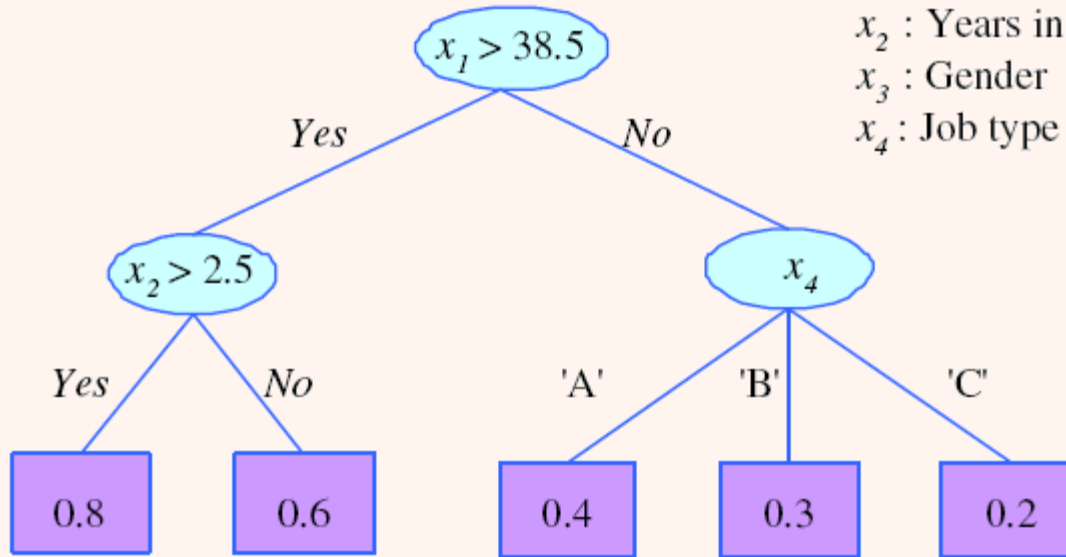
# استخراج دانش

- درخت تصمیم به نوعی کار **استخراج خصیصه** را انجام می‌دهد.
  - ممکن است پس از تشکیل درخت، برخی خصیصه‌ها مورد استفاده قرار نگیرند.
  - می‌توان گفت خصیصه‌هایی که به ریشه نزدیک‌تر هستند، مهمتر می‌باشند.
- درخت تصمیم، قابلیت تفسیرپذیری دارد.
  - مسیر ریشه تا هر برگ، «ترکیب عطفی (conjunction)» یک سری شروط است، چرا که همه‌ی شروط باید برقرار باشند تا داده به گره‌ی برگ مورد نظر برسد.
  - همه‌ی این مسیرها یک «**پایگاه قوانین (rule base)**» را تشکیل می‌دهند.
  - برای هر قانون می‌توان میزان پشتیبانی (نسبت داده‌های پوشش داده شده توسط قانون) را محاسبه نمود.
  - در دسته‌بندی ممکن است، چند برگ تشکیل یک کلاس را بدهند، در این صورت قوانین به صورت «ترکیب فصلی (disjunction)» بیان می‌شوند.
- می‌توان قوانین به دست آمده را نیز هرس کرد؛ پس از هرس کردن دیگر نمی‌توان درخت معادل آن را ساخت.



# استخراج دانش (ادامه...)

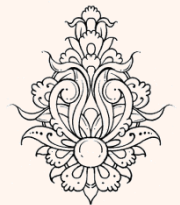
$x_1$  : Age  
 $x_2$  : Years in job  
 $x_3$  : Gender  
 $x_4$  : Job type



- R1: IF (age > 38.5) AND (years-in-job > 2.5) THEN  $y = 0.8$   
R2: IF (age > 38.5) AND (years-in-job  $\leq$  2.5) THEN  $y = 0.6$   
R3: IF (age  $\leq$  38.5) AND (job-type='A') THEN  $y = 0.4$   
R4: IF (age  $\leq$  38.5) AND (job-type='B') THEN  $y = 0.3$   
R5: IF (age  $\leq$  38.5) AND (job-type='C') THEN  $y = 0.2$

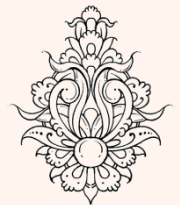
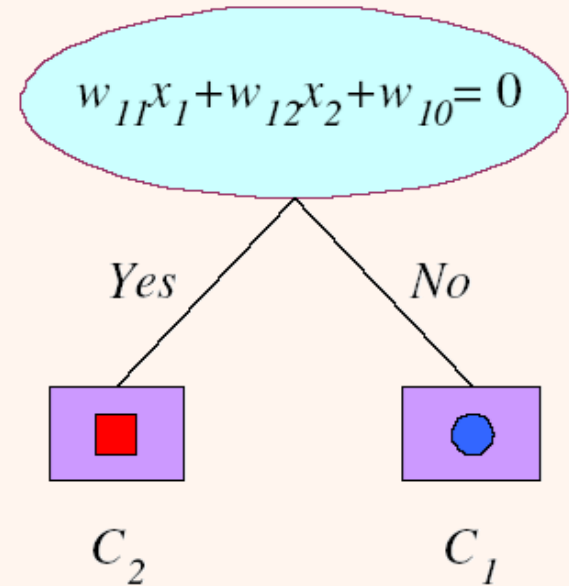
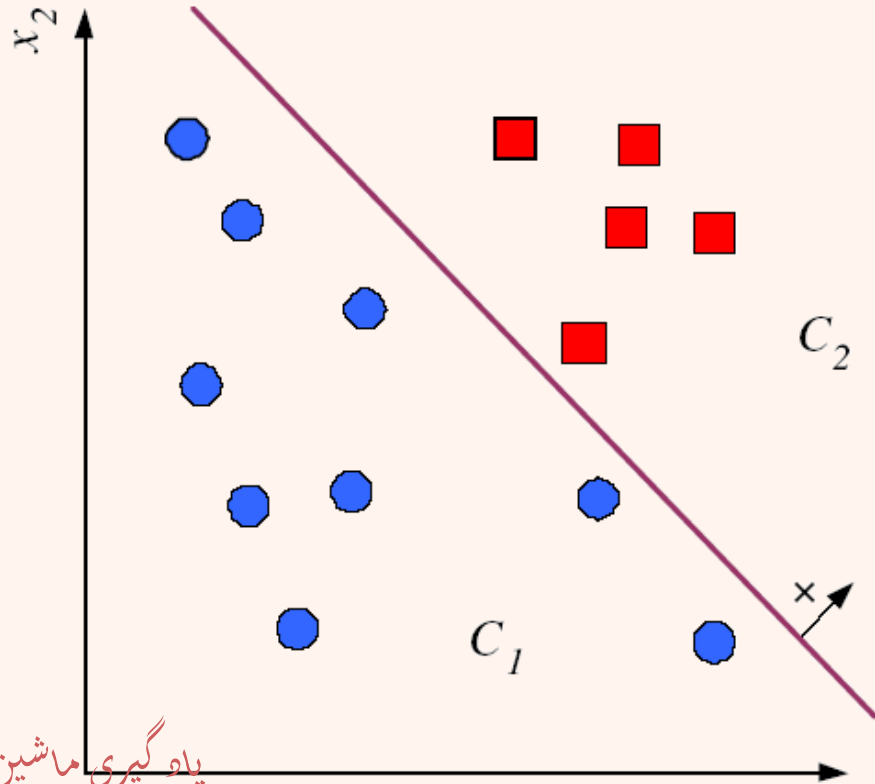
به عنوان مثال در صورتی که تمام افرادی که شغل A را دارند، فروچی 0.4 داشته باشد می‌توان R3 را هرس کرد:

R3': IF (job-type='A') THEN  $y = 0.4$



- در درخت چندمتغیره تمام خصیصه‌های در گره‌های داخلی قابل استفاده هستند، به عنوان مثال برای خصیصه‌های پیوسته:

$$f_m(\mathbf{x}) = \mathbf{w}_m^T \mathbf{x} + w_{m0} > 0$$



# درخت چندمتغیره (ادامه...)

- بدین ترتیب، نوامی محلی با یک چندضلعی محدود می‌شوند. (در حالت تک‌متغیره، جداساز  $i$ ، بر محور  $i$  عمود است)

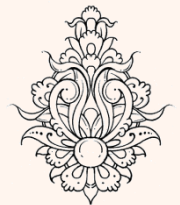
- - در حالت تک‌متغیره، تنها  $d$  راستا برای انتخاب وجود دارد و  $N_m - 1$  مقدار برای جداساز

- در حالت چندمتغیره، امکان استفاده از گره‌های غیرقطبی وجود دارد.  
$$f_m(\mathbf{x}) = \mathbf{x}\mathbf{W}_m^T \mathbf{x} + \mathbf{w}_m^T \mathbf{x} + w_{m0} > 0$$

- همچنین استفاده از MLP در هر گره، پیشنهاد شده است.

- همچنین استفاده از گره‌های کروی (sphere node) پیشنهاد شده است.

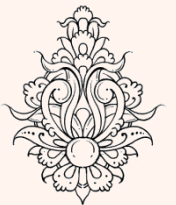
$$f_m(\mathbf{x}) = \|\mathbf{x} - \mathbf{c}_m\| > \alpha_m$$





# درخت چندمتخیره (ادامه...)

- برای CART یک نمونه‌ی چندمتخیره پیشنهاد شده است.
- در برخی روش‌ها، مرحله‌ی پیش‌پردازش کاهش بعد نیز انجام می‌شود.
- در روش تک‌متخیره، فضای فرضیه ساده‌تر است، در نتیجه در فضای چندمتخیره با تعداد گره‌های کم‌تر تخمین بهتری انجام می‌شود.
- - تعداد انشعاب‌ها (branching factor) هم اثر مشابهی دارد.
- بین ابعاد درخت، تعداد انشعاب‌ها و پیچیدگی گره‌ها trade-off وجود دارد.

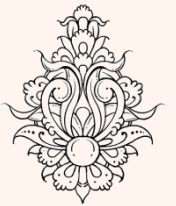


# درخت چندمتغیره (ادامه...)

یک درخت **بزرگ** با گره‌های **ساده** و تعداد انشعاب **کم**  
یا

یک درخت **کوچک** با گره‌های **پیچیده** و تعداد انشعاب **زیاد**

- قابلیت تفسیرپذیری و تصمیم‌پذیری درخت اول بهتر است.
- یکی از مزایای درخت، شکستن مسأله به مسائل کوچک‌تر است، افزایش پیچیدگی در واقع نقض غرض است.



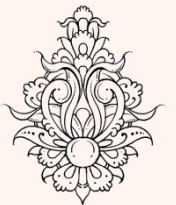
- دارای ساختار ترکیبی است.

- در هر گره، با توجه به نوع و پیچیدگی مسأله، جداساز مطلوب ممکن است متفاوت باشد.

- گره‌ها ممکن است تک‌متخیره، چندمتخیره (خطی و یا غیرخطی) باشد.

- مدل‌های پیچیده تنها زمانی انتخاب می‌شوند که مدل‌های ساده کارایی قابل مقایسه نداشته باشند.

- نتایج نشان داده است که گره‌های نزدیک به ریشه پیچیده‌تر هستند (هرچه از ریشه دور شویم، مسأله ساده‌تر می‌شود).



## مزایای درخت تصمیم

- مشابه سایر روش‌های ناپارامتری، برگ‌ها محادل نوارها هستند؛ با این تفاوت که نه عرض نوارها ثابت است و نه تعداد نمونه‌ها
- تقسیم‌بندی نواحی محلی، تنها بر اساس شباهت نمونه‌ها انجام نمی‌شود، بلکه فروچی هم در نظر گرفته می‌شود.
- سرعت بیشتری دارند.
- نیاز به ذخیره‌ی همه‌ی نمونه‌ها نیست.

