数据挖掘与商务分析

第3讲 分类分析

主讲教师：肖升生

课程导入：分类分析实例

- 例子1：根据诊断数据预测肿瘤是恶性还是良性

良性

恶性

- 例子2：根据信用卡的消费数据判定消费是否异常

- 例子3：根据司机驾驶行为数据判定驾驶风险等级

高风险
中风险
低风险

讲授提纲

01 分类分析基本概念

02 决策树分类模型

03 朴素贝叶斯模型

04 逻辑回归模型

05 分类模型评估

06 商务案例分析

讲授提纲

01 分类分析基本概念

02 决策树分类模型

03 朴素贝叶斯模型

04 逻辑回归模型

05 分类模型评估

06 商务案例分析

什么是分类分析？

- 分类就是通过对已有信息的学习得到一个目标函数f，通过这个目标函数把数据集中每个样本点映射到一个预先定义好的类标号Y中。

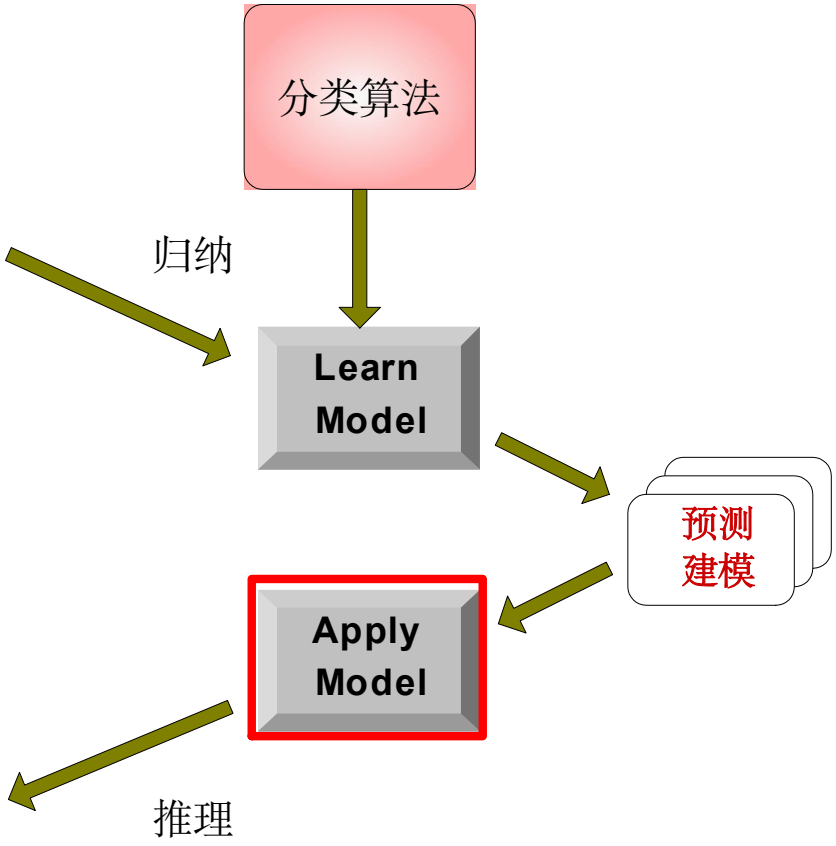
- 分类通常涉及：**训练数据集** + **测试数据集**

- 分类分析的两个主要步骤：
  - 第一步:通过归纳分析训练样本集来建立分类模型得到分类规则
  - 第二步：先用已知的检验样本集评估分类规则的准确率，如果准确率是可以接受的，则使用该模型对未知类标号的样本集进行预测

什么是分类分析？

Tid	偿还借款	婚姻状况	年收入	是否欺诈
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

训练集



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	?

测试集

常用的分类方法

- 经常使用的分类计算与方法：
  - Decision Tree （决策树）
  - Naïve Bayesian（朴素贝叶斯）
  - KNN（K-近邻分类方法）
  - Support Vector Machines（支持向量机）
  - Neural Networks（人工神经网络）
  - …

讲授提纲

01 分类分析基本概念

02 决策树分类模型

03 朴素贝叶斯模型

04 逻辑回归模型

05 分类模型评估

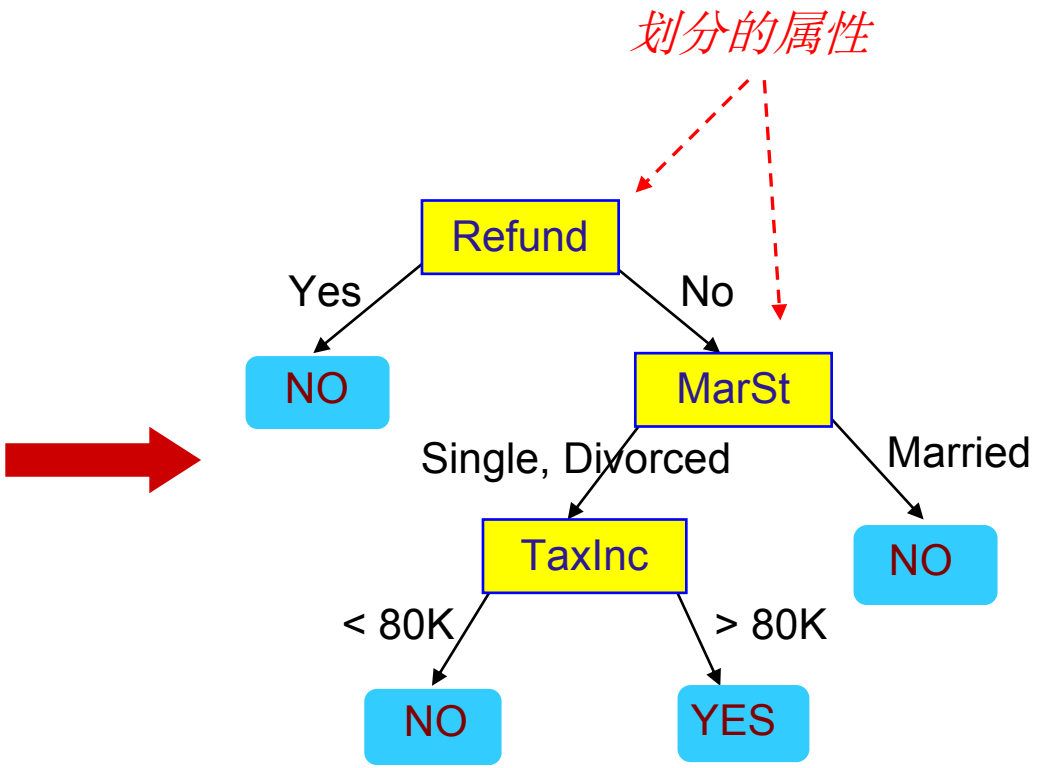
06 商务案例分析

什么是决策树分类模型？

- 决策树是一树状结构，它的每一个树结点可以是叶节点，对应着某一类，也可以对应着一个划分。

	categorical	categorical	continuous	class
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

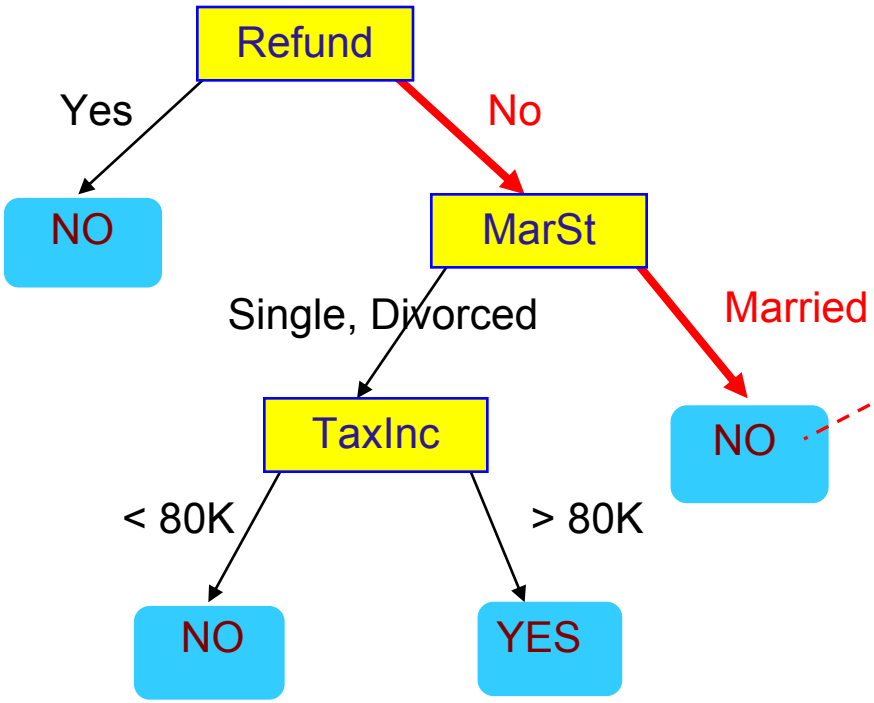
训练数据集



决策树

决策树模型的使用

- 对于训练好的决策树模型如何使用？



训练好的决策树模型

待预测的样本点

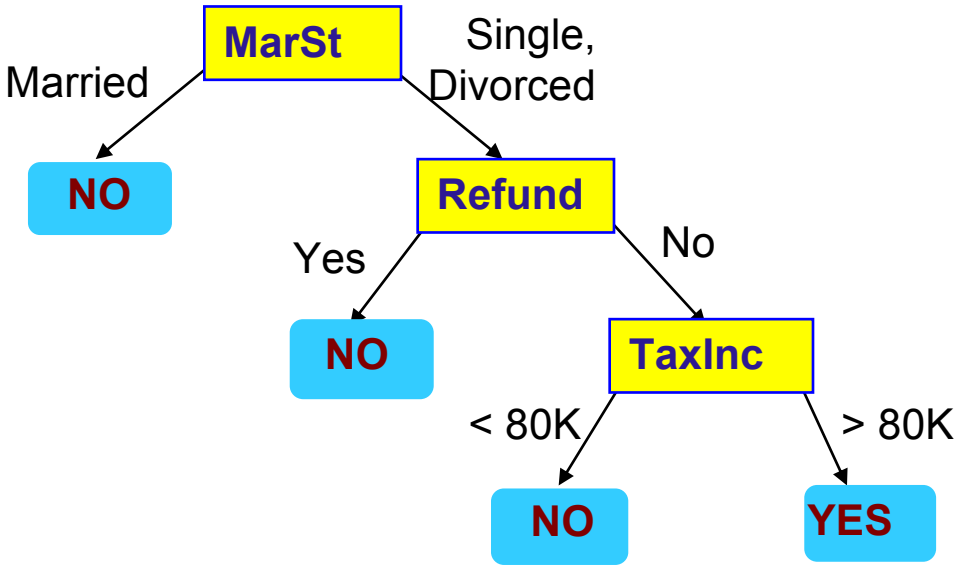
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

判定类标签为 “No”

什么是决策树分类模型？

- 同一个数据集构建的另外一颗决策树

	categorical	categorical	continuous	class
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



给定同一个数据集，可以构建非常多的且符合定义要求的决策树。我们应该如何构建呢？

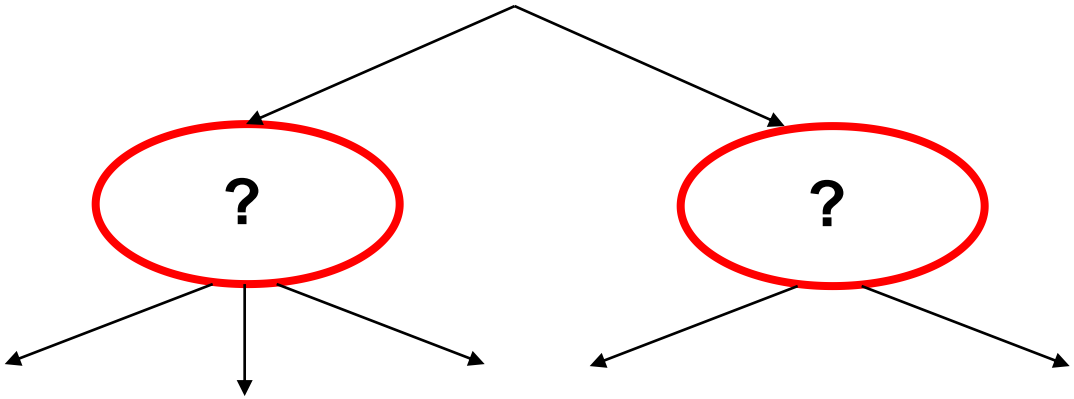
如何构建决策树？

- 决策树的生成需要解决两个问题：
  - **如何对数据中的样本点进行划分？（如何分裂）：**
    - 指定什么属性对样本点进行划分？
    - 给定同一属性的不同划分方式，如何评估其优劣？
  - **什么时候停止样本的划分？(如何终止)**
    - 决策树构建到什么时候应该停止？

如何构建决策树？

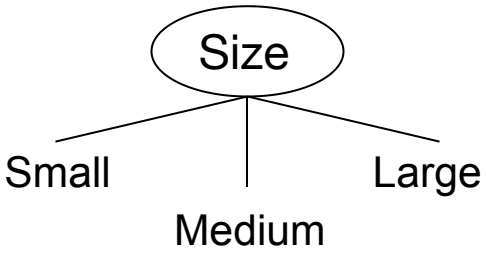
- 例如：给定如下数据集，如何筛选属性对样本点进行划分，进而构建出一颗决策树？

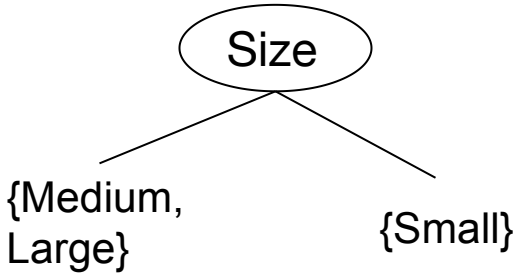
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

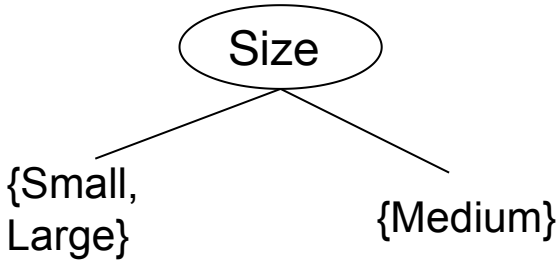


决策树构建：属性划分

- **影响属性划分的两个因素：**
  - 因素1：属性的类型
    - Nominal （标称型）
    - Ordinal （序数型）
    - Continuous （连续型）
  - 因素2：属性划分的分支数量
    - 两划分
    - 多划分

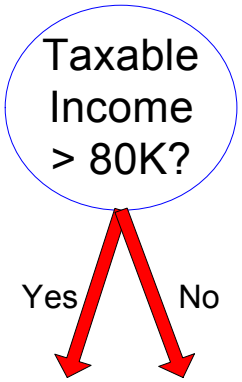




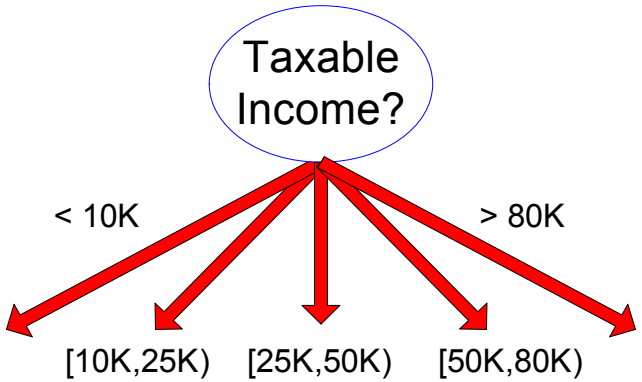


决策树构建：连续属性的划分

- **对于连续型属性而言：**
  - 既可以离散化，将其转换成离散可数的类别型属性
  - 引入单个分割条件，进行二元离散化：$(A < v)$ or $(A \geq v)$



(i) Binary split

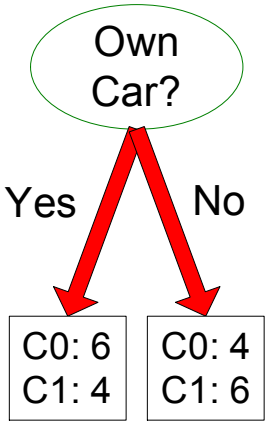


(ii) Multi-way split

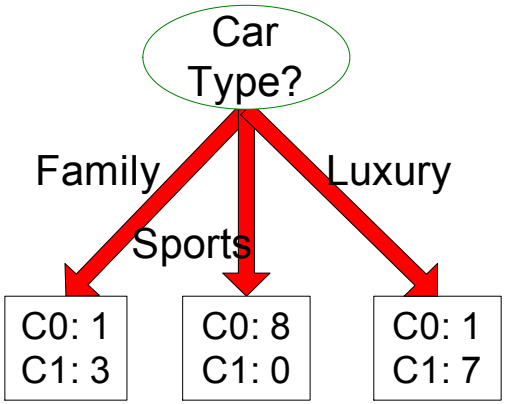
决策树的构建：最优划分的确定

- 给定如下情景，如何确定最优划分?

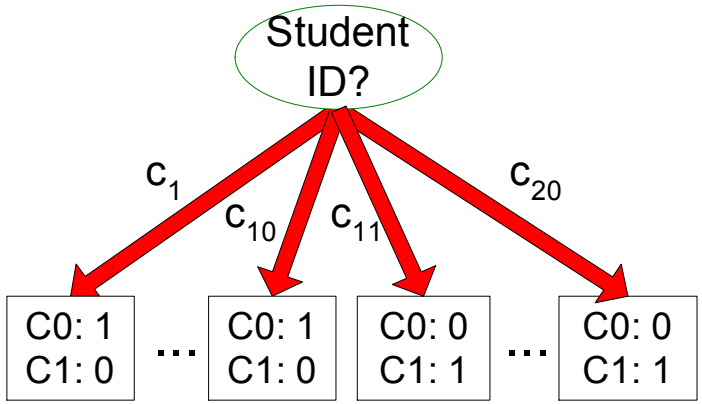
划分前: 10 records of class 0,
10 records of class 1



(1)



(2)



(3)

上述三种方案，哪个是最好的?

针对每个节点，我们需要量化数据在类标签取值上分布的情况

决策树的构建：最优划分的确定

- 决策树上常用的节点不纯性（Impurity）度量指标：
  - Gini Index （Gini 系数）
  - Entropy （信息熵）
  - Misclassification error

节点不纯性测度：Gini Index

- 给定节点t:

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

其中：P(j|t)是指节点t中类别是j的样本点所占比例

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

节点不纯性测度：Entropy

- 给定节点t:

$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

其中：P(j|t)是指节点t中类别是j的样本点所占比例

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

节点不纯性测度：Misclassification error

- 给定节点t:

$$Error(t) = 1 - \max_i P(i \mid t)$$

其中：P(j|t)是指节点t中类别是j的样本点所占比例

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

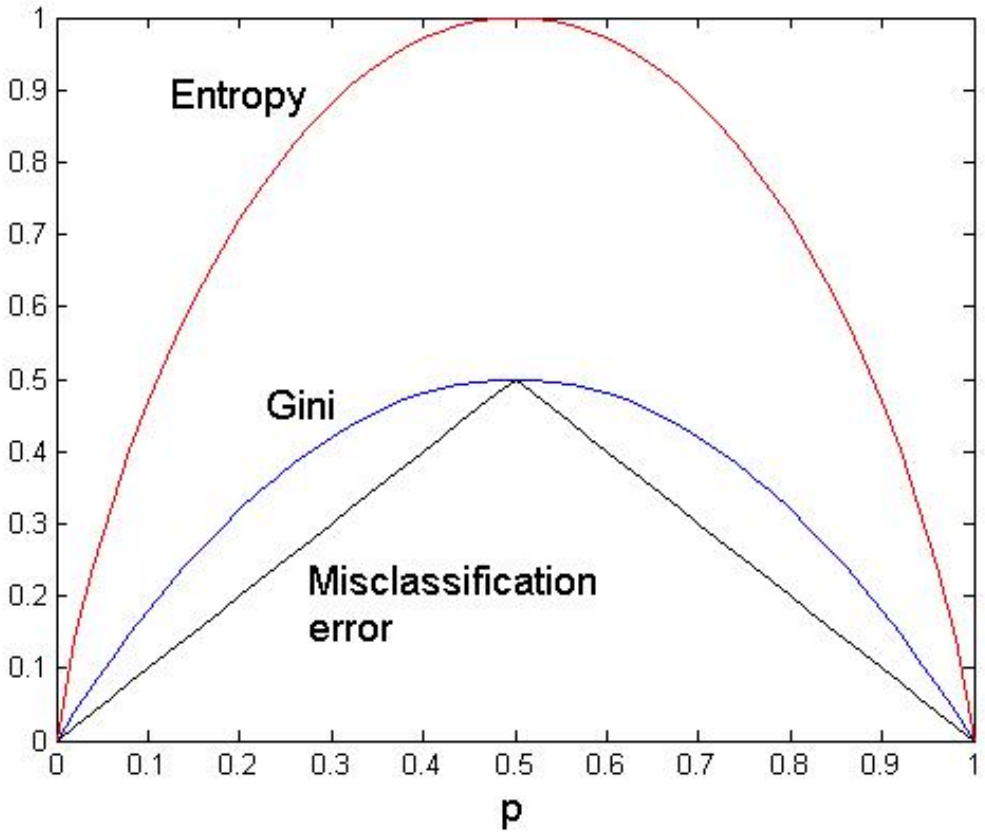
C1	**2**
C2	**4**

P(C1) = 2/6　　P(C2) = 4/6

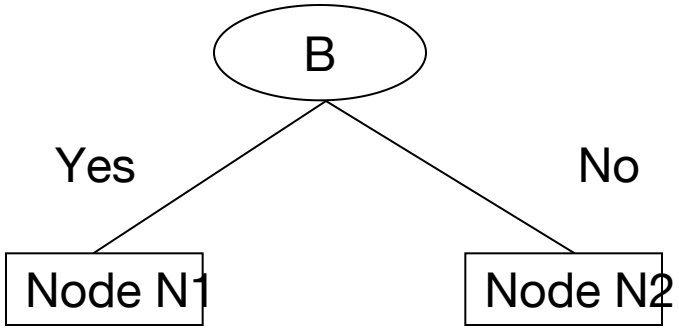
Error = 1 – max (2/6, 4/6) = 1 – 4/6 = 1/3

三个指标的比较

- 给定两分类问题，上述三个不纯性指标的比较：



基于Gini Index 的属性划分



	Parent
C1	**6**
C2	**6**
Gini = 0.500	

	N1	N2
C1	**5**	**1**
C2	**2**	**4**
Gini=0.371		

Gini(N1) = 1 – $(5/7)^2$ – $(2/7)^2$ = 0.408

Gini(N2) = 1 – $(1/5)^2$ – $(4/5)^2$ = 0.320

Gini(Children) = 7/12 * 0.408 + 5/12 * 0.320= 0.371

基于Gini Index 最优划分确定

- 针对给定的不同划分方案计算其Gini Index
- 比较不同方案的Gini Index，从而选择最优划分

Multi-way split

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

Two-way split
(find best partition of values)

	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	

	CarType	
	{ Luxury }	{Family, Sports }
C1	1	3
C2	1	5
Gini	0.475	

如何构建决策树？

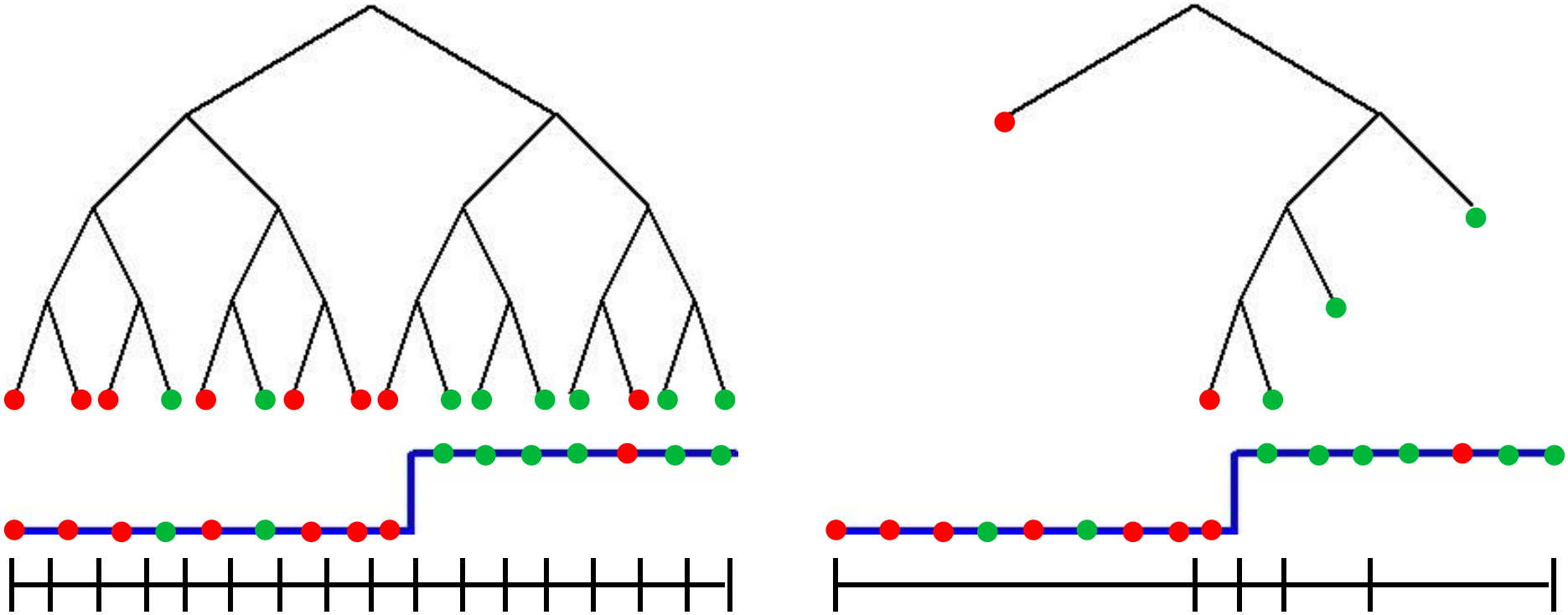
- 决策树的生成需要解决两个问题：
  - 如何对数据中的样本点进行划分？（如何分裂）：

答：根据划分所产生的节点不纯性度量指标来进行划分!

  - 什么时候停止样本的划分？(如何终止)

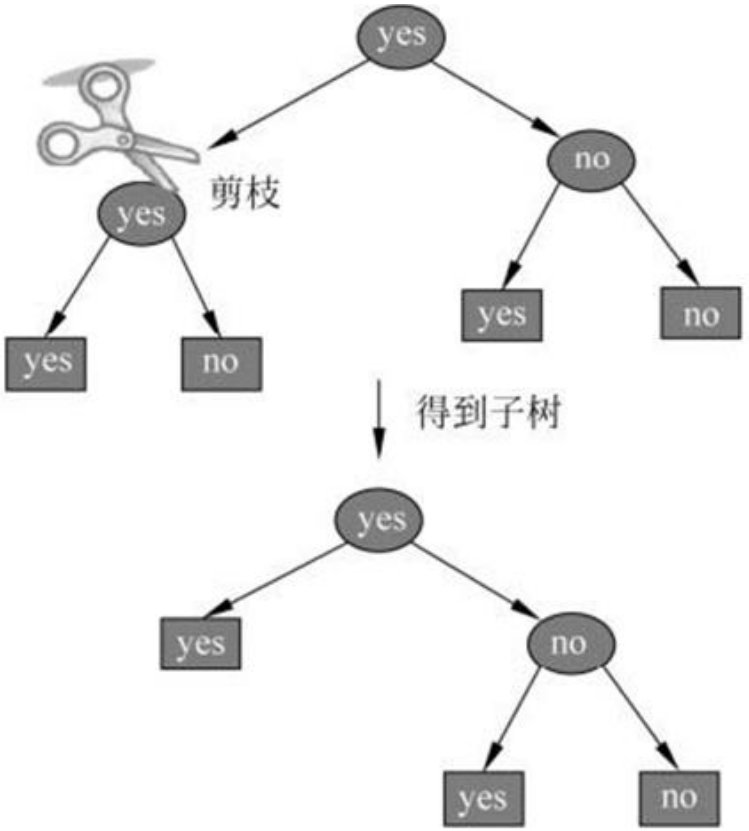
为什么要停止？

- 决策树过分生长和分裂会导致过分拟合：



如何停止分裂？

- 终止决策树过分增长的策略：
  - 先剪枝
    - 预先指定决策树的深度
    - 预先限制决策树的叶子节点数量
  - 后剪枝
    - 采用自下而上的策略
    - 根据模型拟合程度来确定是否剪枝



讲授提纲

01 分类分析基本概念

02 决策树分类模型

03 朴素贝叶斯模型

04 逻辑回归模型

05 分类模型评估

06 商务案例分析

贝叶斯规则

- 条件概率：给定B事件发生的前提下A事件发生的概率

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

- 贝叶斯规则：P(A|B) 与 P(B|A) 的互换：

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

贝叶斯学习

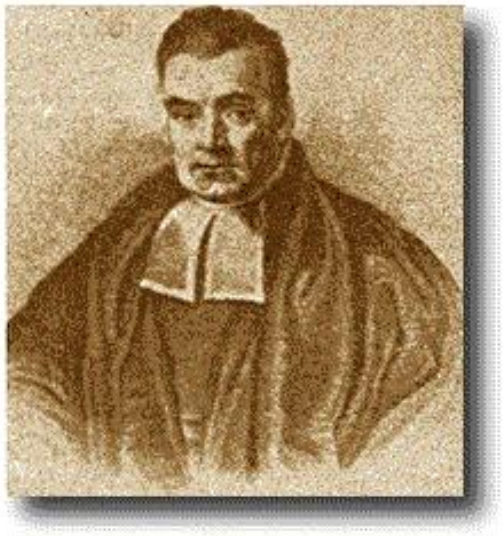
- 分类分析中使用贝叶斯规则：

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

- 等价于：

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

posterior　　likelihood prior



Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

贝叶斯分类模型

- 将数据集的所有变量($X_1, X_2, \ldots, X_n$) 以及类标签变量Y都看成是随机变量

- 给定数据样本点： ($X_1, X_2, \ldots, X_n$)
  - 分类的目的是要预测该样本点的类标签Y的取值
  - 实质上，我们需要找到一个Y的取值，使得 $P(Y| X_1, X_2, \ldots, X_n)$ 最大

- 根据贝叶斯定理和数据，对$P(Y| X_1, X_2, \ldots, X_n)$进行估计

$$P(Y \mid X_1 X_2 \cdots X_n) = \frac{P(X_1 X_2 \cdots X_n \mid Y) P(Y)}{P(X_1 X_2 \cdots X_n)}$$

朴素贝叶斯：条件独立性假设

- 朴素贝叶斯中的条件独立性假设：

$$P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y)$$

- 推广到一般情况下：

$$P(X_1 ... X_d|Y) = \prod_{i=1}^{d} P(X_i|Y)$$

朴素贝叶斯分类模型

- 从数据集中获取如下信息：
  - 类标签的先验分布信息：P(Y)
  - 给定类标签Y时条件独立的属性X
  - 对于每个给定的属性X_i，计算$P(X_i|Y)$

- 贝叶斯决策规则：

$$f_{NB}(\mathbf{x}) = \arg\max_y P(x_1, \ldots, x_d \mid y)P(y)$$

$$= \arg\max_y \prod_{i=1}^{d} P(x_i|y)P(y)$$

朴素贝叶斯分类模型：例子

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

A: attributes

M: mammals

N: non-mammals

$$P(A \mid M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A \mid M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

朴素贝叶斯分类总结

- 对数据中孤立的噪音点表现仍然稳健
- 在数据变量有一定缺失值的情况下也能够工作
- 对数据中不相干属性的引入也表现较为稳健
- 条件独立性假设被满足时，朴素贝叶斯能取得很好效果
- 条件独立性被破坏时，需要使用其他技术，如 Bayesian Belief Networks (BBN)等。

讲授提纲

01 分类分析基本概念

02 决策树分类模型

03 朴素贝叶斯模型

04 逻辑回归模型

05 分类模型评估

06 商务案例分析

逻辑回归模型

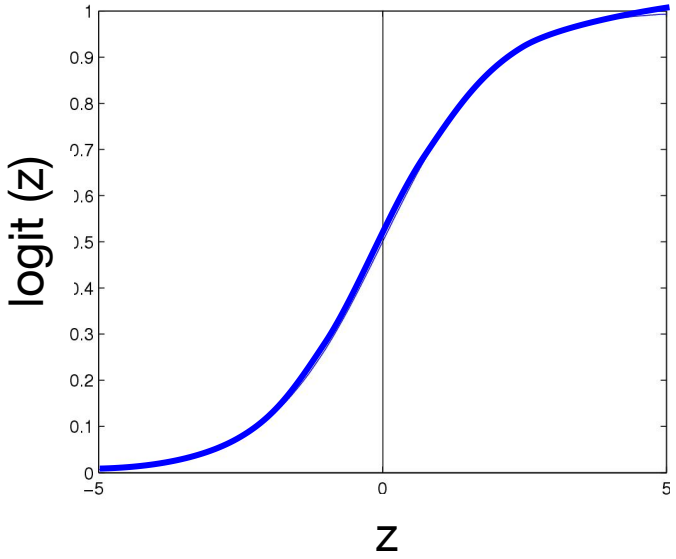
- 若分类变量取值是二元的(1/0)，我们可以假设P（Y|X）：

$$P(Y=1|X)=\frac{\exp(w_0+\sum_i w_i X_i)}{1+\exp(w_0+\sum_i w_i X_i)}$$

Logistic Function (or Sigmoid Function):

$$\sigma(z)=\frac{\exp(z)}{1+\exp(z)}$$

$$z=w_0+\sum_i w_i X_i$$



逻辑回归模型：分类准则

- 逻辑回归的分类准则推导：

$$P(Y=1|X)=\frac{\exp(w_0+\sum_i w_iX_i)}{1+\exp(w_0+\sum_i w_iX_i)}$$

$$\Rightarrow P(Y=0|X)=\frac{1}{1+\exp(w_0+\sum_i w_iX_i)}$$

$$\Rightarrow \frac{P(Y=1|X)}{P(Y=0|X)}=\exp(w_0+\sum_i w_iX_i) \underset{0}{\overset{1}{\gtrless}} 1$$

$$\Rightarrow \boxed{w_0+\sum_i w_iX_i \underset{0}{\overset{1}{\gtrless}} 0}$$

线性的分类模型！！

逻辑回归模型：多分类标签

- 当分类标签$Y \in \{y_1, \ldots, y_K\}$，即存在多分类标签时：

- 若 $k < K$

$$P(Y = y_k|X) = \frac{\exp(w_{k0} + \sum_{i=1}^{d} w_{ki}X_i)}{1 + \sum_{j=1}^{K-1} \exp(w_{j0} + \sum_{i=1}^{d} w_{ji}X_i)}$$

- 若 $k = K$

$$P(Y = y_K|X) = \frac{1}{1 + \sum_{j=1}^{K-1} \exp(w_{j0} + \sum_{i=1}^{d} w_{ji}X_i)}$$

逻辑回归模型：系数估计

- 给定数据集：

$$\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n} \Rightarrow X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$$

- 目标：估算参数 $w_0, w_1, \ldots w_d$

- 解决思路：极大似然估计

$$\hat{\mathbf{w}}_{MCLE} = \arg\max_{\mathbf{w}} \prod_{j=1}^{n} P(Y^{(j)} \mid X^{(j)}, \mathbf{w})$$

逻辑回归模型：系数估计

- 概率表达式：

$$P(Y=1|\mathbf{X},\mathbf{w})=\frac{\exp(w_0+\sum_i w_i X_i)}{1+\exp(w_0+\sum_i w_i X_i)}$$

$$P(Y=0|\mathbf{X},\mathbf{w})=\frac{1}{1+\exp(w_0+\sum_i w_i X_i)}$$

$$P(Y=y|\mathbf{X},\mathbf{w})=\frac{\exp\left(y(w_0+\sum_i w_i X_i)\right)}{1+\exp(w_0+\sum_i w_i X_i)}$$

$$\begin{aligned} l(\mathbf{w}) &\equiv \ln\prod_j P\left(y^{(j)}|\mathbf{x}^{(j)},\mathbf{w}\right) \\ &= \sum_j\left[y^{(j)}\left(w_0+\sum_i^d w_i x_i^{(j)}\right)-\ln\left(1+\exp\left(w_0+\sum_i^d w_i x_i^{(j)}\right)\right)\right] \end{aligned}$$

$$max\ l(w) \rightarrow \min -l(w)$$

梯度下降方法（Gradient Descent）求解该问题!

讲授提纲

01 分类分析基本概念

02 决策树分类模型

03 朴素贝叶斯模型

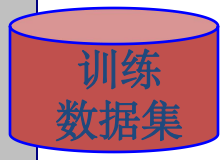
04 逻辑回归模型

05 分类模型评估

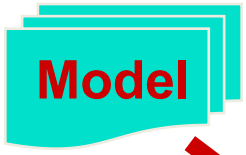
06 商务案例分析

分类模型

	categorical	categorical	continuous	class
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



Refund	**Marital Status**	**Taxable Income**	**Cheat**
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



分类应用

分类模型：损失函数

- 常用损失函数$\text{loss}(Y, f(X))$量分类模型预测值与真实类标签值之间的吻合程度

- 数据分析中常见的损失函数形式如下：

$$\text{loss}(Y, f(X)) = 1_{\{f(X) \neq Y\}}$$ **0/1 loss**

$$L(Y, f(X)) = |Y - f(X)|$$ **Absolute Value loss**

$$\text{loss}(Y, f(X)) = (Y - f(X))^2$$ **square loss**

$$L(Y, P(Y|X)) = -logP(Y|X)$$ **logarithmic loss function**

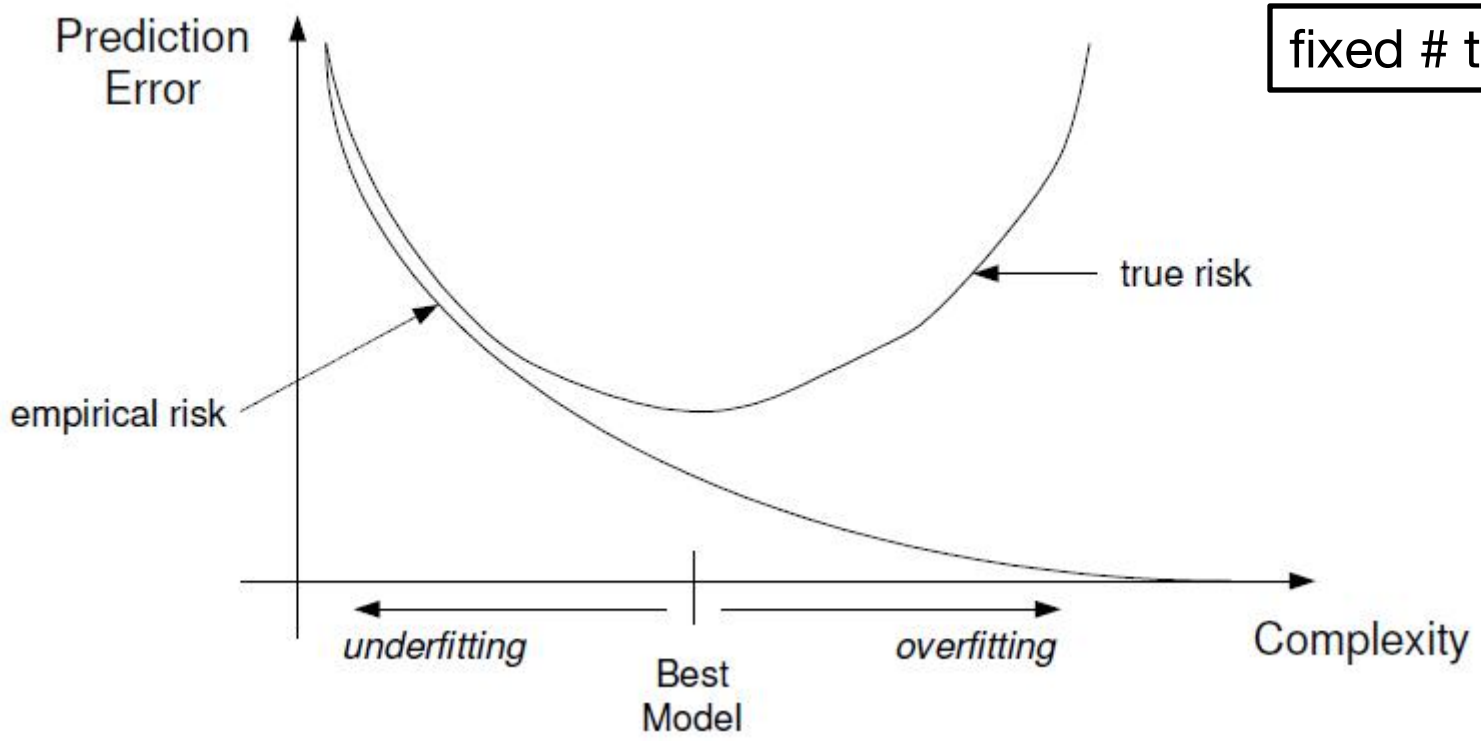
分类模型：真实风险和经验风险

- **真实风险/损失（True Risk）**：将分类模型用于任意一个未知样本点上进行类别判定时产生的期望损失。
  - 例如：分类分析中的错分概率 $P(f(X) \neq Y)$
  - 例如回归分析中的 $\mathbb{E}[(f(X) - Y)^2]$

- **经验风险/损失（Empirical Risk）**：分类模型在训练数据集上产生的期望损失。
  - 训练数据集上的错分率均值 $\frac{1}{n}\sum_{i=1}^{n} 1_{f(X_i)\neq Y_i}$
  - 训练数据集上的离差平方和均值 $\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2$

分类模型：拟合不足与过分拟合

- 模型的拟合不足与过分拟合



分类模型的评估

- 分类结果的混淆矩阵（confusion matrix）

	预测的类标签		
真实的类标签		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

$$\text{Precision (p)} = \frac{a}{a+c}$$

$$\text{Recall (r)} = \frac{a}{a+b}$$

$$\text{F - measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

小结

- 分类模型的基本概念
- 决策树分类模型
  - 如何确定最优的数据划分方案
  - 如何停止分裂
- 朴素贝叶斯分类模型
- 逻辑回归分类模型
- 分类模型的评估
  - 损失函数
  - 过分拟合
  - 分类模型的评估

讲授提纲

01 分类分析基本概念

02 决策树分类模型

03 朴素贝叶斯模型

04 逻辑回归模型

05 分类模型评估

06 商务案例分析