

بسمه تعالی

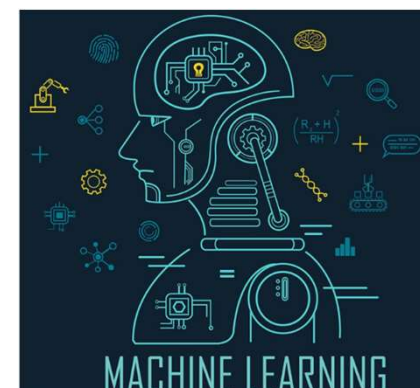
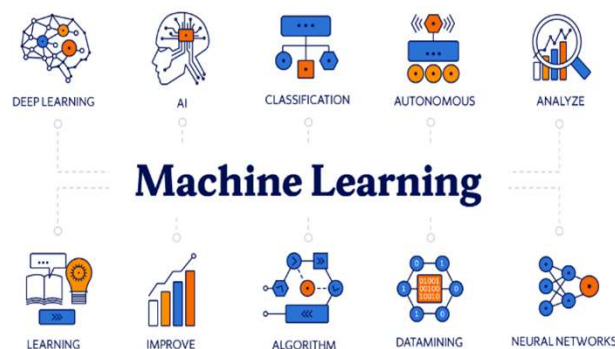
فصل سوم یادگیری ماشین

دسته بندی
Classification

مدرس
فاطمه دارائی

f_daraei@semnan.ac.ir

<https://fdaraei.profile.semnan.ac.ir>





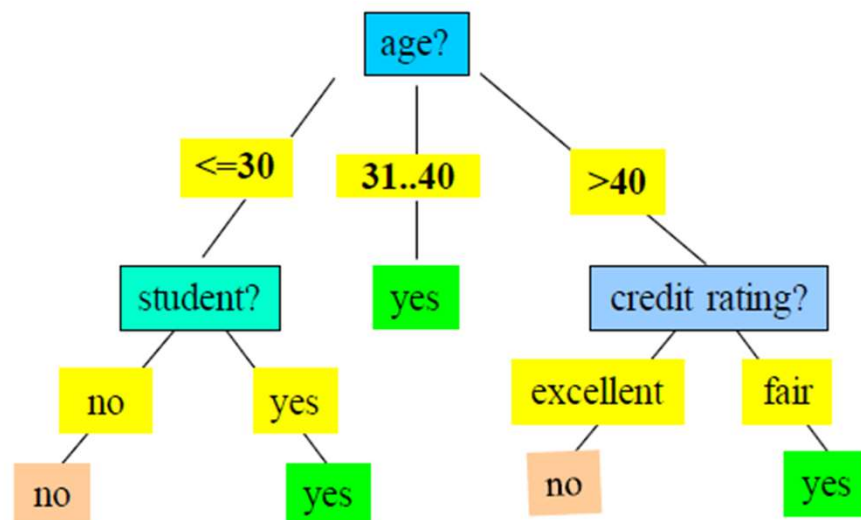
دانشگاه سمنان

دانشگاه سمنان
Semnan University
پردیس فرزنانگان

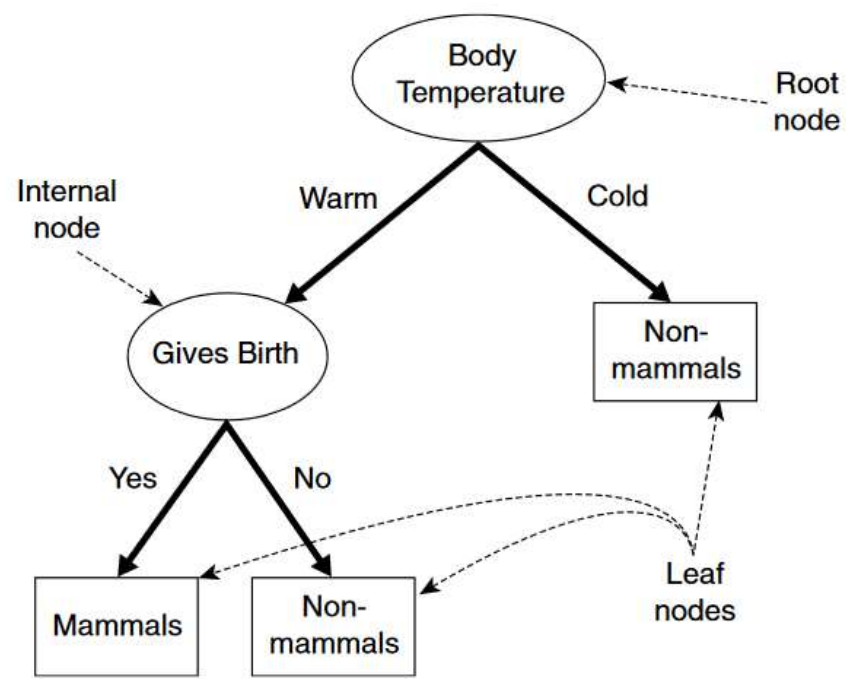
درخت تصمیم Decision Tree

درخت تصمیم، یکی از ابزارهای قوی و متداول برای دسته بندی و پیش بینی می باشد.
درخت تصمیم یک درخت با ساختار فلوچارت مانند است.

age	income	student	credit rating	buys computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

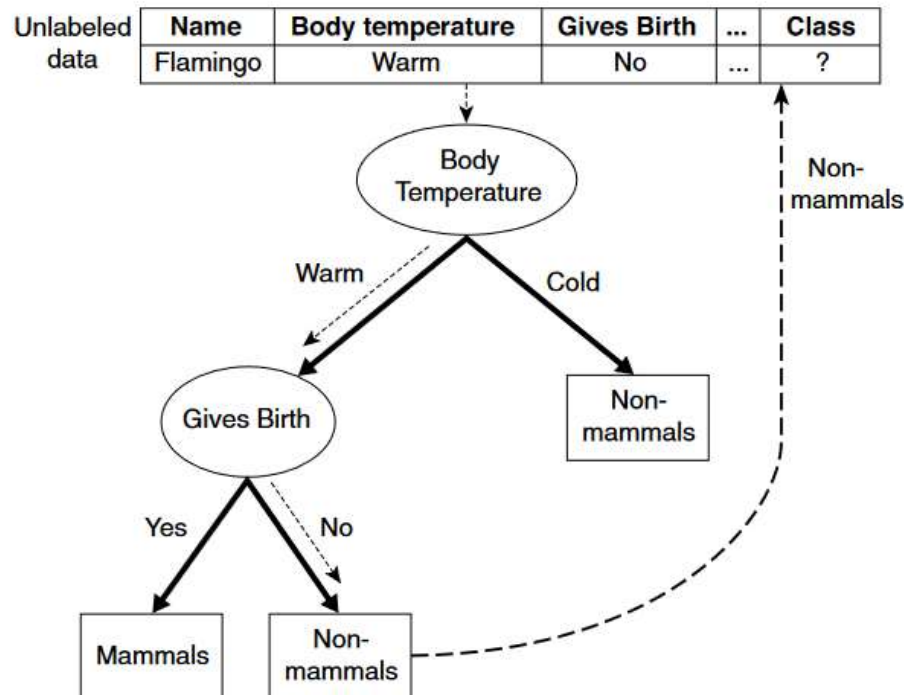


- هر گره غیرپایانی، یک تست (آزمون و سوال) بر روی ویژگی انجام می دهد.
- در هر یک از گره های داخلی، با توجه به یک یا چند صفات خاصه تصمیم گیری صورت می گیرد.
- هر شاخه یا یال یک نتیجه از تست است.
- هر گره برگ، یک برجسب کلاس است.
- بالاترین گره در درخت، گره ریشه است.



نحوه استفاده از درخت تصمیم

- برای یک داده جدید که مقادیر ویژگی آن معلوم است اما برچسبی ندارد.
- بر اساس درخت یک مسیر از ریشه به برگ طی می شود.
- برچسب برگ نشان دهنده کلاس پیش‌بینی شده برای آن موجودیت است.
- با این تفاسیر درخت تصمیم را به راحتی می توان به قوانین دسته بندی تبدیل نمود.



در درخت تصمیم یکسری سوال وجود دارد و با مشخص شدن

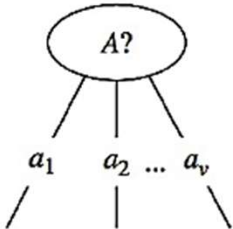
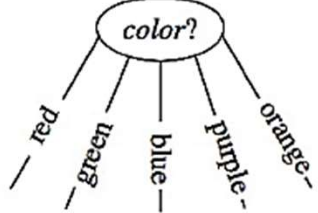
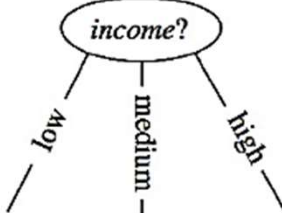
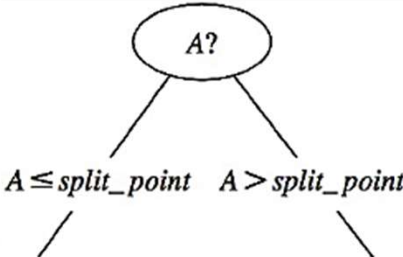
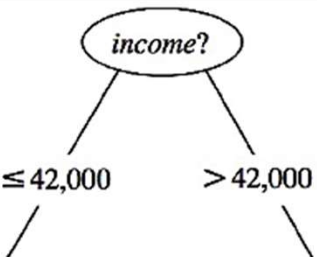
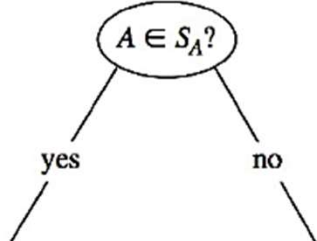
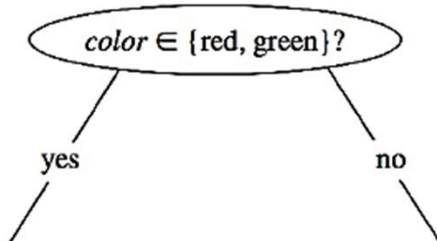
پاسخ هر سوال، یک سوال دیگر پرسیده می شود.

اگر سوالها درست و خوب پرسیده شوند،

با یک مسیر کوتاه پیش بینی دسته رکورد جدید انجام می شود.

انواع تقسیم بندی گره میانی

A is discrete-valued

Partitioning scenarios	Examples	
		
		
		

A is continuous-valued

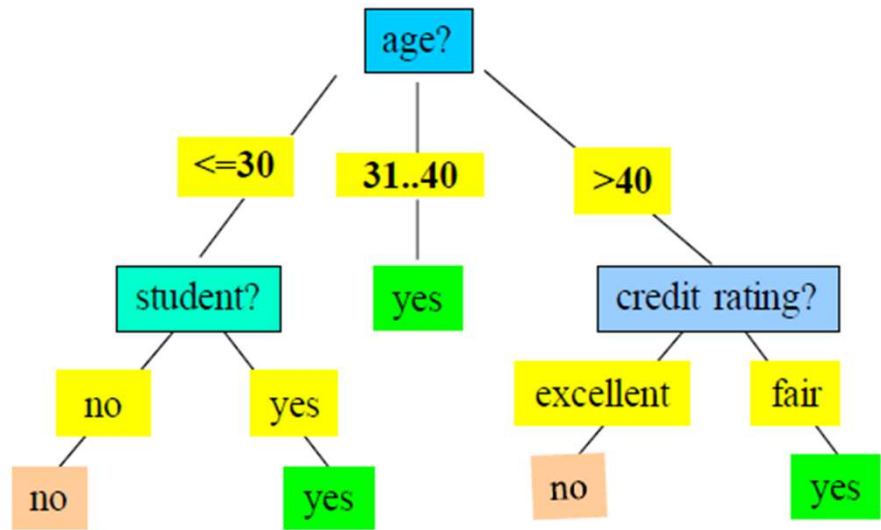
A is discrete-valued
but we want a binary tree



نحوه استفاده از درخت تصمیم

- چگونه درخت‌های تصمیم برای طبقه‌بندی استفاده می‌شوند؟
1. مسیری از ریشه تا یک گره برگ طی می‌شود که پیش‌بینی مربوط به آن صفت را در خود دارد.
 2. ویژگی‌های یک صفت با درخت تصمیم مقایسه می‌شوند.

RID	age	income	student	credit-rating	Class
1	youth	high	no	fair	?





دانشگاه سمنان

دانشگاه سمنان

Semnan University

پردیس فرزندان

چرا درخت تصمیم؟

- (۱) ایجاد درخت تصمیم به هیچ دانش تخصصی یا تنظیم پارامتری احتیاج ندارد، به همین دلیل برای کشف دانش مناسب است.
- (۲) درخت‌های تصمیم‌گیری می‌تواند داده‌های با ابعاد زیاد را پوشش دهند
- (۳) نمایش دانش بدست آمده از ساختار درختی شهودی و به طور کلی آسان است و توسط انسان قابل درک است.
- (۴) آموزش و دسته‌بندی با استفاده از درخت تصمیم‌گیری ساده و سریع می‌باشد.
- (۵) به طور کلی، دسته‌بندی با درخت تصمیم‌گیری از دقت خوبی برخوردار است.
- (۶) الگوریتم درخت تصمیم‌گیری برای دسته‌بندی در بسیاری از زمینه‌های کاربردی قابل استفاده است.

الگوریتم درخت تصمیم: یادگیری

الگوریتم های مختلفی برای ساخت درخت تصمیم وجود دارد:

-  ID3 (Iterative Dichotomiser 3)
-  C4.5 (successor of ID3)
-  CART (Classification And Regression Tree)
-  CHAID (CHi-squared Automatic Interaction Detector)
-  MARS (extends decision trees to handle numerical data better)

- در این الگوریتم ها یک روش حریصانه، بدون بازگشت به عقب، استفاده می شود.
- همچنین راهکار به صورت بالا به پایین و تقسیم و حل است.



دانشگاه سمنان

دانشگاه سمنان

Semnan University

پردیس فرزندان

الگوریتم حریصانه (الگوریتم پایه)

- درخت به صورت بازگشتی با روش تقسیم و حل، از بالا به پایین ساخته می‌شود.
- در آغاز، همه نمونه های آموزشی در ریشه هستند.
- ویژگی ها بر اساس یک معیار آماری انتخاب می‌شوند.
- ویژگی ها چند مقداری هستند (اگر مقادیر پیوسته باشد، گسسته می‌شوند).
- داده ها به صورت بازگشتی و بر اساس ویژگی های انتخاب شده پارتیشن بندی می‌شوند.
- ویژگی های آزمون بر اساس معیارهای اکتشافی و یا آماری انتخاب می‌شوند.

شرایط برای توقف پارتیشن بندی

- همه نمونه ها برای یک گره داده شده متعلق به یک کلاس باشند.
- هیچ ویژگی ای برای پارتیشن بندی بیشتر، باقی نمانده باشد- در این شرایط برچسب اکثریت نمونه های برگ برای دسته بندی به کار گرفته می‌شود. (ناشی از ویژگی های ناکافی یا وجود نویز)
- هیچ نمونه بیشتری وجود نداشته باشد.



دانشگاه سمنان

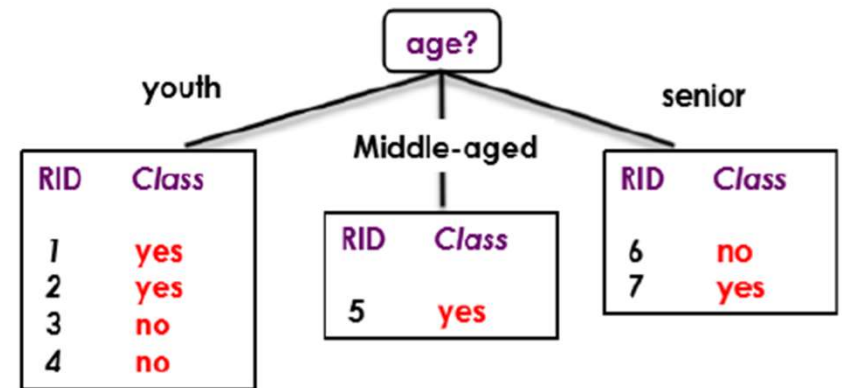
دانشگاه سمنان

Semnan University

پردیس فرزنانگان

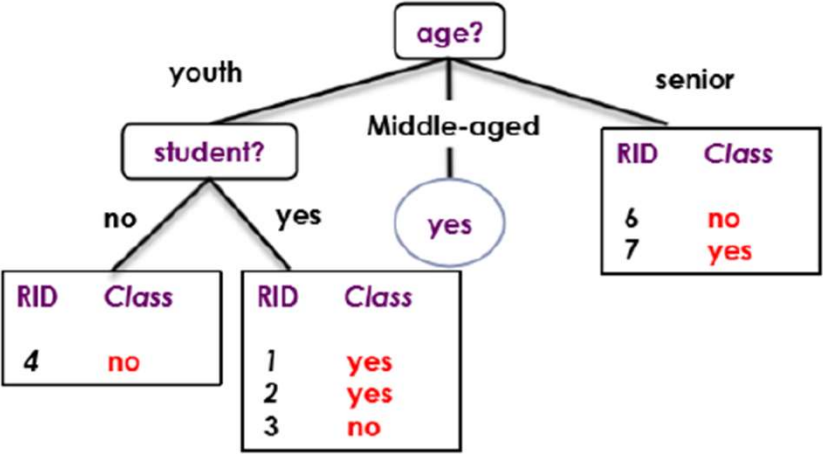
مثال

RID	age	student	credit-rating	Class: buys_computer
1	youth	yes	fair	yes
2	youth	yes	fair	yes
3	youth	yes	fair	no
4	youth	no	fair	no
5	middle-aged	no	excellent	yes
6	senior	yes	fair	no
7	senior	yes	excellent	yes





RID	age	student	credit-rating	Class: buys_computer
1	youth	yes	fair	yes
2	youth	yes	fair	yes
3	youth	yes	fair	no
4	youth	no	fair	no
5	middle-aged	no	excellent	yes
6	senior	yes	fair	no
7	senior	yes	excellent	yes





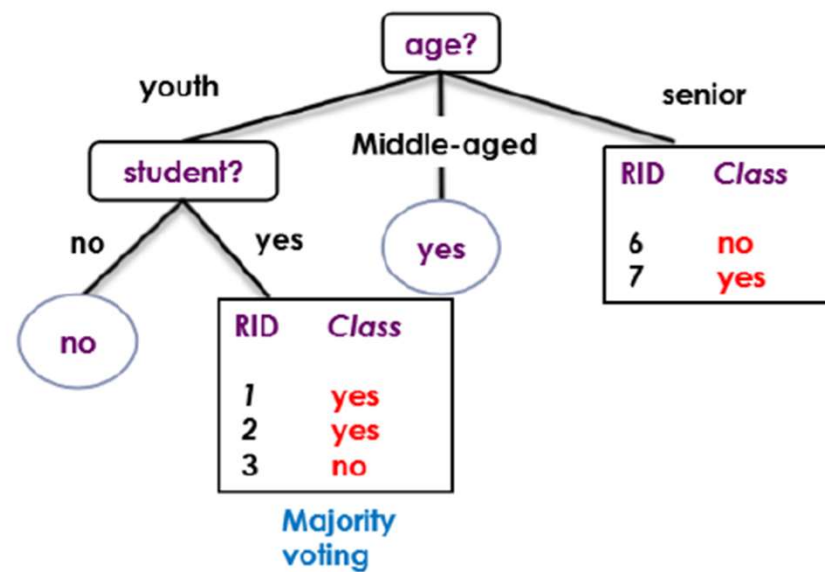
دانشگاه سمنان

دانشگاه سمنان

Semnan University

پردیس فرزنانگان

RID	age	student	credit-rating	Class: buys_computer
1	youth	yes	fair	yes
2	youth	yes	fair	yes
3	youth	yes	fair	no
4	youth	no	fair	no
5	middle-aged	no	excellent	yes
6	senior	yes	fair	no
7	senior	yes	excellent	yes





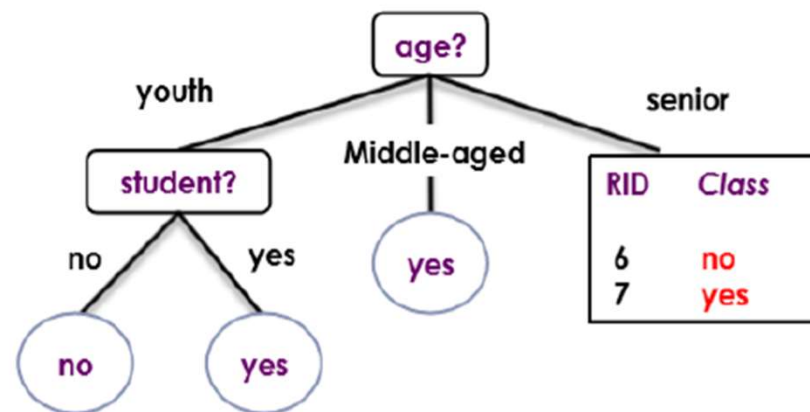
دانشگاه سمنان

دانشگاه سمنان

Semnan University

پردیس فرزنانگان

RID	age	student	credit-rating	Class: buys_computer
1	youth	yes	fair	yes
2	youth	yes	fair	yes
3	youth	yes	fair	no
4	youth	no	fair	no
5	middle-aged	no	excellent	yes
6	senior	yes	fair	no
7	senior	yes	excellent	yes





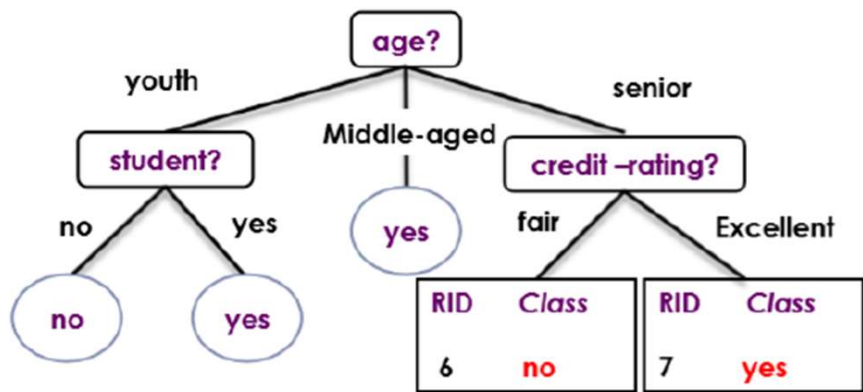
دانشگاه سمنان

دانشگاه سمنان

Semnan University

پردیس فرزندان

RID	age	student	credit-rating	Class: buys_computer
1	youth	yes	fair	yes
2	youth	yes	fair	yes
3	youth	yes	fair	no
4	youth	no	fair	no
5	middle-aged	no	excellent	yes
6	senior	yes	fair	no
7	senior	yes	excellent	yes





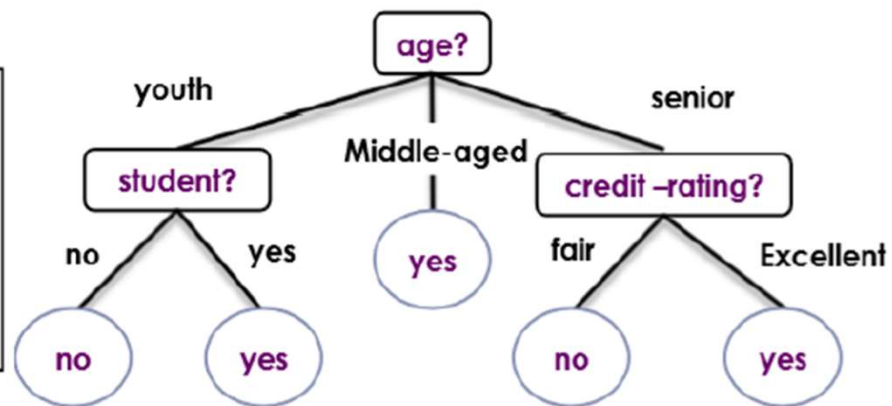
دانشگاه سمنان

دانشگاه سمنان

Semnan University

پردیس فرزندان

RID	age	student	credit-rating	Class: buys_computer
1	youth	yes	fair	yes
2	youth	yes	fair	yes
3	youth	yes	fair	no
4	youth	no	fair	no
5	middle-aged	no	excellent	yes
6	senior	yes	fair	no
7	senior	yes	excellent	yes





معیار انتخاب صفات و ویژگی

معیار انتخاب ویژگی (قاعده‌ی تقسیم) یک روش ابتکاری برای انتخاب معیار تقسیم‌بندی است که به **بهترین** شکل، یک بخش داده را جدا می‌کند. ایده‌آل این است که:

- هر بخش حاصل باید خالص باشد (شامل نمونه‌هایی که همگی به یک کلاس تعلق دارند).
- برای هر ویژگی، امتیازی تعیین شده و ویژگی با بالاترین امتیاز انتخاب می‌شود.
- می‌تواند یک نقطه‌ی تقسیم برای ویژگی‌های پیوسته یا یک زیرمجموعه‌ی تقسیم برای درخت‌های دودویی تعیین کند.

به بررسی معیارهای **Information Gain**، **Gain Ratio** و **Gini Index** خواهیم پرداخت.



Information Gain معیار

بهره اطلاعات همان مفهوم آنتروپی اطلاعات است.

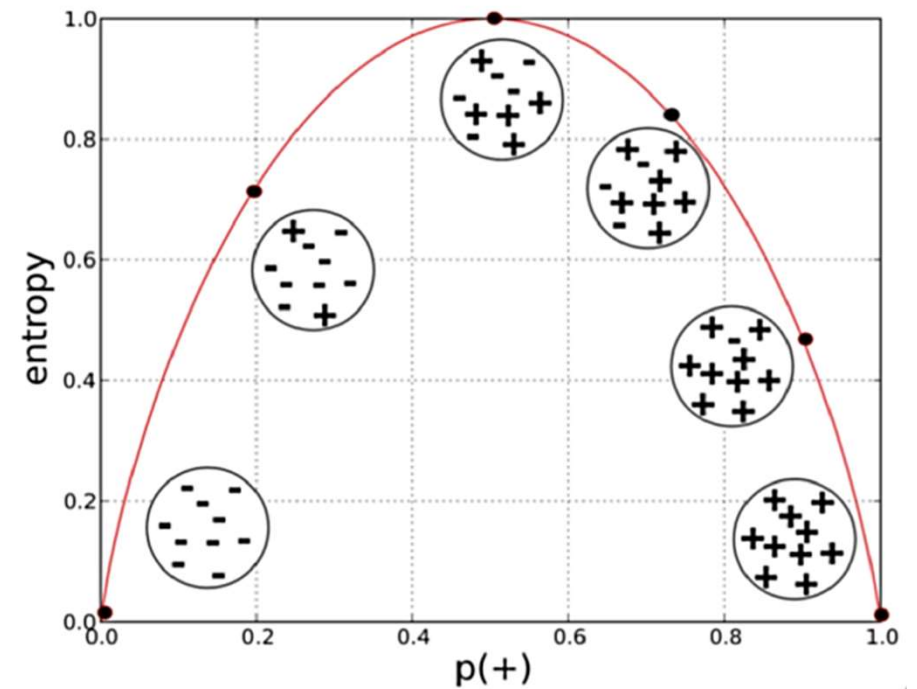
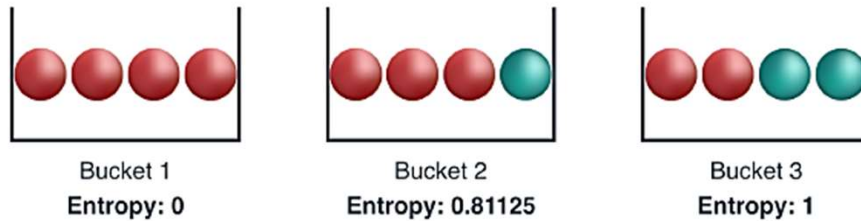
آنتروپی اطلاعات مفهومی از نظریه اطلاعات است. این مفهوم بیان می کند که در یک رویداد چه میزان اطلاعات وجود دارد.

به طور کلی، هر چه یک رویداد نامطمئن تر یا تصادفی تر باشد، میزان اطلاعات موجود در آن بیشتر خواهد بود.

فرمول زیر محاسبه آنتروپی را نشان می دهد که H همان آنتروپی اطلاعات می باشد:

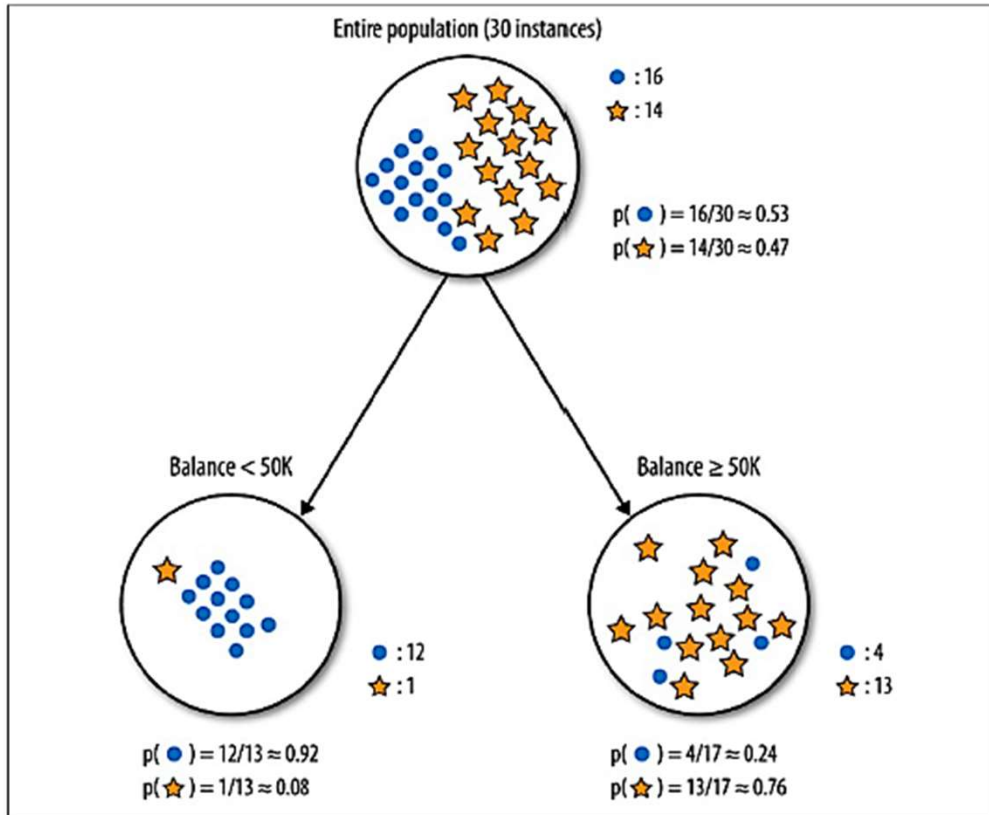
$$\begin{array}{l} X \in \{x_1, x_2, \dots, x_m\} \\ \{p_1, p_2, \dots, p_m\} \end{array} \quad \Rightarrow \quad H(X) = - \sum_{i=1}^m p_i \log_2(p_i)$$

دنبال معیاری هستیم که آنتروپی آن کمتر باشد.
 آنتروپی = عدم قطعیت



<https://medium.com/udacity/shannon-entropy-information-gain-and-picking-balls-from-buckets-5810d35d54b4>

Source: Data Science for Business



Source: Data Science for Business

آنتروپی بالا

- از یک توزیع یکنواخت (هیستوگرام مسطح) آمده است.

- مقادیر نمونه‌گیری شده از آن کمتر قابل پیش‌بینی هستند.

آنتروپی پایین

- از یک توزیع متغیر (هیستوگرام دارای قله‌ها و دره‌های متعدد) آمده است.
- مقادیر نمونه‌گیری شده از آن بیشتر قابل پیش‌بینی هستند.





محاسبه Information Gain

مرحله ۱: محاسبه آنتروپی مجموعه D

آنتروپی، میانگین مقدار اطلاعات مورد نیاز برای شناسایی برچسب کلاس یک نمونه در D را نشان می‌دهد.

m : تعداد کلاس‌ها

p_i : احتمال این که یک نمونه دلخواه در D متعلق به کلاس i باشد. این احتمال با رابطه زیر تخمین زده می‌شود:

$$Info(D) = - \sum_{i=1}^m p_i \log_2(p_i)$$

یک تابع لگاریتم به پایه ۲ استفاده می‌شود زیرا اطلاعات به صورت بیت کدگذاری می‌گردد.



دانشگاه سمنان

دانشگاه سمنان

Semnan University

پردیس فرزندان

RID	age	income	student	credit-rating	class:buy_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle-aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle-aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle-aged	medium	no	excellent	yes
13	middle-aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

9 tuples in class yes
5 tuples in class no



$$Info(D) = -\frac{9}{14} \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \log_2\left(\frac{5}{14}\right) = 0.940 \text{ bits}$$

مرحله ۲: محاسبه آنترופی وزنی برای هر ویژگی پس از تقسیم‌بندی با استفاده از آن ویژگی

فرض کنید می‌خواهیم نمونه‌های موجود در D را بر اساس یک ویژگی A تقسیم‌بندی کنیم.
برای این کار:

تقسیم‌بندی بر اساس ویژگی A : هر مقدار ممکن از A یک بخش جدید ایجاد می‌کند. این بخش‌ها مجموعه‌های فرعی مختلفی از D خواهند بود.

محاسبه آنترופی وزنی برای هر بخش: برای هر بخش ایجادشده توسط مقدارهای مختلف ویژگی A ، آنترופی محاسبه می‌شود و هر آنترופی متناسب با اندازه بخش خود وزن‌دهی می‌شود.

$$Info_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times Info(D_j)$$



دانشگاه سمنان

دانشگاه سمنان

Semnan University

پردیس فرزنانگان

الگوریتم *ID3*

این الگوریتم یکی از ساده‌ترین الگوریتم‌های درخت تصمیم است که از معیار *Information Gain* استفاده می‌کند. در اجرای این الگوریتم دو شرط توقف وجود دارد. یکی این که کلیه‌ی نمونه‌های باقیمانده متعلق به یک کلاس باشند و یا اینکه پس از محاسبه‌ی مقدار معیار *Information Gain* بهترین آن بزرگتر از صفر نباشد. هیچگونه روش هرس کردنی در آن موجود نیست و می‌تواند صفات خاصه‌ی عددی و داده‌های ناقص را به عنوان ورودی بپذیرد.



دانشگاه سمنان

دانشگاه سمنان

Semnan University

پردیس فرزندان

RID	age	income	student	credit-rating	class:buy_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle-aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle-aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle-aged	medium	no	excellent	yes
13	middle-aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

$$Info_{age}(D) = \frac{5}{14} \times \left(-\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} \right) + \frac{4}{14} \times \left(-\frac{4}{4} \log_2 \frac{4}{4} \right) + \frac{5}{14} \times \left(-\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \right) = 0.694 \text{ bits.}$$

مرحله ۳: محاسبه بهره اطلاعاتی **Information Gain**

$$Gain(A) = Info(D) - Info_A(D)$$

ویژگی با بیشترین Information Gain به عنوان ویژگی تقسیم انتخاب می شود. این ویژگی بالاترین قدرت را در جداسازی داده ها دارد و با کاهش بیشترین میزان ابهام، بیشترین اطلاعات را فراهم می کند.

$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit_rating) = 0.048$$



دانشگاه سمنان

دانشگاه سمنان
Semnan University
پردیس فرزانهگان

مثال

جدول ۲-۷: نمونه‌ای از یک داده‌ی آزمایشی

ID	Age	Income	Job	Computer
1	Old	Medium	Student	No
2	Middle	High	Teacher	No
3	Old	Low	Teacher	No
4	Young	Medium	Teacher	Yes
5	Young	Low	Teacher	Yes
6	Old	Medium	Student	Yes
7	Middle	Medium	Student	Yes
8	Young	High	Teacher	No
9	Old	High	Student	No
10	Middle	High	Student	No

$$Entropy(D) = -\frac{4}{10} \log_2\left(\frac{4}{10}\right) - \frac{6}{10} \log_2\left(\frac{6}{10}\right) = 0.970$$

$$Domain(Age) = \{Old, Middle, Young\}$$

$$Domain(Income) = \{High, Medium, Low\}$$

$$Domain(Job) = \{Teacher, Student\}$$

$$Entropy_{Age}(D) = \frac{4}{10} \times \left(-\frac{3}{4} \log_2\left(\frac{3}{4}\right) - \frac{1}{4} \log_2\left(\frac{1}{4}\right)\right) + \frac{3}{10} \times \left(-\frac{2}{3} \log_2\left(\frac{2}{3}\right) - \frac{1}{3} \log_2\left(\frac{1}{3}\right)\right) + \frac{3}{10} \times \left(-\frac{1}{3} \log_2\left(\frac{1}{3}\right) - \frac{2}{3} \log_2\left(\frac{2}{3}\right)\right) = 0.875$$

$$Entropy_{Income}(D) = \frac{4}{10} \times \left(-\frac{4}{4} \log_2\left(\frac{4}{4}\right) - \frac{0}{4} \log_2\left(\frac{0}{4}\right)\right) + \frac{4}{10} \times \left(-\frac{1}{4} \log_2\left(\frac{1}{4}\right) - \frac{3}{4} \log_2\left(\frac{3}{4}\right)\right) + \frac{2}{10} \times \left(-\frac{1}{2} \log_2\left(\frac{1}{2}\right) - \frac{1}{2} \log_2\left(\frac{1}{2}\right)\right) = 0.524$$

$$Entropy_{Job}(D) = \frac{5}{10} \times \left(-\frac{3}{5} \log_2\left(\frac{3}{5}\right) - \frac{2}{5} \log_2\left(\frac{2}{5}\right)\right) + \frac{5}{10} \times \left(-\frac{3}{5} \log_2\left(\frac{3}{5}\right) - \frac{2}{5} \log_2\left(\frac{2}{5}\right)\right) = 0.970$$



دانشگاه سمنان

دانشگاه سمنان

Semnan University

پردیس فرزنانگان

$$Entropy(D) = -\frac{4}{10} \log_2\left(\frac{4}{10}\right) - \frac{6}{10} \log_2\left(\frac{6}{10}\right) = 0.970$$

$$InformationGain(Age) = 0.970 - 0.875 = 0.095$$

$$InformationGain(Income) = 0.970 - 0.524 = 0.446$$

$$InformationGain(Job) = 0.970 - 0.970 = 0$$

به دلیل اینکه صفت خاصه‌ی درآمد دارای بیشترین مقدار است، برای ریشه‌ی درخت انتخاب می‌شود. چون در دامنه‌ی این صفت خاصه می‌توان ۳ مقدار متمایز یافت، بنابراین سه شاخه از این گره منشعب می‌شود که هر یک با مقادیر سه گانه‌ی این صفت خاصه برچسب خورده‌اند

در واقع داده‌های آموزشی به ۳ زیرمجموعه افراز می‌شوند. برای هر یک از این زیرمجموعه‌ها همانند قبل بهترین صفت خاصه جهت انشعاب محاسبه می‌شود. این بار صفت خاصه‌ی درآمد در محاسبات شرکت نمی‌کند و از میان دیگر صفات یکی انتخاب می‌شود. این کار تا هنگامی که یکی از شروط توقف الگوریتم محقق شود، ادامه می‌یابد.

Gain Ratio

Information Gain به ویژگی‌هایی که دارای تعداد زیادی مقادیر هستند تمایل دارد. به عنوان مثال، این اندازه‌گیری تمایل به انتخاب ویژگی‌هایی مانند شماره دانشجویی دارد که اصلاً ویژگی مفیدی برای طبقه‌بندی نیست.

الگوریتم C4.5 از نسبت بهره Gain Ratio استفاده می‌کند تا این مشکل را برطرف کند (نرمال‌سازی بهره اطلاعات).

$$SplitInfo_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times \log_2 \left(\frac{|D_j|}{|D|} \right)$$

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$



مثال Gain Ratio

RID	age	income	student	credit-rating	class:buy_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle-aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle-aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle-aged	medium	no	excellent	yes
13	middle-aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

$$SplitInfo_{income}(D) = -\frac{4}{14} \times \log_2 \left(\frac{4}{14} \right) - \frac{6}{14} \times \log_2 \left(\frac{6}{14} \right) - \frac{4}{14} \times \log_2 \left(\frac{4}{14} \right) = 1.557$$

$$Gain(income) = 0.029 \quad \rightarrow \quad GainRatio(income) = 0.029/1.557 = 0.019.$$



معیار Gini Index

شاخص جینی در الگوریتم CART استفاده می‌شود و میزان ناخالصی (یا عدم خلوص) یک مجموعه داده را اندازه‌گیری می‌کند.

اگر مجموعه داده D شامل نمونه‌هایی از m کلاس مختلف باشد، $\text{gini}(D)$ به صورت زیر تعریف می‌شود:

$$\text{Gini}(D) = 1 - \sum_{i=1}^m p_i^2,$$

اگر مجموعه داده D بر اساس ویژگی A به دو زیرمجموعه D_1 و D_2 تقسیم شود، به صورت زیر تعریف می‌شود:

$$\text{Gini}_A(D) = \frac{|D_1|}{|D|} \text{Gini}(D_1) + \frac{|D_2|}{|D|} \text{Gini}(D_2)$$

$$\Delta \text{Gini}(A) = \text{Gini}(D) - \text{Gini}_A(D)$$



RID	age	income	student	credit-rating	class:buy_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle-aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle-aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle-aged	medium	no	excellent	yes
13	middle-aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

$$Gini_{income \in \{low, medium\}}(D)$$

$$= \frac{10}{14} Gini(D_1) + \frac{4}{14} Gini(D_2)$$

$$= \frac{10}{14} \left(1 - \left(\frac{7}{10} \right)^2 - \left(\frac{3}{10} \right)^2 \right) + \frac{4}{14} \left(1 - \left(\frac{2}{4} \right)^2 - \left(\frac{2}{4} \right)^2 \right)$$

$$= 0.443$$

$$= Gini_{income \in \{high\}}(D).$$

$Gini_{\{low, high\}}$ is 0.458; $Gini_{\{medium, high\}}$ is 0.450. Thus, split on the $\{low, medium\}$ (and $\{high\}$) since it has the lowest Gini index



دانشگاه سمنان

دانشگاه سمنان

Semnan University

پردیس فرزانهگان

سن، داشتن شغل و منزل مسکونی و همچنین اعتبار مشتریان در تصمیم‌گیری برای اعطای وام موثر هستند.

از آنجا که ۱۰ نمونه داده به دو کلاس *Yes* و *No* به ترتیب به نسبت ۶ و ۴ توزیع شده است، داریم:

$$Gini(D) = 1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2 = 0.48$$

جهت یافتن بهترین صفت‌خاصه، معیار *Gini Index* برای کلیه صفات‌خاصه محاسبه می‌شود. در مجموعه دامنه‌ی دو صفت‌خاصه‌ی شغل و منزل مسکونی می‌توان دو مقدار *True* و *False* را یافت و به همین دلیل محاسبه‌ی معیار خیلی سخت نیست.

$$Gini_{Job}(D) = \frac{7}{10} \times \left(1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2\right) + \frac{3}{10} \times \left(1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2\right) = 0.343$$

$$Gini_{House}(D) = \frac{5}{10} \times \left(1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2\right) + \frac{5}{10} \times \left(1 - \left(\frac{0}{5}\right)^2 - \left(\frac{5}{5}\right)^2\right) = 0.160$$

جدول ۷-۳ اطلاعات مشتریانی که درخواست وام داشتند را نشان می‌دهد.

جدول ۷-۳: اطلاعاتی جهت تصمیم‌گیری برای اعطای وام

ID	Age	Job	House	Credit	Class
1	Old	False	True	Excellent	Yes
2	Old	False	True	Good	Yes
3	Middle	False	False	Fair	No
4	Middle	True	True	Good	Yes
5	Young	False	False	Fair	No
6	Old	False	False	Fair	No
7	Middle	False	True	Excellent	Yes
8	Young	True	False	Good	Yes
9	Young	True	True	Fair	Yes
10	Middle	False	False	Good	No



دانشگاه سمنان

دانشگاه سمنان

Semnan University

پردیس فرزنانگان

اما صفات خاصه‌ی سن و اعتبار هر یک دارای سه مقدار هستند و از آنجا که معیار *Gini* *Index* یک انشعاب دودویی را برای هر یک از این صفات خاصه می‌سازد، باید کلیه‌ی حالات تقسیم مقادیر برای این صفات خاصه بررسی شوند.

$Gini_{Age}(D)=0.476$ وقتی $\{Old\}$, $\{Middle, Young\}$

$Gini_{Age}(D)=\underline{0.467}$ وقتی $\{Middle\}$, $\{Old, Young\}$

$Gini_{Age}(D)=0.476$ وقتی $\{Young\}$, $\{Middle, Old\}$

$Gini_{Credit}(D)=0.400$ وقتی $\{Excellent\}$, $\{Good, Fair\}$

$Gini_{Credit}(D)=0.450$ وقتی $\{Good\}$, $\{Excellent, Fair\}$

$Gini_{Credit}(D)=\underline{0.317}$ وقتی $\{Fair\}$, $\{Excellent, Good\}$

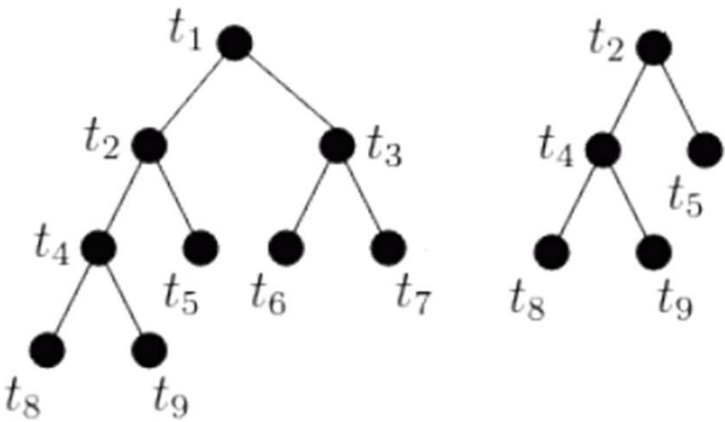
همانطور که ملاحظه می‌کنید بهترین انشعاب برای صفت خاصه‌ی سن هنگامی است که مقادیر پیر و جوان در یک گروه و میانسال در گروه دیگر قرار می‌گیرند و مقدار *Gini* *Index* برابر با ۰/۴۶۷ است. بطور مشابه مقدار این معیار برای صفت خاصه‌ی اعتبار برابر با ۰/۳۱۷ می‌باشد. در مرحله‌ی اول از میان صفات خاصه، منزل مسکونی با مقدار ۰/۱۶ انتخاب می‌شود. با تکرار عملیات فوق در نهایت ما با یک درخت تصمیم دودویی روبرو خواهیم بود.



هرس کردن Pruning

برای ساده‌تر کردن مدل و کاهش احتمال بیش‌برازش overfitting
اصل تیغ اوکام: Occam's razor

از بین دو توضیح برای چیزی، توضیحی که به احتمال زیاد درست است، ساده‌ترین آن‌هاست.



پیش هرس و پس هرس Pre-pruning vs. Post-pruning

پیش-هرس کردن

قبل از اینکه الگوریتم به یک درخت کامل رشد کند، متوقفش کنید.

پس-هرس کردن

درخت تصمیم را به طور کامل رشد دهید.
گره‌های درخت تصمیم را به صورت از پایین به بالا هرس کنید.
اگر پس از هرس کردن، خطای تعمیم بهبود یافت، زیردرخت را با یک گره برگ جایگزین کنید.
برچسب کلاس گره برگ از کلاس اکثریت نمونه‌ها در زیردرخت تعیین می‌شود.

پیش-هرس کردن سریع‌تر است اما دقت کمتری دارد.

ارزیابی Evaluation

برای مقایسه‌ی مدل‌های مختلف

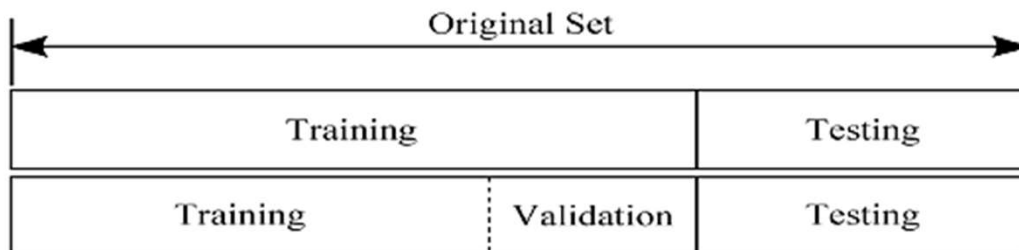
برای تنظیم اَبرمُتغیرها مانند:

K در KNN، تعداد لایه‌ها در شبکه‌های عصبی، بهترین هرس در درخت تصمیم و غیره.

هدف اصلی در یادگیری ماشین، **تعمیم generalization** است. ما می‌خواهیم توانایی تعمیم مدل خود را اندازه‌گیری کنیم.

روش نگه‌داشتن **Hold-out**: شما مدل را روی داده‌های آموزش، آموزش می‌دهید و آن را روی داده‌های اعتبارسنجی **validation** ارزیابی می‌کنید. وقتی مدل آماده شد، یکبار دیگر آن را روی داده‌های آزمون نهایی تست می‌کنید.

داده‌ها را قبل از تقسیم به هم بزنید (Shuffle کنید)



ارزیابی Evaluation

چرا فقط آموزش و تست را انجام ندهیم؟

ما این تنظیمات را با استفاده از عملکرد مدل روی داده‌های اعتبارسنجی به عنوان مرحله بازخورد انجام می‌دهیم.



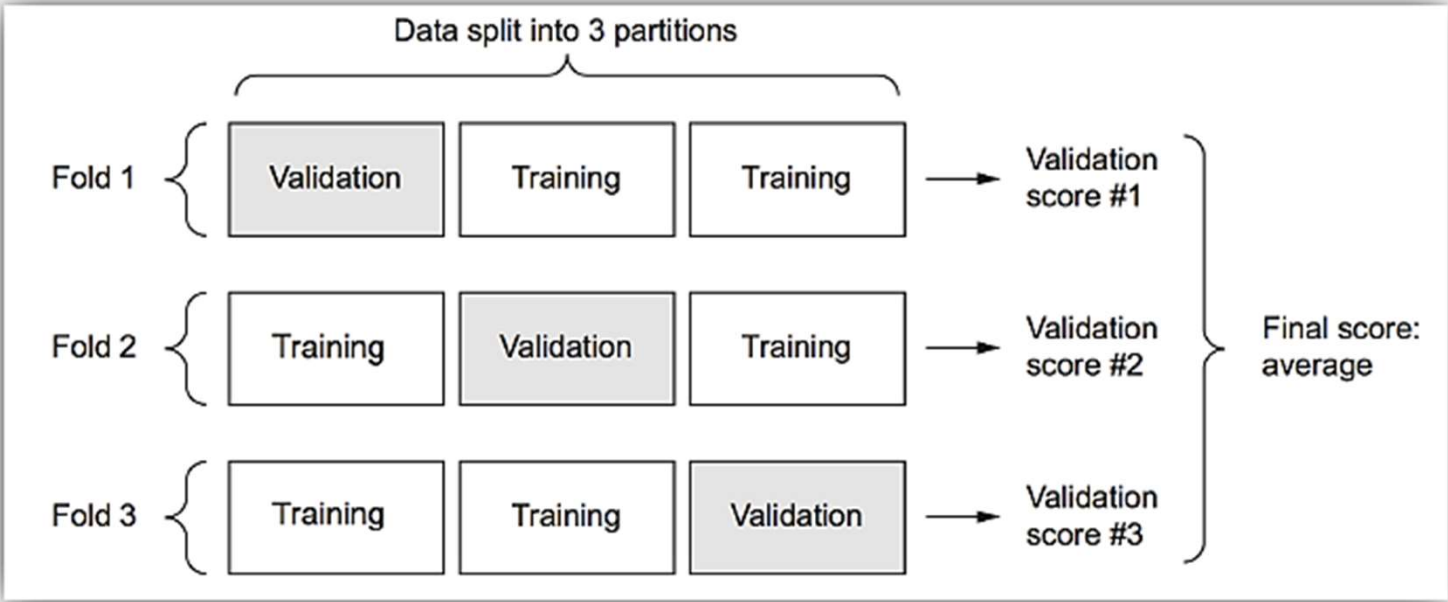
اما این کار می‌تواند به سرعت منجر به بیش‌برازش **overfitting** نسبت به مجموعه‌ی اعتبارسنجی شود (نشت اطلاعات).

در نهایت، ممکن است مدلی داشته باشید که روی داده‌های اعتبارسنجی به صورت مصنوعی عملکرد خوبی دارد، اما روی داده‌های کاملاً جدید خوب عمل نمی‌کند.



k-fold Cross Validation

وقتی داده‌های کمی دارید، مجموعه‌ی اعتبارسنجی بسیار کوچک خواهد بود. این موضوع مانع از ارزیابی قابل اعتماد مدل می‌شود. بنابراین، از **k-fold Cross Validation** استفاده می‌کنیم. مقادیر معمول برای k ۵ و ۱۰ هست.





دانشگاه سمنان

دانشگاه سمنان
Semnan University
پردیس فرزنانگان

معیارهای ارزیابی مدل دسته بندی

تعداد سوابق آزمونی که به درستی (یا نادرست) توسط مدل طبقه بندی پیش بینی شده‌اند.

ماتریس سردرگمی: Confusion matrix

- مثبت صحیح: True Positive شخص بیمار، به درستی بیمار تشخیص داده شود.
- مثبت کاذب: False Positive شخص سالم، به اشتباه بیمار تشخیص داده شود.
- منفی صحیح: True Negative شخص سالم، به درستی سالم تشخیص داده شود.
- منفی کاذب: False Negative شخص بیمار، به اشتباه سالم تشخیص داده شود.

اگر m کلاس داشته باشیم، $CM(i,j)$ در یک ماتریس سردرگمی، تعداد تاپل های کلاس i را نشان می دهد که توسط طبقه بندی کننده به عنوان کلاس j برچسب گذاری شده اند.

Predict class True class	Positive	Negative
Positive	TP	FN
Negative	FP	TN



معیارهای ارزیابی مدل دسته بندی

Accuracy صحت: درصد نمونه هایی که درست دسته بندی شده اند.

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

Actual Class	Predicted Class	
	Class = 1	Class = 0
	Class = 1	Class = 0
Class = 1	f_{11}	f_{10}
Class = 0	f_{01}	f_{00}

Error rate: $1 - accuracy$, or
Error rate = (FP + FN)/All

دقت Accuracy = $\frac{\text{\# correct predictions}}{\text{total \# of predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$

نرخ خطا Error rate = $\frac{\text{\# wrong predictions}}{\text{total \# of predictions}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$

حساسیت و تشخیص

- مشکل عدم تعادل کلاس: یک کلاس ممکن است نادر باشد، به عنوان مثال HIV مثبت
- اکثریت قابل توجه طبقه منفی و اقلیت طبقه مثبت یا برعکس
- در این حالتها از معیارهای زیر استفاده می کنیم:
- **Sensitivity:** True Positive recognition rate
 - **Sensitivity = TP/P**
- **Specificity:** True Negative recognition rate
 - **Specificity = TN/N**

A\P	C	-C	
C	TP	FN	P
-C	FP	TN	N
	P'	N'	All

به بیان ریاضی، حساسیت حاصل تقسیم موارد مثبت واقعی به حاصل جمع موارد مثبت واقعی و موارد منفی کاذب است.
تشخیص حاصل تقسیم موارد منفی واقعی به حاصل جمع موارد منفی واقعی و مثبت کاذب است.



Precision

دقت (Precision): چند درصد از تاپل هایی که طبقه بندی کننده به عنوان مثبت برچسب گذاری کرده است در واقع مثبت هستند.

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

Example

Actual Class\Predicted class	cancer = yes	cancer = no	Total
cancer = yes	90	210	300
cancer = no	140	9560	9700
Total	230	9770	10000

$$Precision = 90/230 = 39.13\%$$



Recall

فراخوانی (Recall): چند درصد از تاپل های مثبت طبقه بندی کننده به عنوان مثبت برچسب زده است؟

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

Example

Actual Class\Predicted class	cancer = yes	cancer = no	Total
cancer = yes	90	210	300
cancer = no	140	9560	9700
Total	230	9770	10000

$$Recall = 90/300 = 30\%$$

F-measures

F-measures (F1 , F score) میانگین هارمونیک بین Precision و Recall

$$F = \frac{2 \times precision \times recall}{precision + recall} = \frac{2TP}{2TP + FP + FN}$$

معیار وزنی Precision و Recall
 β بار از وزن را به درست بودن اختصاص می دهد

$$F_{\beta} = \frac{(1 + \beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$$

معیارهای ارزیابی Regression

$$MAE = \frac{1}{m} \sum_{i=1}^m |y^{(i)} - \hat{y}^{(i)}|$$

$$MSE = \frac{1}{m} \sum_{i=1}^m (y^{(i)} - \hat{y}^{(i)})^2$$