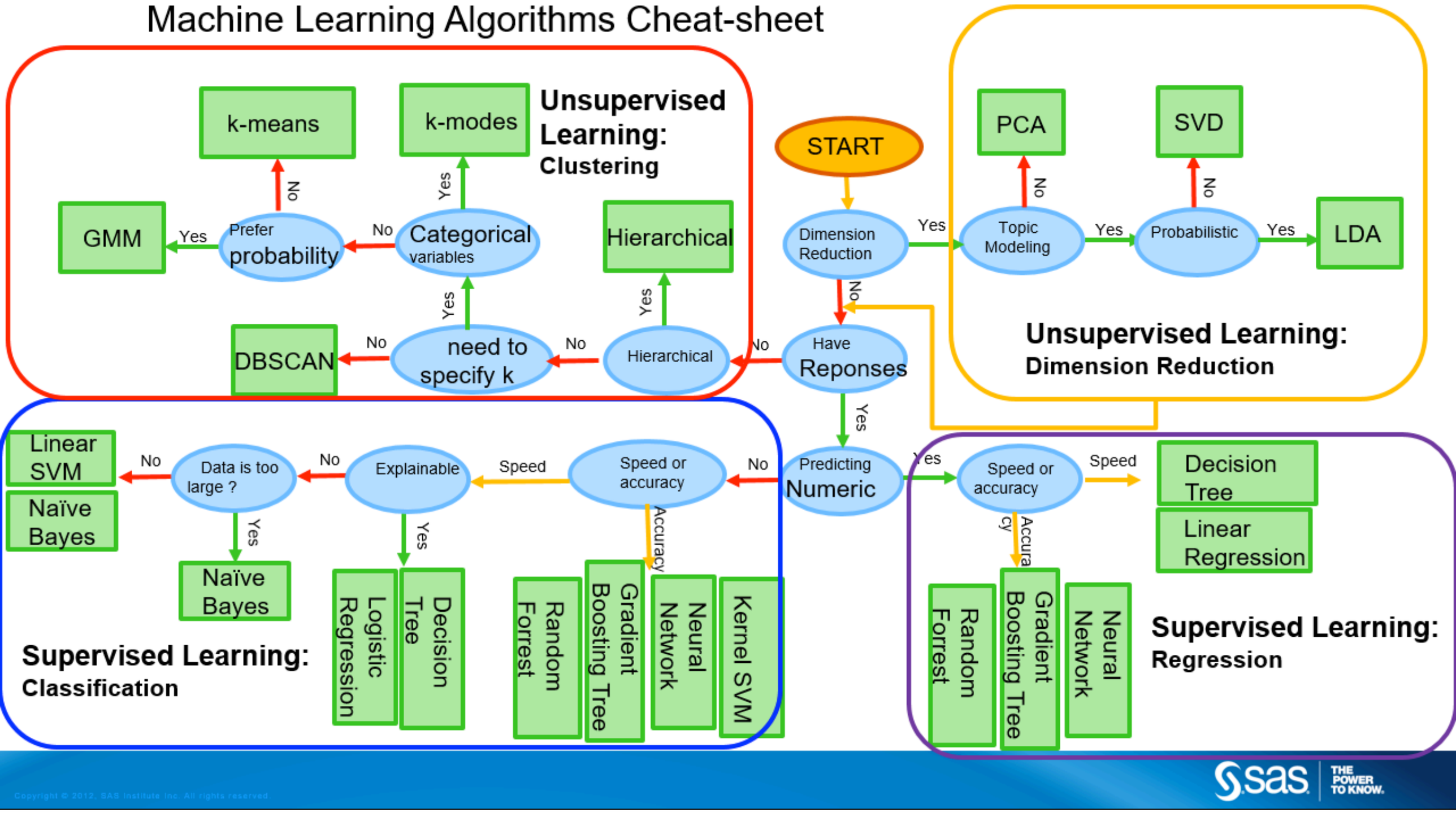
贝叶斯分类器

Group Study, ML: 7

周晓清 2020年7月24日



图片来源：https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/

监督学习

监督学习的任务是学习一个模型，应用这一模型，对给定的输入预测对应的输出。

模型的一般形式为决策函数

$$Y = f(X)$$

或者条件概率分布

$$P(Y \mid X)$$

按照模型求解方法，可分为两类：

- 生成模型
- 判别模型

监督学习

- 生成模型：由数据学习联合概率分布 $P(X, Y)$，然后求出条件概率分布

$$P(Y|X) = \frac{P(X, Y)}{P(X)}$$

  - 这样的方法之所以称为生成方法，是因为模型表示了给定输入 X 产生输出 Y 的生成关系 [李航.统计学习方法:18]
  - 朴素贝叶斯法实际上学习到生成数据的机制，所以属于生成模型 [李航.统计学习方法:48]
  - 例：朴素贝叶斯法、隐马尔可夫模型

- 判别模型：由数据直接学习决策函数 $f(X)$ 或者条件概率分布 $P(Y|X)$
  - 例：决策树、BP 神经网络、SVM

监督学习

在监督学习中，生成模型和判别模型各有优缺点，适用于不同条件下的学习问题。

- 生成方法的特点：
  - 可以还原出联合概率分布 $P(X, Y)$；
  - 学习收敛速度更快，即当样本容量增加时，学到的模型可以更快收敛于真实模型；
  - 当存在隐变量时，仍可以用生成方法学习。
- 判别方法的特点：
  - 直接学习条件概率 $P(Y|X)$ 或决策函数 $f(X)$，往往学习准确率更高；
  - 由于直接学习 $P(Y|X)$ 或 $f(X)$，可以对数据进行各种程度上的抽象、定义特征并使用特征，因此可以简化学习问题。

生成模型和判别模型的另一种定义

- 生成模型：对条件概率 $P(X \mid Y)$ 建模，根据标签值 Y 生成随机的样本数据 X。
  - 例：给定一系列狗的图片，要求生成一张新的，不在已有数据集里的狗的图片。
- 判别模型：对条件概率 $P(Y \mid X)$ 建模，根据样本特征向量 X 的值判断它的标签值 Y。
  - 例：给定一张图片，判断这张图片里的动物是猫还是狗。

https://en.wikipedia.org/wiki/Generative_model#Definition

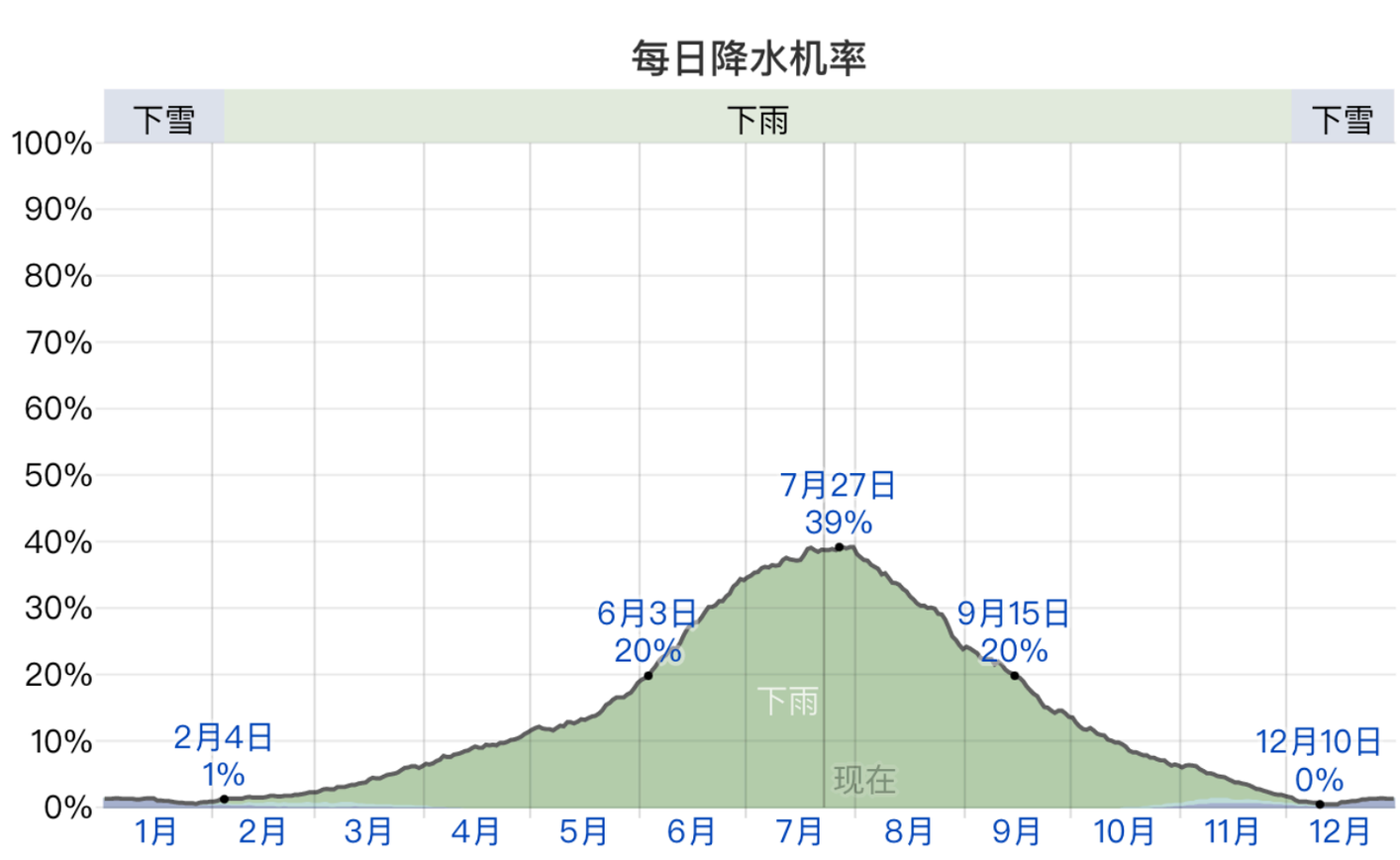
1 朴素贝叶斯分类器

2 半朴素贝叶斯分类器

3 贝叶斯网

先验概率（Prior Probability）

- 在贝叶斯统计推断中，不确定量的先验概率分布（通常简称为先验）是指在考虑某些证据（evidence）之前，相信该量的概率分布
- 例：7月某天降雨或不降雨的概率
- 先验概率记作 $P(y = c_i)$
  - $\sum_{i=1}^{N} P(y = c_i) = 1$



证据（Evidence）

- 已知的变量观测值称为证据

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	NO
sunny	hot	high	TRUE	NO
overcast	hot	high	FALSE	YES
rainy	mild	high	FALSE	YES
rainy	cool	normal	FALSE	YES
rainy	cool	normal	TRUE	NO
overcast	cool	normal	TRUE	YES
sunny	mild	high	FALSE	NO
sunny	cool	normal	FALSE	YES
rainy	mild	normal	FALSE	YES
sunny	mild	normal	TRUE	YES
overcast	mild	high	TRUE	YES
overcast	hot	normal	FALSE	YES
rainy	mild	high	TRUE	NO

后验概率（Posterior Probability）

- 给定观测向量 $\mathbf{x}$，估计出某个特定类别的概率 $P(c_i \mid \mathbf{x})$ 称为后验概率
- 由贝叶斯定理，获得后验概率

$$P(c \mid \mathbf{x}) = \frac{P(\mathbf{x}, c)}{P(\mathbf{x})} = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})}$$

  - $P(c)$ 类先验概率
  - $P(\mathbf{x} \mid c)$ 样本 x 相对于类标记 c 的类条件概率，称为似然
  - $P(\mathbf{x})$ 用于归一化的证据因子，保证各类别的后验概率总和为1从而满足概率条件

- 估计 $P(c \mid \mathbf{x})$ 的问题转化为如何基于训练数据 D 来估计先验 $P(c)$ 和似然 $P(\mathbf{x} \mid c)$

朴素贝叶斯法的学习

- 类先验概率 $P(c)$
  - 表达了样本空间中各类样本所占的比例
  - 根据大数定律，当训练集 D 包含充足的独立同分布样本时，概率可用频率估计

$$P(c) = \frac{|D_c|}{|D|}$$

  - 其中 D_c 表示训练集 D 中第 c 类样本组成的集合.

	Play
Y	9/14
N	5/14

朴素贝叶斯法的学习

- 似然 $P(\mathbf{x}\,|\,c) = P(x_1, x_2, \ldots, x_n\,|\,c)$
  - 在数理统计学中，似然函数是一种关于统计模型中的参数的函数，表示模型参数中的似然性
  - 概率，用于在已知一些参数的情况下，预测接下来在观测上所得到的结果；似然性，则是用于在已知某些观测所得到的结果时，对有关事物之性质的参数进行估值
  - 在这种意义上，似然函数可以理解为条件概率 $P(c\,|\,\mathbf{x})$ 的逆反
  - 由于涉及关于 $\mathbf{x}$ 所有属性的联合概率，直接根据样本出现的频率来估计会遇到严重困难
    - 假设 x_i 可取值有 S_i 个，y 可取值有 K 个，那么参数个数为 $K\prod_{i=1}^{n} S_i$
    - 参数个数可能大于训练样本数，无法直接用频率估计
- 基于有限样本直接估计联合概率，在计算上将遭遇组合爆炸问题，在数据上将遭遇样本稀疏问题；属性数越多，问题越严重

朴素贝叶斯法的学习

- 条件独立性假设：似然 $P(\mathbf{x}|c) = P(x_1, x_2, \ldots, x_n|c) = \prod_{j=1}^{n} P(x_j|c)$
  - 假设每一个属性独立地对分类结果发生影响，估计联合似然变成了估计每一个维度上的似然
  - 做了极大的简化，朴素贝叶斯法由此得名
  - 这一假设使朴素贝叶斯法变得简单，但有时会牺牲一定的分类准确率
- 对离散属性，$P(x_i|c) = \frac{|D_{c,x_i}|}{|D_c|}$，其中 D_{c,x_i} 表示 D_c 中在第 i 个属性上取值为 x_i 的样本组成的集合

Outlook			**Temperature**			**Humidity**			**Windy**		
	Y	N		Y	N		Y	N		Y	N
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	F	6/9	2/5
overcast	4/9	0/5	mild	4/9	2/5	norm	6/9	1/5	T	3/9	3/5
rainy	3/9	2/5	cool	3/9	1/5						

朴素贝叶斯法的学习

- 条件独立性假设：似然 $P(\mathbf{x}\,|\,c) = P(x_1, x_2, \ldots, x_n\,|\,c) = \prod_{j=1}^{n} P(x_j\,|\,c)$
  - 对连续属性可考虑概率密度函数，假定 $p(x_i\,|\,c) \sim \mathcal{N}(\mu_{c,i}, \sigma_{c,i}^2)$，其中 $\mu_{c,i}$ 和 $\sigma_{c,i}^2$ 分别是第 c 类样本在第 i 个属性上取值的均值和方差，则有

$$p(x_i\,|\,c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}\right)$$

朴素贝叶斯法的分类

- 后验概率最大化：朴素贝叶斯法分类时，对给定的输入 $\mathbf{x}$，通过学习到的模型计算后验概率分布 $P(c \mid \mathbf{x})$，将后验概率最大的类作为预测结果

$$f(\mathbf{x}) = \arg\max_i P(c_i \mid \mathbf{x})$$

$$P(Y)P(Outlook = S, Temp = C, Humidity = H, Windy = T \mid Y)$$
$$= P(Y)P(Outlook = S \mid Y)P(Temp = C \mid Y)P(Humidity = H \mid Y)P(Windy = T \mid Y)$$
$$= \frac{9}{14}\frac{2}{9}\frac{3}{9}\frac{3}{9}\frac{3}{9} = 0.005$$

$$P(N)P(Outlook = S, Temp = C, Humidity = H, Windy = T \mid N)$$
$$= P(N)P(Outlook = S \mid N)P(Temp = C \mid N)P(Humidity = H \mid N)P(Windy = T \mid N)$$
$$= \frac{5}{14}\frac{3}{5}\frac{1}{5}\frac{4}{5}\frac{3}{5} = 0.021$$

Outlook			**Temperature**			**Humidity**			**Windy**		
	Y	N		Y	N		Y	N		Y	N
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	F	6/9	2/5
overcast	4/9	0/5	mild	4/9	2/5	norm	6/9	1/5	T	3/9	3/5
rainy	3/9	2/5	cool	3/9	1/5						

	Play
Y	9/14
N	5/14

后验概率最大化的含义

- 朴素贝叶斯法将实例分到后验概率最大的类中，这等价于期望风险最小化
- 假设选择 0-1 损失函数：

$$L(Y, f(X)) = \begin{cases} 1, Y \neq f(X) \\ 0, Y = f(X) \end{cases},$$

式中 $f(X)$ 是分类决策函数

- 这时，期望风险函数为

$$R_{exp}(f) = E_X \sum_{k=1}^{K} [L(c_k, f(X))] P(c_k | X),$$

期望是对联合分布 $P(X, Y)$ 取的

后验概率最大化的含义

- 为了使期望风险最小化，只需对 $X = \mathbf{x}$ 逐个极小化，由此得到：

$$
\begin{aligned}
f(x) &= \arg\min_{y \in \mathcal{Y}} L(c_k, y) P(c_k | X = x) \\
&= \arg\min_{y \in \mathcal{Y}} P(y \neq c_k | X = x) \\
&= \arg\min_{y \in \mathcal{Y}} (1 - P(y = c_k | X = x)) \\
&= \arg\max_{y \in \mathcal{Y}} P(y = c_k | X = x)
\end{aligned}
$$

- 于是根据期望风险最小化准则，得到了后验概率最大化准则

$$
f(x) = \arg\max_{c_k} P(c_k | X = x)
$$

平滑（Smoothing）

- 问题：其他属性携带的信息，可能被训练集中未出现的属性值抹去

$$P(N)P(Outlook=O,Temp=C,Humidity=H,Windy=T|N)$$
$$=P(N)P(Outlook=S|N)P(Temp=C|N)P(Humidity=H|N)P(Windy=T|N)$$
$$=\frac{5}{14}\frac{0}{5}\frac{1}{5}\frac{4}{5}\frac{3}{5}=0$$

当 Outlook=overcast 时，无论其它属性取什么值，最终算得的后验概率都是 0，使分类产生偏差

- 拉普拉斯平滑：在随机变量各个取值的频数上加 1

$$P(Y=c_k)=\frac{\sum_{i=1}^{N}I(y_i=c_k)+1}{N+K}$$

$$P(x_j=a_{jl}|c_i)=\frac{\sum_{i=1}^{N}I(x_j=a_{jl},y_i=c_k)+1}{\sum_{i=1}^{N}I(y_i=c_k)+S_j}$$

Outlook	
	N
sunny	(3+1)/(5+3)
overcast	**(0+1)/(5+3)**
rainy	(2+1)/(5+3)

1 朴素贝叶斯分类器

2 半朴素贝叶斯分类器

3 贝叶斯网

半朴素贝叶斯分类器

- 为降低估计后验概率 $P(c \mid \mathbf{x})$ 的困难，朴素贝叶斯分类器采用了属性条件独立性假设，但在现实任务中这个假设很难成立

- 半朴素贝叶斯分类器 适当考虑一部分属性间的相互依赖信息，从而既不需进行完全联合概率计算，又不至于彻底忽略了比较强的属性依赖关系

- 独依赖估计（One-Dependent Estimator, ODE）假设每个属性在类别之外最多依赖于一个其它属性，即

  $$P(c \mid \mathbf{x}) \propto P(c) \prod_{i=1}^{d} P(x_i \mid c, pa_i)$$

  其中 pa_i 为属性 x_i 所依赖的属性，称为 x_i 的父属性

- 问题的关键转化为如何确定每个属性的父属性

SPODE（Super-Parent ODE）

- 最直接的做法是假设所有属性都依赖于同一属性，称为超父（super-parent），然后通过交叉验证等模型选择方法来确定父属性，由此形成了 SPODE（Super-Parent ODE）方法

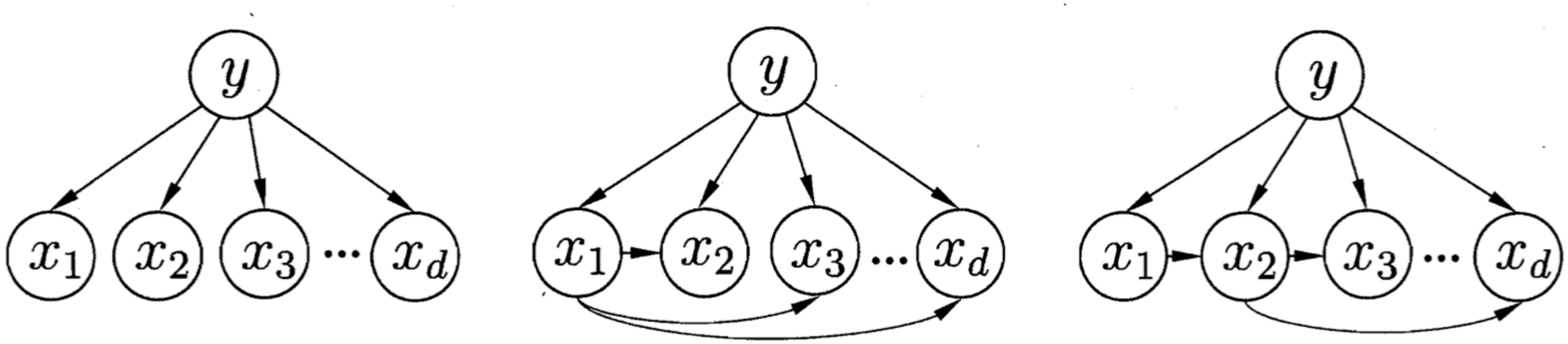


图 7.1 朴素贝叶斯与两种半朴素贝叶斯分类器所考虑的属性依赖关系

TAN（Tree Augmented Naïve Bayes）

- TAN (Tree augmented Naïve Bayes) 则在最大带权生成树算法的基础上，通过以下步骤将属性间依赖关系简约为树形结构

  - 计算任意两个属性之间的条件互信息

$$I(x_i, x_j | y) = \sum_{x_i, x_j; c \in \mathcal{Y}} P(x_i, x_j | c) \log \frac{P(x_i, x_j | c)}{P(x_i | c) P(x_j | c)};$$

  - 以属性为结点构建完全图，任意两个结点之间边的权重设为 $I(x_i, x_j | y)$；
  - 构建此完全图的最大带权生成树，挑选根变量，将边设为有向；
  - 加入类别节点 y，增加从 y 到每个属性的有向边。

- 条件互信息 $I(x_i, x_j | y)$ 刻画了属性 x_i 和 x_j 在已知类别情况下的相关性，因此，通过最大生成树算法，TAN 实际上保留了强相关属性之间的依赖性

AODE（Averaged ODE）

- AODE（Averaged One-Dependent Estimator）是一种基于集成学习机制、更为强大的独依赖分类器。

  - AODE 尝试将每个属性作为超父来构建 SPODE，然后将那些具有足够训练数据支撑的 SPODE 集成起来作为最终结果，即

$$P(c \mid \mathbf{x}) \propto \sum_{i=1, |D_{x_i}| \geq m'}^{d} P(c, x_i) \prod_{j=1} P(x_j \mid c, x_i)$$

  其中 D_{x_i} 是个第 i 个属性上取值为 x_i 的样本的集合，m’ 为阈值常数

- 类似朴素贝叶斯，AODE 的训练过程也是“计数”，即在训练集上对符合条件的样本进行计数

1 朴素贝叶斯分类器

2 半朴素贝叶斯分类器

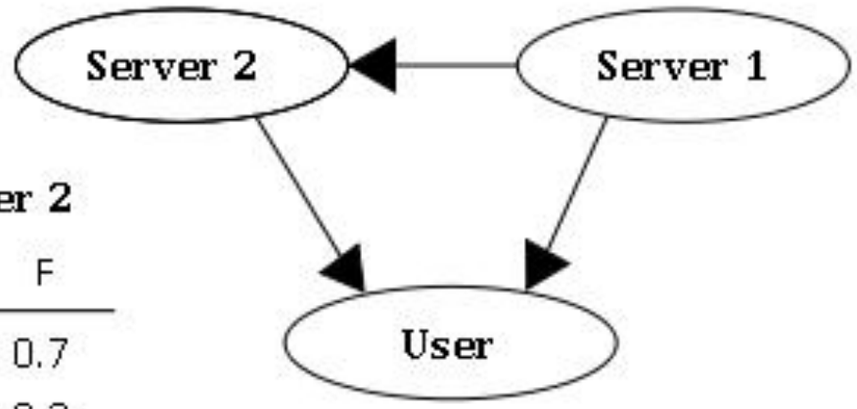
3 贝叶斯网

贝叶斯网（Bayesian Network）

- 贝叶斯网络（Bayesian network）是帮助人们把概率统计应用于复杂领域、进行不确定性推理的工具
  - 联合概率太复杂（随变量个数指数增长）
  - 贝叶斯网把概率分解成一系列简单模块，从而降低复杂度
- 借由有向无环图来刻画属性之间的依赖关系，其中结点代表属性，每个结点都附有一个概率分布，根结点 X 所附的是它的边缘分布 P(X)，而非根结点所附的是条件概率分布 $P(X \mid \pi(X))$，其中 $\pi(X)$ 指 X 的父结点集
- $B = < G, \Theta >$
  - 网络结构 G 是有向无环图，每个结点对应一个属性，若两个属性有直接依赖关系，则用一条有向边连接起来
  - 参数 Θ 定量描述属性间的依赖关系，包含了每个属性的条件概率表

贝叶斯网（Bayesian Network）

- 例：假设有两个服务器(S_1, S_2)，会发送数据包到用户端（以U表示），但是第二个服务器的数据包发送成功率会与第一个服务器发送成功与否有关，因此该贝叶斯网的结构图可以表示成如下图的形式。就每个数据包发送而言，只有两种可能值：T（成功）或 F（失败）。则此贝叶斯网络之联合概率分布可以表示成：
  $P(U, S_1, S_2) = P(U \mid S_1, S_2)P(S_2 \mid S_1)P(S_1)$



Server 1
T	F
0.4	0.6

Server 1	Server 2	
	T	F
F	0.3	0.7
T	0.7	0.3

Server 2	Server 1	User	
		T	F
F	F	0	1
F	T	1	0
T	F	1	0
T	T	1	0

利用条件独立降低复杂度

$$P(X_1, \ldots, X_n) = P(X_1)P(X_2 \mid X_1)\ldots P(X_n \mid X_1, X_2, \ldots, X_{n-1})$$
$$= \prod_{i=1}^{n} P(X_i \mid X_1, X_2, \ldots, X_{i-1})$$

对任意 X_i，若存在 $\pi(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$，使得给定 $\pi(X_i)$，X_i 与 $\{X_1, \ldots, X_{i-1}\}$ 中的其他变量条件独立，即

$$P(X_i \mid X_1, \ldots, X_{i-1}) = P(X_i \mid \pi(X_i))$$

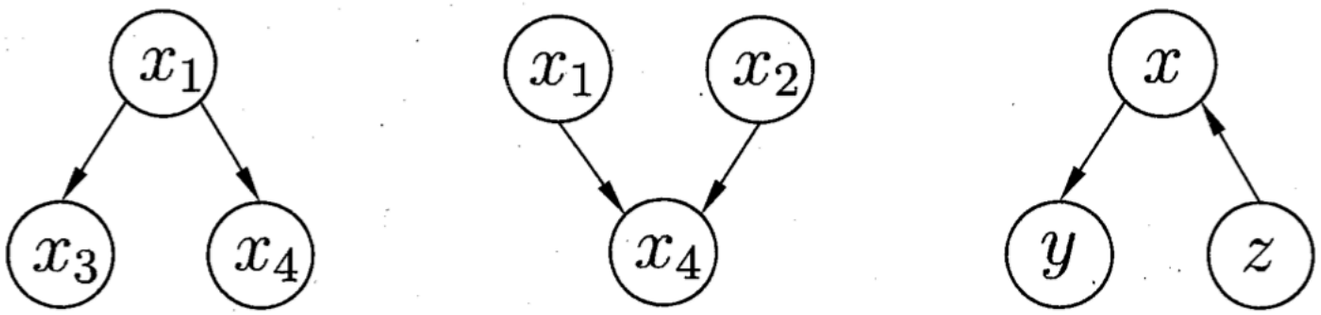
则

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \pi(X_i))$$

假设对任意 X_i，$\pi(X_i)$ 最多包含 m 个变量，则右端的独立参数最多为 $n2^m$ 个

贝叶斯网的结构

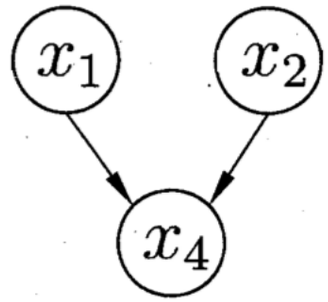
- 贝叶斯网结构有效地表达了属性间的条件独立性。给定父节点集，贝叶斯网假设每个属性与它的非后裔属性独立
- 贝叶斯网中三个变量之间的典型依赖关系



同父结构　　V型结构　　顺序结构

- 同父结构中，给定父结点 x_1 的取值，则 x_3 与 x_4 条件独立
- 顺序结构中，给定 x 的取值，则 y 和 z 条件独立
- V 型结构中，给定子结点 x_4 的取值，则 x_1 与 x_2 必不独立；若 x_4 的取值完全未知，则 x_1 与 x_2 则是相互独立的　← 边际独立性

贝叶斯网的结构



V型结构

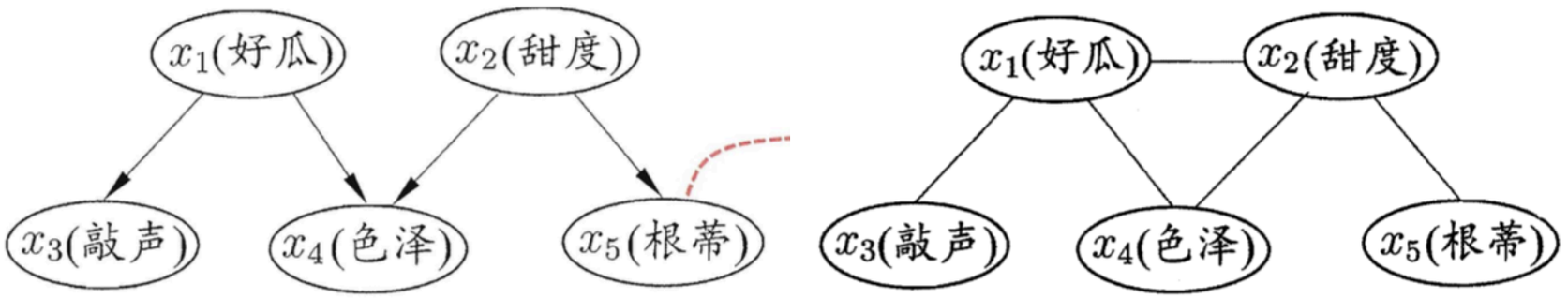
- 边际独立性（marginal independence）
  - 证明：若 x_4 的取值完全未知，则 V 型结构下 x_1 与 x_2 是相互独立的

$$P(x_1, x_2) = \sum_{x_4} P(x_1, x_2, x_4)$$
$$= \sum_{x_4} P(x_4 \mid x_1, x_2) P(x_1) P(x_2)$$
$$= P(x_1) P(x_2)$$

- 一个变量取值的确定与否，能对另两个变量间的独立性发生影响

贝叶斯网的结构

- 道德图（moral graph）：为了分析有向图中变量间的条件独立性，可使用有向分离（D-separation），将一个有向图转化为无向图
  - 将所有 V 型结构的两个父节点相连
  - 将所有有向边改为无向边



- 基于道德图能直观、迅速找到变量间的条件独立性. 假设道德图中有变量 x, y 和变量集合 $\mathbf{z} = \{z_i\}$，若 x 和 y 在图上能被 $\mathbf{z}$ 分开，即从道德图中将 $\mathbf{z}$ 去除后，x 和 y 属于两个连通分支，则称变量 x 和 y 被 $\mathbf{z}$ 有向分离，$x \perp y | \mathbf{z}$ 成立

贝叶斯网的结构

- 确定网络结构
  1. 选定一组刻画问题的随机变量 $\{X_1, X_2, \ldots, X_n\}$；
  2. 选择一个变量顺序 $\alpha = \langle X_1, X_2, \ldots, X_n \rangle$；
  3. 从一个空图出发，按照顺序 α 逐个将变量加入 $\mathcal{G}$ 中；
  4. 在加入变量 X_i 时，$\mathcal{G}$ 中的变量包括 $X_1, X_2, \ldots, X_{i-1}$：
     1. 利用问题的背景知识，在这些变量中选择一个尽可能小的子集 $\pi(X_i)$，是的假设“给定 $\pi(X_i)$，X_i 与 $\mathcal{G}$ 中的其他变量条件独立”合理；
     2. 从 $\pi(X_i)$ 中的每一个节点添加一条指向 X_i 的有向边。

  - 不同的变量顺序导致不同的网络结构，不同的网络结构表示了联合分布的不同分解，而不同的分解则意味着不同的复杂度
  - 建议用因果关系来决定变量顺序，原因在前，结果在后

- 确定网络参数：通过数据分析或从问题的特性直接得到

贝叶斯网的学习

- 贝叶斯网学习的首要任务是根据训练数据集来找出结构最“恰当”的贝叶斯网
- “最小描述长度”（Minimal Description Length, MDL）准则
  - 贝叶斯网描述了一个在训练数据上的概率分布，自有一套编码机制能使那些经常出现的样本有更短的编码。于是，应选择综合编码长度（包括描述网络和编码数据）最短的贝叶斯网
- 给定训练集 $D = \{x_1, x_2, \ldots, x_m\}$，贝叶斯网 $B = < G, \Theta >$ 在 D 上的评价函数可以写为

$$S(B \mid D) = f(\theta) \mid B \mid - LL(B \mid D)$$

- 其中，|B| 是贝叶斯网的参数个数；$f(\theta)$ 表示描述每个参数 θ 所需的字节数，而

$$LL(B \mid D) = \sum_{i=1}^{m} logP_B(x_i)$$

- 是贝叶斯网的对数似然

贝叶斯网的推断

- 通过已知变量观测值来推测待推测查询变量的过程称为“推断”（inference），已知变量观测值称为证据（evidence）
- 最理想的是根据贝叶斯网络定义的联合概率分布来精确计算后验概率。当网络结点较多、连接稠密时，难以进行精确推断，已被证明是 NP 难的
- 在现实应用中，贝叶斯网的近似推断常使用吉布斯采样（Gibbs sampling）来完成
  - 吉布斯采样随机产生一个与证据 $E = e$ 一致的样本 q^0 作为初始点，然后每步从当前样本出发产生下一个样本。每个样本 q^i 的产生都依赖前一个样本 q^{i-1}，且 q^i 和 q^{i-1} 最多只有一个非证据变量的取值不同
  - 假定经过 T 次采样的得到与 q 一致的样本共有 n_q 个，则可近似估算出后验概率

$$P(Q = q \mid E = e) \simeq \frac{n_q}{T}$$

贝叶斯网的推断

- 吉布斯采样算法

输入： 贝叶斯网 $B = \langle G, \Theta \rangle$;
采样次数 T;
证据变量 $\mathbf{E}$ 及其取值 $\mathbf{e}$;
待查询变量 $\mathbf{Q}$ 及其取值 $\mathbf{q}$.

过程：

$n_q = 0$
$\mathbf{q}^0 =$ 对 $\mathbf{Q}$ 随机赋初值
for $t = 1, 2, \ldots, T$ **do**
  for $Q_i \in \mathbf{Q}$ **do**
    $\mathbf{Z} = \mathbf{E} \cup \mathbf{Q} \setminus \{Q_i\}$;
    $\mathbf{z} = \mathbf{e} \cup \mathbf{q}^{t-1} \setminus \{q_i^{t-1}\}$;
    根据 B 计算分布 $P_B(Q_i \mid \mathbf{Z} = \mathbf{z})$;
    $q_i^t =$ 根据 $P_B(Q_i \mid \mathbf{Z} = \mathbf{z})$ 采样所获 Q_i 取值;
    $\mathbf{q}^t =$ 将 $\mathbf{q}^{t-1}$ 中的 q_i^{t-1} 用 q_i^t 替换
  end for
  if $\mathbf{q}^t = \mathbf{q}$ **then**
    $n_q = n_q + 1$
  end if
end for

输出： $P(\mathbf{Q} = \mathbf{q} \mid \mathbf{E} = \mathbf{e}) \simeq \frac{n_q}{T}$
