第8章 分类算法

《人工智能算法》

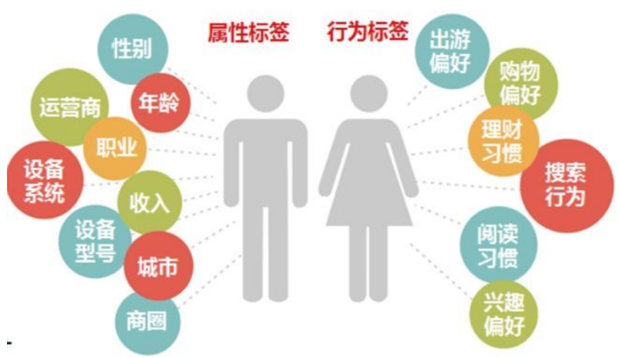
清华大学出版社
2022年7月

提纲

- 分类算法概述
- 决策树
- 支持向量机
- 贝叶斯分类
- 总结

分类算法概述 (1)

- **电商平台面临的实际问题**
  - 如何快速精准地实现用户分群？（预测流失或VIP客户）
  - 如何预测新产品的销量及喜爱该产品的客户？
  - 如何对客户的某些特征进行分类（圈选具有共同特征的用户，提供个性化的购物体验）



分类算法概述 (2)

- **数据分类（Classification）**

  目的：根据新数据样本的属性为其分配一个正确的类别

  应用：图片识别、信誉证实、医疗诊断、异常检测、情感分析…

- **经典的单一分类算法**
  - 决策树（Decision Tree）
  - k-近邻（k-Nearest Neighbor）
  - 支持向量机（Support Vector Machine, SVM）
  - 贝叶斯（Bayesian）分类
  - 人工神经网络（Neural Network）
  - 关联分类（Association Classification）

监督学习（Supervised Learning）

提纲

- 分类算法概述
- 决策树
- 支持向量机
- 贝叶斯分类
- 总结

决策树 (1)

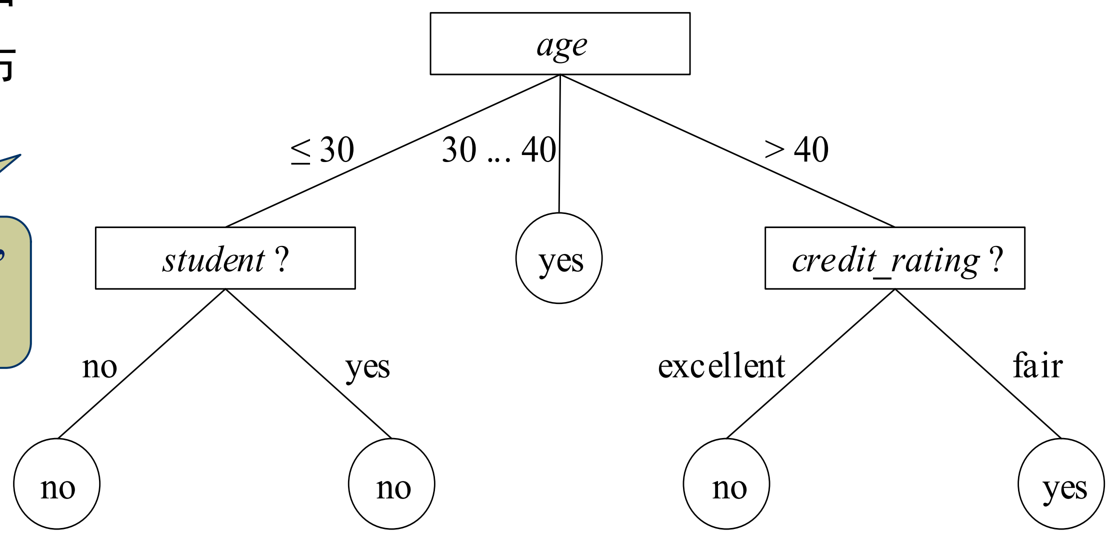
- **基本概念**
  - 从实例中构造表示分类规则的决策树（描述属性与类别的关系）
  - 类似于流程图的树结构

    内部节点：一个属性变量上的测试

    分枝：一个测试输出

    叶子节点：类或分布

描述“buys_computer”的决策树



决策树 (2)

- **决策树构造算法**

基本思想：基于贪心法递归地分裂输入变量空间的各个单元

关键步骤（以ID3为例）：

（1）选择测试变量

① 如果样本都在同一个类，则该节点成为叶子节点

② 否则，选择信息增益最高的变量作为该节点的测试变量

（2）递归分裂

（3）递归分裂步骤停止，仅当下列条件之一成立

① 给定节点的所有样本属于同一类

② 没有剩余属性变量可用来进一步分裂样本

③ 一个分枝没有样本

决策树 (3)

- **决策树构造算法**

  选择测试变量：

  设S是s个带类标记的数据样本的集合，F是n个属性变量的集合，有m个类$\{c_1,\ldots,c_m\}$，s_i是类c_i中的样本数，对每个样本分类的期望信息为：

$$I(s_1,\cdots,s_m)=-\sum_{i=1}^{m}(p_i\cdot\log p_i)$$

  设属性变量$A=\{a_1,\cdots,a_v\}$（$A\in F$），可用A将S划分为v个子集$\{S_1,\cdots,S_v\}$的熵或期望信息为：

$$E(A)=\sum_{j=1}^{v}\frac{s_{1j}+\cdots+s_{mj}}{s}I\left(s_{1j}+\cdots+s_{mj}\right)$$

  以A作为测试变量（即在A上分裂）所获得的信息增益为：

$$Gain(A)=I\left(s_1,\cdots,s_m\right)-E(A)$$

决策树 (4)

- **决策树构造算法**

输入： S，带有类标记的训练数据集；
F：属性变量集；ε：信息增益阈值
输出： T，决策树（类标记）

步骤：

1. **If** S所有样本属于同个类c_k **Then**
2. T为单节点树，c_k为该节点的类标记
3. **Return** T
4. **End if**
5. **If** $F=\varnothing$ **Then**
6. T为单节点树，将S中实例最多的类c_k作为该节点的类标记
7. **Return** T
8. **End If**
9. $A_g \leftarrow \arg\max\{Gain(A), A \in F\}$
10. **If** $Gain(A_g)<\varepsilon$ **Then**
11. T为单节点树，通过多数表决将S中实例最多的类c_k作为该节点的类标记
12. **Return** T
13. **Else**
14. **For** $j=0$ **To** v **Do** //考察A_g的每一个可能取值
15. 得到S中在A_g上具有a_j值得样本集S_j
16. createDecisionTree(S_j, $F\backslash\{A_0\}$, ε)
17. **End For**

复杂度$O(sn|T|)$

决策树 (5)

- **决策树构造算法示例**

由表中的训练数据集构造概念“buys_computer”的决策树

age	*income*	*student*	*credit_rating*	**Class:** ***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
⋮	⋮	⋮	⋮	⋮
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

“buys_computer”有2个类：c_1代表“yes”，c_2代表“no”；c_1有9个样本， c_2有5个样本。

决策树 (6)

- **决策树构造算法示例**

（1）根据公式$I(s_1,\cdots,s_m)$计算对给定的样本进行分类所需要的期望值

$$I(s_1,s_2)=I(9,5)=-\left(\frac{9}{14}\right)\times\log\left(\frac{9}{14}\right)-\left(\frac{5}{14}\right)\times\log\left(\frac{5}{14}\right)$$

（2）根据公式$E(A)$和$Gain(A)$计算“*age*”变量的信息增益

- 对于“ 30”：$s_{11}=2$，$s_{21}=3$，则$I(s_{11},s_{21})=0.971$
- 对于“31…40”：$s_{12}=4$，$s_{22}=0$，则$I(s_{12},s_{22})=0$
- 对于“>40”：$s_{13}=3$，$s_{23}=2$，则$I(s_{13},s_{23})=0.971$

所以，$E(age)=\left(\frac{5}{14}\right)\times I(s_{11},s_{21})+\left(\frac{4}{14}\right)\times I(s_{21},s_{22})+\left(\frac{5}{14}\right)\times I(s_{13},s_{23})=0.694$。

相应地，$Gain(age)=I(s_1,s_2)-E(age)=0.246$。

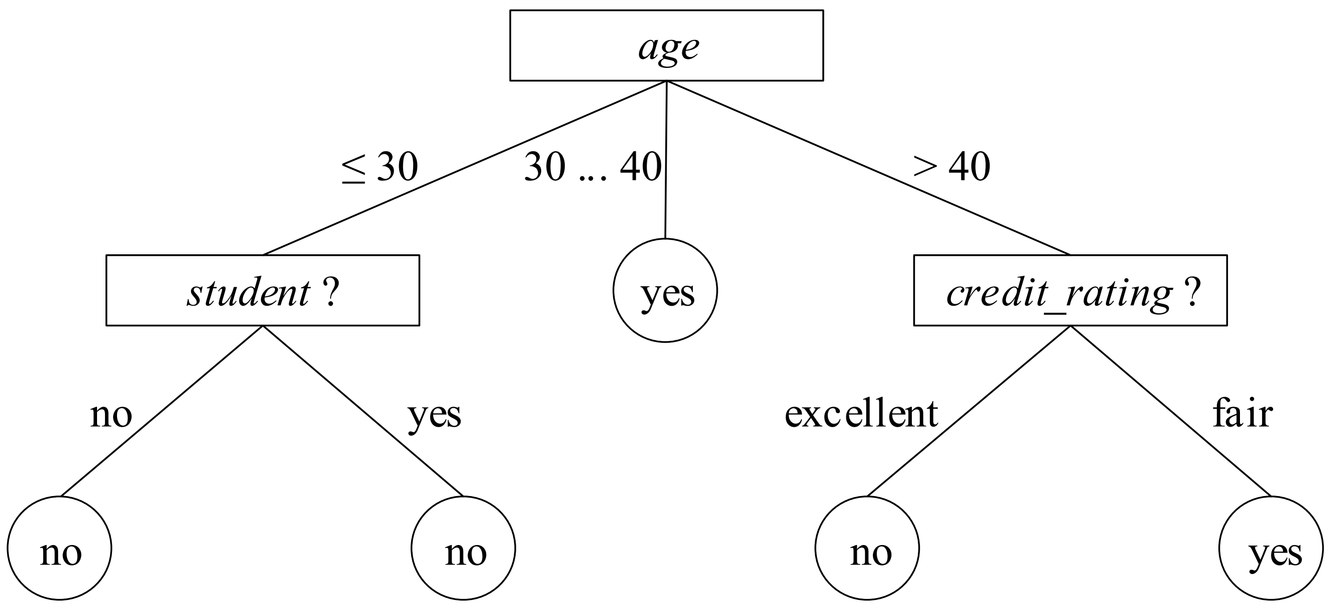
类似地，可以计算出：

Gain(*income*)=0.029，*Gain*(*student*) =0.151，*Gain*(*credit_rating*)= 0.048

决策树 (7)

- **决策树构造算法示例**

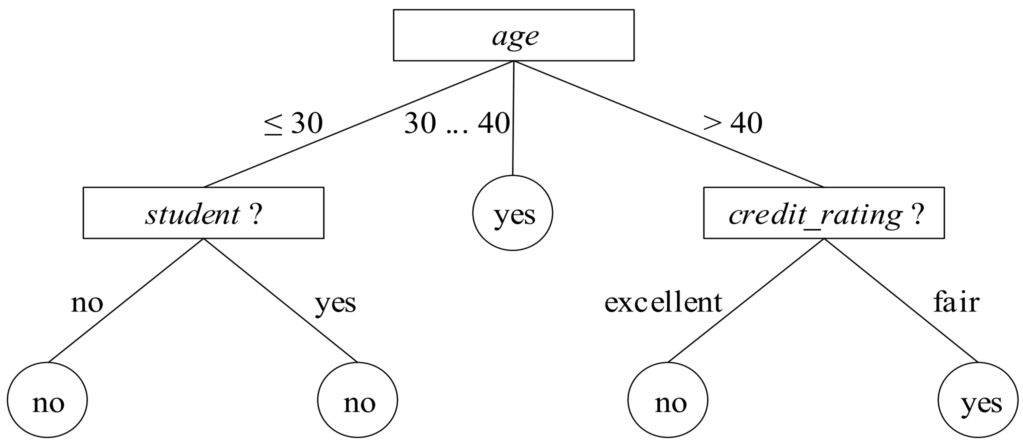
（3）由于“*age*”具有最高的信息增益，则选择它为测试变量，创建“*age*”节点，进行第一次分裂，最终可构造出决策树



决策树 (8)

- **分类规则提取**

  可从已经构造好的决策树中提取形如If-Then的分类规则，每条从根节点到叶子节点的路径对应一个规则



沿着从根节点到叶子节点的路径，可提取出如下的分类规则：

If *age*="≤30" and *student*="no" **Then** *buys_computers*="no"

If *age*="≤ 30" and *student*="yes" **Then** *buys_computers*="yes"

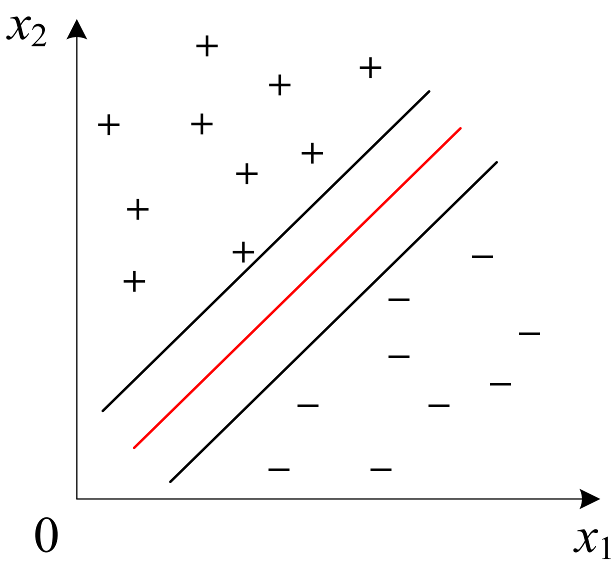
If *age*="31 · · · 40" **Then** *buys_computers*="yes"

提纲

- 分类算法概述
- 决策树
- 支持向量机
- 贝叶斯分类
- 总结

支持向量机 (1)

- **基本概念**
  - ✓ 二分类模型：在样本空间中找出一个超平面来对数据进行分类，并使分类误差尽可能小。
  - ✓ 分离超平面：比所在数据空间小一维的空间，在二维数据空间中是一条直线，在三维数据空间中就是一个平面。



分离超平面将两类训练样本分开

训练集有两个特征和两类标签：

- 特征一用x_1表示，特征二用x_2表示
- 用“+”表示正例，用“–”表示负例

支持向量机 (2)

- **训练算法**

（1）训练数据

一个特征空间上线性可分的数据集$D=\{(x_1,y_1),\ldots,(x_n,y_n)\}$，其中$x_i\in\mathbb{R}^n$，$y_i=\{+1,-1\}$，$i=1, 2, \ldots, n$。$x_i$为第$i$个训练样本的特征向量， y_i为x_i的类标记，(x_i, y_i)称为样本点。当$y_i=+1$时，称x_i为正例；当$y_i=-1$时，称x_i为负例。

（2）寻找最大间隔超平面

通过线性方程$\boldsymbol{w}^T x+b=0$来描述分离超平面，其中 $\boldsymbol{w}=(\boldsymbol{w_1};\ \ldots;\boldsymbol{w_d})$为决定超平面方向的**法向量**，$b$为决定超平面与原点之间距离的**位移项**。

分类策略函数为：

$$f(x)=\mathrm{sign}(\boldsymbol{w}^{\mathrm{T}}x+b)\text{，}\mathrm{sign}(\cdot)\text{ 为符号函数。}$$

支持向量机 (3)

- **训练算法**

给定数据集D和超平面$(\mathbf{w}, b)$，超平面关于样本点(x_i, y_i)的**几何间隔**为：

$$\gamma_i = \frac{y_i(\boldsymbol{w}^T x^i + b)}{\|\boldsymbol{w}\|}, i = 1, \ldots, n$$

若超平面$(\mathbf{w}, b)$能将所有样本点正确分类，则$y_i\left(\boldsymbol{w}^T x_i + b\right) > 0$

若y_i=+1，则正例x_i满足约束条件$\boldsymbol{w}^T x_i + b > 0$；

若y_i=–1，则负例x_i满足约束条件$\boldsymbol{w}^T x_i + b < 0$。

令$y_i(w^T x_i + b) \geq 1$，则**约束条件**表示为：

$$\begin{cases} \boldsymbol{w}^T x_i + b \geq +1, y_i = +1 \\ \boldsymbol{w}^T x_i + b \leq -1, y_i = -1 \end{cases}$$

支持向量机 (4)

- **训练算法**

支持向量：与超平面几何间隔最小且满足约束条件的样本点 $\min\limits_{i=1,\dots,n} \gamma_i$

样本点到超平面的最小几何间隔为$\frac{1}{\|\boldsymbol{w}\|}$

两个异类支持向量到超平面距离之和为$\frac{2}{\|\boldsymbol{w}\|}$，称为间隔

求解最大间隔分离超平面，可表示为以下最优化问题：

$$\max_{\boldsymbol{w},b} \frac{2}{\|\boldsymbol{w}\|} \quad \text{s.t. } y_i\left(\boldsymbol{w}^{\boldsymbol{T}} x_i + b\right) \geq 1,\ i = 1, \dots, n$$

由于$\max\limits_{\boldsymbol{w},b} \frac{2}{\|\boldsymbol{w}\|}$和$\min\limits_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$等价，训练SVM的最优化问题如下：

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{s.t. } \left(\boldsymbol{w}^{\boldsymbol{T}} x_i + b\right) \geq 1,\ i = 1, \dots, n$$

支持向量机 (5)

- **训练算法**

（3）软间隔最大化

实际情况下
几乎不存在

硬间隔：分离超平面能正确划分所有样本

软间隔：

允许某些点不满足约束，可对每个样本点(x_i, y_i)引入松弛变量$\xi_i \geq 0$，则约束条件变为： $y_i\left(\boldsymbol{w}^T x_i + b\right) \geq 1 - \xi_i$

$\xi_i = l_{0/1}(y_i\left(\boldsymbol{w}^T x_i + b\right) - 1)$ 0/1损失函数 $l_{0/1}(z) = \begin{cases} 1, & z < 0 \\ 0, & z \geq 0 \end{cases}$

目标函数变为：$\min\limits_{\boldsymbol{w},b,\xi_i} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i$，正常数$C$称为惩罚系数

优化目标：使$\frac{1}{2}\|\boldsymbol{w}\|^2$尽量小（间隔尽量大），同时使误差分类点的个数尽量少

支持向量机 (6)

- **训练算法**

0/1损失函数：非凸和非连续，是单位跃迁函数，目标函数求解难。

常用凸连续函数替代损失函数来取代“0/1损失函数”。

常用的替代损失函数：

1. hinge损失：$l_{\text{hinge}}(z) = \max(0, 1 - z)$

2. 指数损失：$l_{\exp}(z) = \exp\,(-z)$

3. 对率损失：$l_{\log}(z) = \exp(1 + \exp(-z))$

支持向量机 (7)

- **训练算法**

输入： D，训练数据集；C，惩罚系数

输出： $f(x)$，分类决策函数

步骤：

1. 构造线性支持向量机原始最优化问题：

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{s.t. } y_i(\mathbf{w}^T x_i + b) \geq 1, i = 1, \ldots, n$$

2. 使用拉格朗日乘子求解对偶问题：

$$\max_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i * x_j) - \sum_{i=1}^{N}\alpha_i$$
$$\text{s.t. } \sum_{i=1}^{N}\alpha_i\alpha_j = 0，\ 0 \leq \alpha_i \leq C，\ i = 1, \ldots, n$$

3. 计算法向量$\mathbf{w}$*和位移项b*：

$$\mathbf{w}^* = \sum_{i=1}^{N}\alpha_i^* y_i x_i$$
$$b^* = y_i - \sum_{i=1}^{N} y_i \alpha_i^* (x_i * x_j)$$

4. 计算最大间隔分离超平面以及分类决策函数：

$$\mathbf{w}^* x + b^* = 0$$
$$f(x) = \text{sign}(\mathbf{w}^* x + b^*)$$

Return $f(x)$

时间复杂度 $O(n^3)$
空间复杂 度$O(n^2)$

支持向量机 (8)

- **核函数**
  - 原始样本空间可能不存在能正确划分两类样本的超平面
  - 经过空间转换，在高维空间解决线性问题等价于在低维空间中解决非线性问题

名称	表达式	参数
线性核	$k(x_i, x_j) = x_i^T x_j$	/
多项式核	$k(x_i, x_j) = (x_i^T x_j)^d$	$d \geq 1$为多项式的次数
高斯核	$k(x_i, x_j) = \exp(-\frac{\left\Vert x_i - x_j \right\Vert^2}{2\sigma^2})$	$\sigma > 0$为高斯核的带宽
拉普拉斯核	$k(x_i, x_j) = \exp(-\frac{\left\Vert x_i - x_j \right\Vert}{\sigma})$	$\sigma > 0$
Sigmoid	$k(x_i, x_j) = \tanh(\beta x_i^T x_j + \theta)$	$\beta > 0, \theta < 0$

提纲

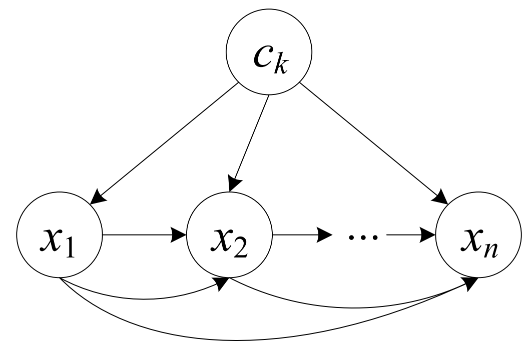
- 分类算法概述
- 决策树
- 支持向量机
- 贝叶斯分类
- 总结

贝叶斯分类 (1)

- **基本概念**
  - 一类以贝叶斯定理为基础、用概率论和统计学知识进行分类的算法
  - 包括朴素贝叶斯分类、链增强朴素贝叶斯分类、树增强朴素贝叶斯分类等
- **朴素贝叶斯分类**
  - 贝叶斯分类器中最简单、应用最为广泛的算法之一
  - 由于假设特征之间相互独立，所以称为“朴素贝叶斯”
  - 分类时对每个类别计算$P(c_k)P(x_i|c_k)$，以$P(c_k)P(x_i|c_k)$的最大项作为待预测样本X所属的类别

贝叶斯分类 (2)

- **贝叶斯分类的基本思想**
  - 设数据集$D=\{x_1, \ldots, x_i, \ldots, x_{n(D)}\}$，样本$x_i$的属性集合$X_i=\{x_{i1}, \ldots, x_{in}\}$，类别集合$C=\{c_1, c_2 \ldots, c_m\}$，即样本可分为$m$个类别



  - 网络结构含有属性集合$X=\{x_1, \ldots, x_i, \ldots, x_n\}$和类别集合$C=\{c_1, \ldots, c_k, \ldots, c_m\}$。对于属性集合为$\{x_1, x_2, \ldots, x_n\}$的待预测数据样本$X$，使$P(c_k|x_1, \ldots, x_n)$最大的分类任务称为贝叶斯分类：

$$C(X) = \arg\max_{c_k \in C}\{P(c_k|x_1, \ldots, x_n)\}$$

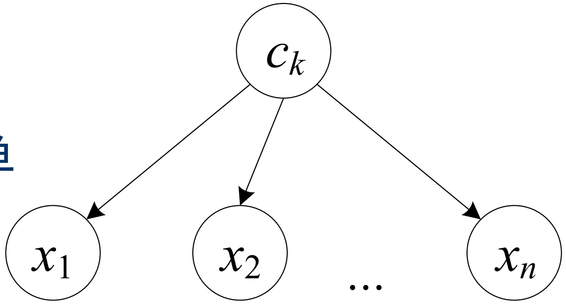
c_k的后验概率为：

$$P(c_k|x_1, \ldots, x_n) = \frac{P(c_k, x_1, \ldots, x_n)}{P(x_1, \ldots, x_n)} = \frac{P(x_1, \ldots, x_n|c_k)P(c_k)}{P(x_1, \ldots, x_n)}$$

没有变量独立假设时计算需指数时间

贝叶斯分类 (3)

- **朴素贝叶斯分类的基本思想**
  - 设在给定类别变量下属性变量之间条件独立，朴素贝叶斯分类使$P(c_k|x_1, \ldots, x_n)$最大
  - 在条件独立性假设下，朴素贝叶斯分类具有简单的星形结构网络结构
  - 每个属性只有唯一的类c_k作为其父节点，这意味着给定类c_k时，$x_1, x_2, \ldots, x_n$条件独立，即



$$P(x_1, \ldots, x_n|c_k) = \prod_{i=1}^{n} P(x_i|c_k)$$

贝叶斯分类 (4)

- **朴素贝叶斯分类的基本思想**
  - 为了降低$P(c_k \mid x_1, \ldots, x_n)$的计算复杂度，根据条件独立性将联合概率分解为：

$$P(c_k, x_1, \ldots, x_n) = P(c_k)P(x_1, \ldots, x_n | c_k) = P(c_k)\prod_{i=1}^{n} P(x_i | c_k)$$

  - 根据联合概率的分解形式，对于给定的待预测样本X，朴素贝叶斯分类形式表示为：

$$C(X) = \arg\max_{c_k \in C}\left\{P(c_k)\prod_{i=1}^{n} P(x_i | c_k)\right\}$$

贝叶斯分类 (5)

- **朴素贝叶斯分类的训练算法**
  - 关键步骤：

    ① 确定特征属性、获取样本数据集

    ② 训练分类器，分别计算每个类别的概率$P(c_k)$和每个属性在该类别下的条件概率$P(x_i|c_k)$

    ③ 对每个类别计算$P(c_k)\prod_{i=1}^{n}P(x_i|c_k)$，以$P(c_k)\prod_{i=1}^{n}P(x_i|c_k)$的最大项作为$X$所属的类别

步骤②中的参数估计，包括**类别概率估计**$\hat{P}(c_k)$和**条件概率估计**$\hat{P}(c_k\,|\,x_i)$

贝叶斯分类 (6)

- **朴素贝叶斯分类的训练算法**
  - 属性值为离散型：

    类别概率估计： $\hat{P}(c_k)=n(c_k)/n(D)$；其中，$n(c_k)$为第c_k类中样本的数量，$n(D)$为样本总数

    条件概率估计： $\hat{P}(x_i|c_k)=n(x_i|c_k)/n(c_k)$；其中，$n(x_i|c_k)$为第$c_k$类中属性为$x_i$的样本数量
  - 属性值为连续型：

    类别概率估计： $\hat{P}(c_k)=n(c_k)/n(D)$；其中，$n(c_k)$为第c_k类中样本的数量，$n(D)$为样本总数

    条件概率估计： $\hat{P}(x_i|c_k)=\frac{1}{\sqrt{2\pi}\sigma_{c_k}}exp\left\{-\frac{\left(x_i-\mu_{c_k}\right)^2}{2\sigma_{c_k}^2}\right\}$；其中，

    $\hat{P}(x_i|c_k)\sim N(\mu_{c_k},\sigma_{c_k}^2)$，$\mu_{c_k}$和$\sigma_{c_k}^2$分别为$c_k$类中$x_i$的均值和方差

贝叶斯分类 (7)

- **朴素贝叶斯分类的训练算法**

输入： D，数据样本集；X，待预测数据的属性集合；C，类别集合

输出： $C(X)$ //以$P(x|y_i)P(y_i)$最大项作为X所属类别

步骤：

1. 统计D中样本的总数$n(D)$
2. 统计D中每类样本的数量$n(c_k)$
3. 统计D中第c_k类中属性为x_i的样本数量$n(x_i|c_k)$
4. 统计X中属性的总数n
5. $\hat{P}(x_i|c_k) \leftarrow 1$
6. $P(c_k|X) \leftarrow \varnothing$

7. **For** k=0 **To** $n(c_k)$ **Do**
8. $\quad\hat{P}(c_k) \leftarrow n(c_k)/n(D)$ //类别概率估计
9. $\quad$**For** j=i **To** n **Do**
10. $\quad\quad\hat{P}(x_j|c_k) \leftarrow (n(x_j|c_k)/n(c_k)) \times \hat{P}(x_j|c_k)$
11. $\quad\quad\hat{P}(c_k|X) \leftarrow \hat{P}(c_k) \times \hat{P}(x_j|c_k)$
12. $\quad\quad P(c_k|X) \leftarrow P(c_k|X) \cup \hat{P}(c_k|X)$
13. $\quad$**End For**
14. **End For**
15. $C(X) = \arg\max\{P(c_k|X)\}$

时间复杂度
$O(n(c_k)\times n)$

贝叶斯分类 (8)

- **朴素贝叶斯分类示例**

任务： 已知某人身高“高”、体重“中”和鞋码“中”，预测其性别。

设“男”和“女”为2个类别，用c_1和c_2表示；属性集合为“身高”、“体重”和“鞋码”，用x_1、x_2和x_3表示。分类步骤如下：

编号	身高	体重	鞋码	性别
1	高	重	大	男
2	高	重	大	男
3	中	中	大	男
4	中	中	中	男
5	矮	轻	小	女
6	矮	轻	小	女
7	矮	中	中	女
8	中	中	中	女

① 类别概率估计：

类别为“男”的概率为$\hat{P}(c_1)=1/2$，类别为“女”的概率为$\hat{P}(c_2)=1/2$。

② 条件概率估计：

性别为“男”、身高“高”、体重“中”、鞋码“中”的概率为

$\hat{P}(x_1, x_2, x_3|c_1)=\hat{P}(x_1|c_1)\hat{P}(x_2|c_1)\hat{P}(x_3|c_1)$

$=(1/2)\times(1/2)\times(1/4)=1/16$

性别为“女”、身高“高”、体重“中”、鞋码“中”的概率为

$\hat{P}(x_1, x_2, x_3|c_2)=\hat{P}(x_1|c_2)\times\hat{P}(x_2|c_2)\times\hat{P}(x_3|c_2)=0$

③ 类别预测：

由于$\hat{P}(c_1)\times\hat{P}(x_1, x_2, x_3|c_1)>\hat{P}(c_2)\times\hat{P}(x_1, x_2, x_3|c_2)$，此人性别为“男”。

提纲

- 分类算法概述
- 决策树
- 支持向量机
- 贝叶斯分类
- 总结

总结

- **决策树**

  优点：易于理解和解释，可处理混合类型的变量，对缺失值不敏感且灵活性好。

  缺点：贪心法的解可能不是最优值，会出现过拟合。

- **支持向量机**

  优点：全局最优值，泛化能力强，算法简单且鲁棒。

  缺点：样本数量大时，存储和计算的开销较大。

- **朴素贝叶斯**

  优点：时空开销小，能处理多分类任务，对缺失数据不太敏感，结果可解释、容易理解。

  缺点：决策存在错误率，对输入数据的表达形式很敏感。

结语

谢谢！