贝叶斯分决策论与贝叶斯类器

主讲人：屠恩美

《机器学习与知识发现》

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

概率回顾

- 三个重要概率
  - 联合概率P(A, B): 事件A和B同时发生的概率
  - 条件概率P(A|B): 已知事件B发生的情况下，事件A发生的概率
  - 边缘概率P(B): 事件B的发生概率，无论A发生与否
- 三个重要关系
  - P(A, B)=P(B, A)=P(A|B)P(B)=P(B|A)P(A) (Bayes公式)
  - A, B独立事件 $\Leftrightarrow$ P(A, B)=P(A)P(B) , P(A|B)=P(A), P(B|A)=P(B)
  - 边缘分布 $\mathrm{P(B)}=\sum_{\mathrm{A}}\mathrm{P(A,B)}$ $\quad p(B)=\int_B p(A,B)dx_b$
- 期望：随机变量 x 的期望 $E[x]=\sum_i x_i P(x_i)$ ， $E[x]=\int_x xp(x)dx$

贝叶斯公式

- 流感通常伴有发烧咳嗽，是不是观察到发烧咳嗽就一定是得了流感？
- 如果流感患者中86%都伴有发烧咳嗽，那么某人出现发烧咳嗽，是流感的概率有多大？

$A=$ 流感，　　$B=$ 发烧咳嗽

已知 $P(B|A)=0.86$，求 $P(A|B)$??

- 由 $P(A,B)=P(B,A)=P(A|B)P(B)=P(B|A)P(A)$ 可得贝叶斯公式：

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

- 含义：观察到事件B的情况下，事件A发生的概率多大！
- 机器学习中最重要的公式之一，也是数学中最优美的公式之一

贝叶斯公式

- 贝叶斯公式等价形式

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

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

$$=\frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^c)P(A^c)};\qquad \left(P(A^c)=1-P(A)\right)$$

贝叶斯公式

- 假设人群中10%人可能得流感，得流感的人中86%有发烧咳嗽，而没得流感的人也有5%有发烧咳嗽（其他原因导致）
- 那么一个人观察到发烧咳嗽，得流感的概率多大？

$A=$ 流感， $B=$ 发烧咳嗽

$$\begin{cases} P(A)=0.1 \\ P(B\mid A)=0.86 \\ P(B\mid A^c)=0.05 \end{cases} \Rightarrow \quad P(A\mid B)=?$$

贝叶斯理论

- 由已知概率可写出如下概率表

条件变量

概率变量

	A = 得流感	A^c = 没流感	
B = 有发烧咳嗽	$0.86 = P(B\|A)$	$0.05 = P(B\|A^c)$	已知给出
B^c = 无发烧咳嗽	$0.14 = P(B^c\|A)$	$0.95 = P(B^c\|A^c)$	计算得到

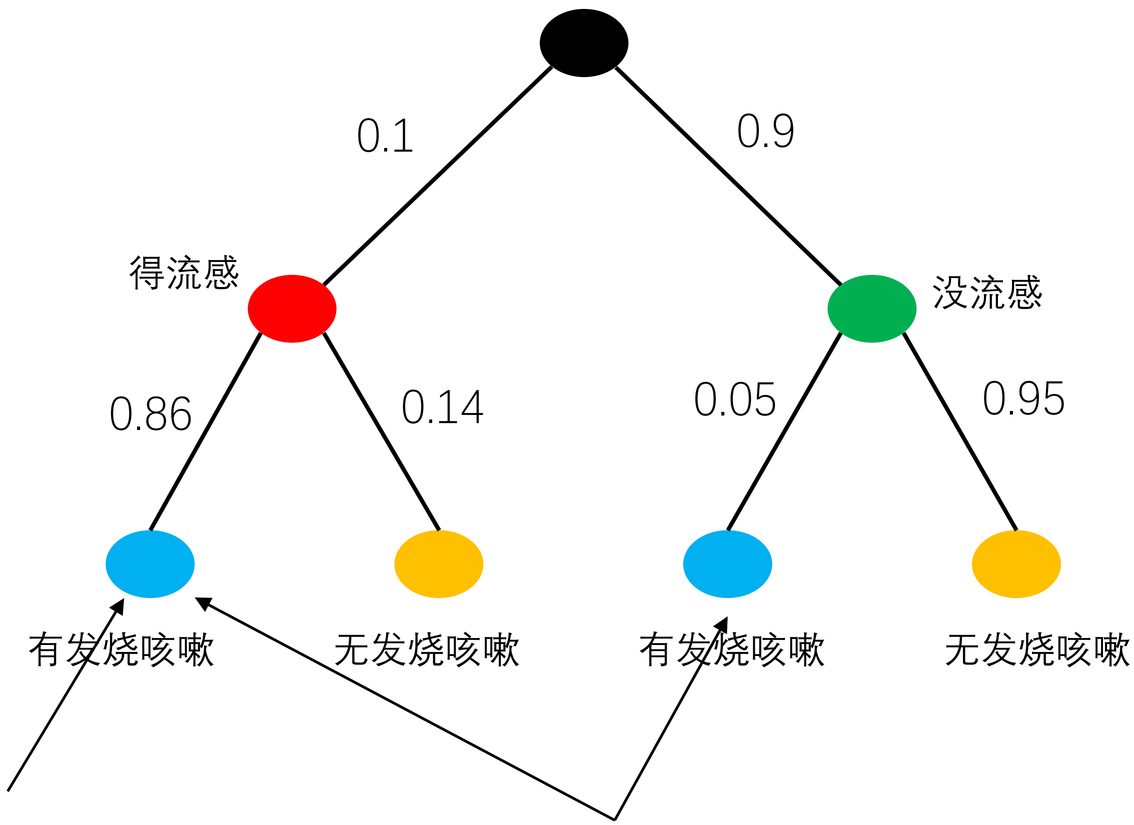
$$P(A)=0.1,\quad P(A^c)=1-P(A)=0.9$$

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}=\frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^c)P(A^c)}$$

$$=\frac{0.86\times0.1}{0.86\times0.1+0.05\times0.0.9}\approx0.656$$

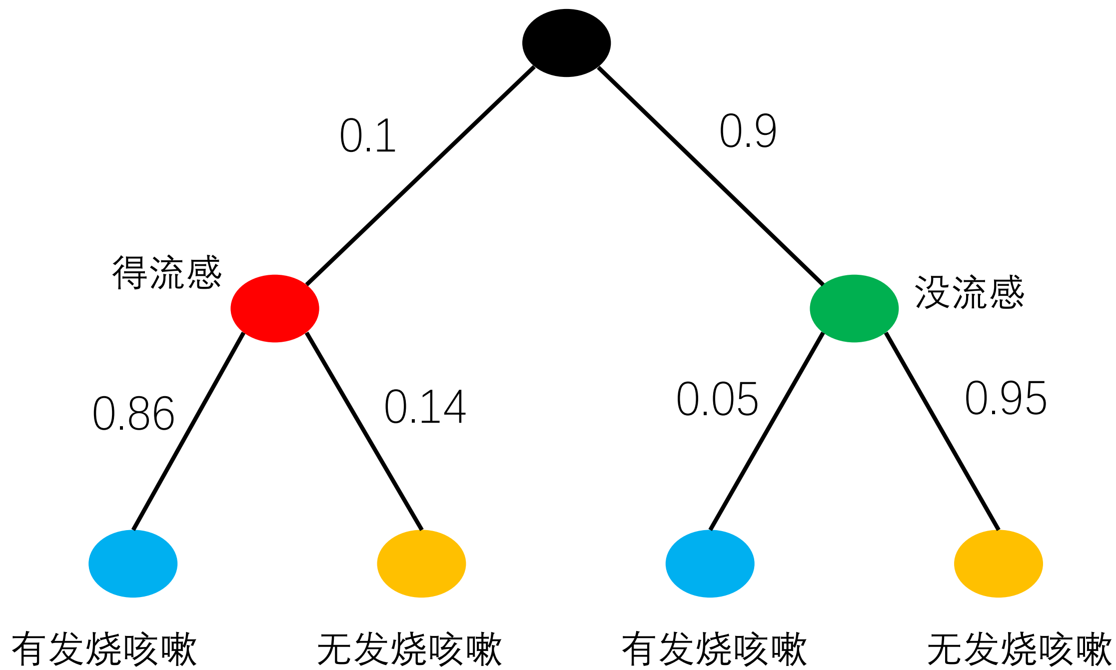
- 问:为什么上表列和为1，而行和不为1？取同一值的概率，不是概率分布

贝叶斯理论



- 流感有发烧咳嗽的概率 ≠ 发烧咳嗽是流感的概率
- 贝叶斯决策：有一些观察量(发烧咳嗽)，计算个体是否属于某一类概率

贝叶斯理论



- 如果记 x :发烧咳嗽(有1无0)， y :流感(有1无0)

采集的样本　　　　　　类别标记

则前面求解的问题就是分类判决问题: $P(y=1|x=1)$

贝叶斯决策论

- 任何的决策都是有风险的!

发烧咳嗽 —判决为→ 无流感 —实际有流感→ 延误病情

发烧咳嗽 —判决为→ 有流感 —实际无流感→ 错误治疗

- 最优的决策就是：使风险最小化!
- 记 c_1:无流感， c_2:无流感，

引入判决损失$\lambda_{ij} > 0$： 判断类别是c_i， 真实类别是c_j

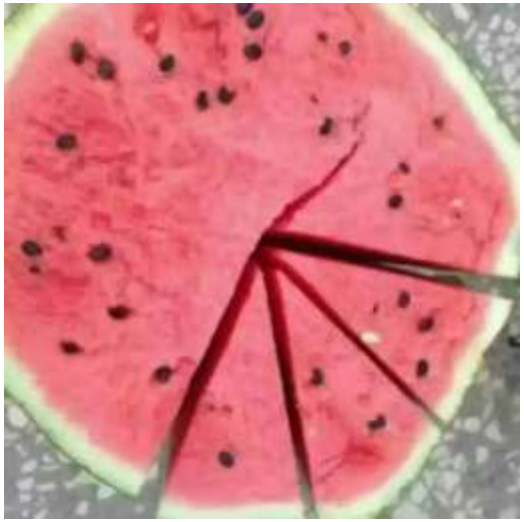
- 那么某次判决的平均风险

$$R = P(A = c_1 \mid B)\lambda_{12} + P(A = c_2 \mid B)\lambda_{21}$$

- 最优判决：最小化 R

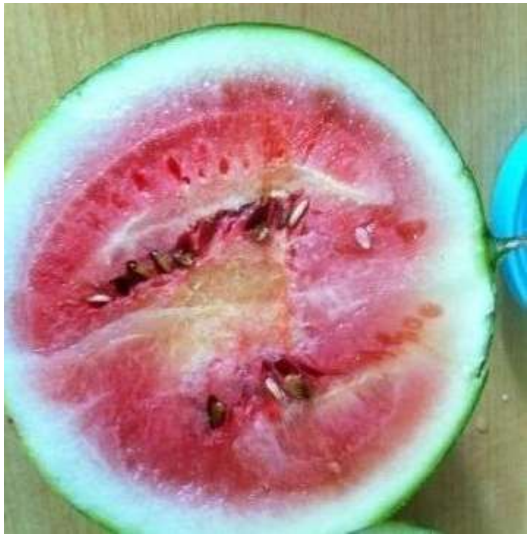
贝叶斯决策论

- 假设数据共有N个类 $\{\mathcal{X}, \mathcal{Y}\}=\{(\mathbf{x}_i, y_i)\}$ ，其中 $y_i \in \mathcal{Y}=\{c_1, c_2, \ldots, c_N\}$
- 例如有三种西瓜: 优质瓜，普通瓜和劣质瓜



第1类 c_1=优质瓜

第2类 c_2=普通瓜



第3类 c_3=劣质瓜

- 样本 $\mathbf{x}$ 的真实类别为 $y=c_j$，则把它误分为类别 c_i 产生的损失记为 $\lambda_{ij}>0$
- 通常正确分类不产生损失，即 $\lambda_{ii}=0$ （j 代表真实类，i 代表判别类）

贝叶斯决策论

- 将样本 $\mathbf{x}$ 判别为第 c_i 类所产生的期望损失（expected loss），也称在样本上的“条件风险”（conditional risk）

$$R(y=c_i \mid \mathbf{x}) = \sum_{j=1}^{N} \lambda_{ij} P(y=c_j \mid \mathbf{x})$$

期望定义：损失的值 x 损失的概率，再求和，也即平均损失

如果真实类是c_j的概率

- **贝叶斯判定准则**（Bayes decision rule）：找到一个判决准则$h:\mathcal{X}\rightarrow\mathcal{Y}$

$$h^* = \arg\min_{h(\mathbf{x})\in\mathcal{Y}} R\left(y=h(\mathbf{x}) \mid \mathbf{x}\right)$$

称为贝叶斯最优分类器，对应的分类精度是机器学习模型精度理论上限

贝叶斯决策论 – 0-1损失

- 错分损失 $\lambda_{i,j}$ 用户设定。作为特例，通常说的分类错误率对应的损失

$$\lambda_{i,j} \begin{cases} 0, & \text{if } i = j; \\ 1, & \text{otherwise}, \end{cases}$$

即分对损失为0，分错损失为1，也称0-1损失

- 0-1损失对应的判别c_i类产生的期望损失为

$$R(y=c_i \mid \mathbf{x}) = \sum_{\substack{j=1 \\ j \neq i}}^{N} \Big(\underbrace{1}_{\lambda_{i,j}} \times P(y=c_j \mid \mathbf{x})\Big) = 1 - P(y=c_i \mid \mathbf{x})$$

判别为c_i类的损失=1- c_i类后验概率

- 于是，最小分类错误率(0-1损失)**对应的**贝叶斯最优分类器(判决准则)为

$$h^*(\mathbf{x}) = \arg\min_{y \in \mathcal{Y}} R(y = h(\mathbf{x}) \mid \mathbf{x}) = \arg\max_{y \in \mathcal{Y}} P(y = h(\mathbf{x}) \mid \mathbf{x})$$

即对样本 $\mathbf{x}$，选择**后验概率** $P(y \mid \mathbf{x})$ **最大**的类，可使期望损失最小。

- **如无特别说明，后面都是针对0-1损失推导**

贝叶斯决策论 – 0-1损失

- 例如，对于两类情况的判决准则（多类判决过程相似）

$$y=\begin{cases} c_1, & \text{if} \quad P(y=c_1\,|\,x)>P(y=c_2\,|\,x) \\ c_2, & \text{if} \quad P(y=c_1\,|\,x)<P(y=c_2\,|\,x) \end{cases}$$

- 由此可见，分类的**关键**在于计算出各类的后验概率 $P(y=c_k|\mathbf{x}), k=1, 2, \ldots, N$
- 给定一个样本$\mathbf{x}$，如何计算它属于每类的后验概率 $P(y=c_i|\mathbf{x})$ 呢？ 有两种基本策略：生成式和判别式
  - 生成式：先对 $P(\mathbf{x}, y=c_i)$ 进行建模，然后利用条件概率公式计算 $P(y=c_i|\mathbf{x})$

$$P(y=c_i\,|\,\mathbf{x})=\frac{P(\mathbf{x}, y=c_i)}{P(\mathbf{x})}$$

贝叶斯决策论 – 0-1损失

- 例如，对于两类情况的判决准则（多类判决过程相似）

$$y=\begin{cases} c_1, & \text{if} \quad P(y=c_1\mid x)>P(y=c_2\mid x) \\ c_2, & \text{if} \quad P(y=c_1\mid x)<P(y=c_2\mid x) \end{cases}$$

- 由此可见，分类的**关键**在于计算出各类的后验概率 $P(y=c_k|\mathbf{x}), k=1, 2, \ldots, N$
- 给定一个样本$\mathbf{x}$，如何计算它属于每类的后验概率 $P(y=c_i|\mathbf{x})$ 呢？ 有两种基本策略：生成式和**判别式**
  - **判别式**：先极大似然估计<u>类条件概率</u> $P(\mathbf{x}\mid y=c_i)$，然后利用贝叶斯公式计算 $P(y=c_i\mid\mathbf{x})$；

$$P(y=c_i\mid\mathbf{x})=\frac{P(\mathbf{x}\mid y=c_i)P(c_i)}{P(\mathbf{x})}$$

- 常见的监督学习，例如线性模型、决策树、SVM等算法属于判别式。

类条件概率估计

- 现在我们测量了一个瓜的特征$\mathbf{x}$，怎么判断属于哪类？由贝叶斯公式知

$$P(y=c_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y=c_i)P(y=c_i)}{p(\mathbf{x})}$$

类条件概率估计

- 现在我们测量了一个瓜的特征$\mathbf{x}$，怎么判断属于哪类？由贝叶斯公式知

$$P(y=c_i\,|\,\mathbf{x}) = \frac{p(\mathbf{x}\,|\,y=c_i)P(y=c_i)}{p(\mathbf{x})}$$

$p(\mathbf{x}\,|\,y=c_i)$ → 类条件概率	**核心问题**
$P(y=c_i)$ → 先验概率	容易获得
$p(\mathbf{x})$ → 证据因子	不太关心
$P(y=c_i\,|\,\mathbf{x})$ → 后验概率	最终目标

- 先验概率 $P(y=c_i)$：训练样本中c_i类样本占比。例如

类别	优质瓜	普通瓜	劣质瓜
占比	30%	60%	10%
先验概率	0.3	0.6	0.1

- 类条件概率分布 $p(\mathbf{x}\,|\,y=c_i)$：其他条件相同情况下，$p(\mathbf{x}\,|\,y=c_i)$ 值较大时，c_i 是真实类别的可能性，因此类条件概率又称似然(likelihood)
- **极大似然法**估计类条件概率 $p(\mathbf{x}|\,y=c_i)$：先假定其具有某种确定的**概率分布形式**$P(\mathbf{x}\,|\,y=c_i,\boldsymbol{\theta})$，再基于训练样本对**概率分布参数 $\boldsymbol{\theta}$** 进行估计。

极大似然估计

- **符号约定**：$y=c\ (c\in\mathcal{Y})$ 类的类条件概率 $P(\mathbf{x}\,|\,y=c,\boldsymbol{\theta})$ 简记为 $P(\mathbf{x}\,|\,\boldsymbol{\theta}_c)$，其中 $\boldsymbol{\theta}_c$ 是待定参数，我们的任务就是利用训练数据集 D 估计参数 $\boldsymbol{\theta}_c$。
- 例如，常见的参数待定概率分布：

分布名称	概率/概率密度	待估计参数 $\boldsymbol{\theta}_c$
Poisson $\mathrm{Pois}(\lambda)$	$P(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}$ $k\in\{0,1,2,\ldots\}$	λ
Uniform $\mathrm{Unif}(a,b)$	$f(x)=\frac{1}{b-a}$ $x\in(a,b)$	a,b
Normal $\mathcal{N}(\mu,\sigma^2)$	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$ $x\in(-\infty,\infty)$	μ,σ
Exponential $\mathrm{Expo}(\lambda)$	$f(x)=\lambda e^{-\lambda x}$ $x\in(0,\infty)$	λ

极大似然法

- 给定训练数据 D， 其中属于 c 类的样本记为 $D_c = \{(\mathbf{x}_i, y_i = c_i)\}_{i=1}^{m}$。
- 假设D_c中的样本都是**独立随机**采样，定义参数 $\boldsymbol{\theta}_c$ 相对于D_c的似然函数

$$L(\boldsymbol{\theta}_c) = \prod_{i=1}^{m} p(\mathbf{x}_i \mid \boldsymbol{\theta}_c)$$

关于$\boldsymbol{\theta}_c$ 的函数

独立事件：连乘表示同时出现的概率。
似然函数的含义：在某个参数$\boldsymbol{\theta}_c$下，所有样本同时出现的概率

- 极大似然：找到使所有样本同时出现**可能性最大**的一组参数值，即

$$\boldsymbol{\theta}_c^* = \arg\max L(\boldsymbol{\theta}_c) = \arg\max \prod_{i=1}^{m} p(\mathbf{x}_i \mid \boldsymbol{\theta}_c)$$

极大似然法

- 为了求解方便，通常利用对数性质把连乘转换为求和

$$LL(\boldsymbol{\theta}_c) = \log L(\boldsymbol{\theta}_c) = \sum_{i=1}^{m} \log p(\mathbf{x}_i \mid \boldsymbol{\theta}_c)$$

- 求解过程
  - 如果$LL(\boldsymbol{\theta}_c)$对参数$\boldsymbol{\theta}_c$可导，则极大值在一阶导数为0处取得

$$\frac{\partial LL(\boldsymbol{\theta}_c)}{\partial \boldsymbol{\theta}_c} = \sum_{i=1}^{m} \frac{\partial \log p(\mathbf{x}_i \mid \boldsymbol{\theta}_c)}{\partial \boldsymbol{\theta}_c}$$

$$= \sum_{i=1}^{m} \frac{1}{p(\mathbf{x}_i \mid \boldsymbol{\theta}_c)} \frac{\partial p(\mathbf{x}_i \mid \boldsymbol{\theta}_c)}{\partial \boldsymbol{\theta}_c}$$

  - 如果$LL(\boldsymbol{\theta}_c)$对参数$\boldsymbol{\theta}_c$不可导，则具体分析表达式的可能极值点

极大似然法 – 正态分布例子

- 假设 $y=c\in\mathcal{Y}$ 类是正太分布为例，则需要估计的参数是 $\boldsymbol{\theta}_c=(\mu,\sigma)$

$$p(\mathbf{x}\mid\boldsymbol{\theta}_c)=\frac{1}{\sqrt{2\pi\sigma}}e^{\frac{-(x-\mu)^2}{2\sigma^2}},\quad(\sigma>0)$$

- 随机采用一组样本 $D_c=\{(x_i,y_i=c)\}_{i=1}^m$， 似然函数

$$LL(\boldsymbol{\theta}_c)=\sum_{i=1}^m\log p(x_i\mid\boldsymbol{\theta}_c)=\sum_{i=1}^m\log\left(\frac{1}{\sqrt{2\pi\sigma}}e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}\right)$$

$$=\sum_{i=1}^m\left(\log\frac{1}{\sqrt{2\pi\sigma}}+\frac{-(x_i-\mu)^2}{2\sigma^2}\right)$$

$$=\sum_{i=1}^m\left(-\log\left(\sqrt{2\pi}\sigma\right)+\frac{-x_i^2-\mu^2+2x_i\mu}{2\sigma^2}\right)$$

极大似然法 – 正态分布例子

- $LL(\boldsymbol{\theta}_c)$对$(\mu,\sigma)$ 均可导

$$\begin{cases} \dfrac{\partial LL(\boldsymbol{\theta}_c)}{\partial \mu} = \sum_{i=1}^{m} \dfrac{-2x_i + 2\mu}{2\sigma} = 0 \\ \dfrac{\partial LL(\boldsymbol{\theta}_c)}{\partial \sigma} = \sum_{i=1}^{m} -\dfrac{1}{\sigma} + \dfrac{-x_i^2 - \mu^2 + 2x_i\mu}{\sigma^3} = 0 \end{cases}$$

$$\Downarrow$$

$$\begin{cases} \mu = \dfrac{1}{m}\sum_{i=1}^{m} x_i \\ \sigma = \dfrac{1}{m}\sum_{i=1}^{m} (x_i - \mu)^2 \end{cases}$$

与直观相符合

多维高斯分布

$$\begin{cases} \boldsymbol{\mu} = \dfrac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i \\ \boldsymbol{\Sigma} = \dfrac{1}{m}\sum_{i=1}^{m} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \end{cases}$$

理一理思路……

- 0-1分类器的期望损失 $R(y=c_i\,|\,\mathbf{x}) = 1 - P(y=c_i\,|\,\mathbf{x})$
- 因此最小化损失就要最大化后验概率，即最优分类器 $c = h^*(\mathbf{x})$ 要满足

$$h^*(\mathbf{x}) = \arg\max_{y\in\mathcal{Y}} P(y = h(\mathbf{x})\,|\,\mathbf{x})$$

- 由贝叶斯公式可知，要计算 $P(y=c\,|\,\mathbf{x})$ ，就要知道 $P(\mathbf{x}\,|\,y=c), P(c)$

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

- $P(c)$ 通常容易计算（各类样本占比）
- 假设 $P(\mathbf{x}\,|\,y=c)$ 是具有未知参数的某种分别（如高斯分别），则可利用**极大似然法**从**训练数据中**估计出未知参数

举个栗子……

- 17个样本，每个样本2个特征，分两类
- 训练集：样本1-16，测试集：样本17
- 假设每类后验概率服从两维高斯分别

$$p(\mathbf{x} \mid y=c_i;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)=\frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}_i|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right)$$

$i=1,2$

$\boldsymbol{\mu}_i\in\mathbb{R}^2,\boldsymbol{\Sigma}_i\in\mathbb{R}^{2\times2}$ 是待估计参数

- 构造似然函数

$$LL(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2)=\sum_{j=1}^{16}\log\left(p(\mathbf{x}_j \mid y=c_i;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\right)$$

- 令 LL 对 $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2$ 的导数等于0，可得

$$\begin{cases}\boldsymbol{\mu}_1=0.125\sum_{j=1}^{8}\mathbf{x}_j\\ \boldsymbol{\Sigma}_1=0.125\sum_{j=1}^{8}\left(\mathbf{x}_j-\boldsymbol{\mu}_1\right)\left(\mathbf{x}_j-\boldsymbol{\mu}_1\right)^T\end{cases}\qquad\begin{cases}\boldsymbol{\mu}_2=0.125\sum_{j=9}^{16}\mathbf{x}_j\\ \boldsymbol{\Sigma}_2=0.125\sum_{j=9}^{16}\left(\mathbf{x}_j-\boldsymbol{\mu}_2\right)\left(\mathbf{x}_j-\boldsymbol{\mu}_2\right)^T\end{cases}$$

	密度	含糖率	好瓜
1	0.697	0.46	是
2	0.774	0.376	是
3	0.634	0.264	是
4	0.608	0.318	是
5	0.556	0.215	是
6	0.403	0.237	是
7	0.481	0.149	是
8	0.437	0.211	是
9	0.666	0.091	否
10	0.243	0.267	否
11	0.245	0.057	否
12	0.343	0.099	否
13	0.639	0.161	否
14	0.657	0.198	否
15	0.36	0.37	否
16	0.593	0.042	否
17	0.719	0.103	否

举个栗子……

- 由贝叶斯公式可计算属于每个类的后验概率

$$P(y_{17}=c_1\mid\mathbf{x}_{17})=\frac{p(\mathbf{x}_{17}\mid y_{17}=c_1)p(c_1)}{p(\mathbf{x}_{17})}$$

$$=\frac{p(\mathbf{x}_{17}\mid\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)\times 0.5}{p_{\times}}$$

$$P(y_{17}=c_2\mid\mathbf{x}_{17})=\frac{P(\mathbf{x}_{17}\mid y_{17}=c_2)P(c_2)}{P(\mathbf{x}_{17})}$$

$$=\frac{p(\mathbf{x}_{17}\mid\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)\times 0.5}{p_{\times}}$$

	密度	含糖率	好瓜
1	0.697	0.46	是
2	0.774	0.376	是
3	0.634	0.264	是
4	0.608	0.318	是
5	0.556	0.215	是
6	0.403	0.237	是
7	0.481	0.149	是
8	0.437	0.211	是
9	0.666	0.091	否
10	0.243	0.267	否
11	0.245	0.057	否
12	0.343	0.099	否
13	0.639	0.161	否
14	0.657	0.198	否
15	0.36	0.37	否
16	0.593	0.042	否
17	0.719	0.103	否

- 比较二者大小，把样本17归入概率最大的一类（因为分类只比较大小，因此 p_x可不计算）

朴素贝叶斯分类器

- 朴素贝叶斯分类器(Naïve Bayes Classifier)采用了“**属性条件独立性假设**”，即每个属性独立地对分类结果发生影响
- 假设样本 $\mathbf{x}$ 有 d 个属性，为 x_i 在第 i 个属性上的取值，则后验概率为

$$P(c \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid c) P(c)}{p(\mathbf{x})} = \frac{P(c)}{p(\mathbf{x})} \prod_{i=1}^{d} p(x_i \mid c)$$

- 因为$p(\mathbf{x})$对所有属性都一样，因此对分类判别没有作用。利用贝叶斯判断准，朴素贝叶斯分类器

$$c^* = \arg\max_{c \in \mathcal{Y}} P(c) \prod_{i=1}^{d} p(x_i \mid c)$$

朴素贝叶斯分类器

- 朴素贝叶斯分类器

$$c^* = \arg\max_{c \in \mathcal{Y}} P(c) \prod_{i=1}^{d} p(x_i \mid c)$$

- 对于给定的一组样本 $D_c = \{(\mathbf{x}_i, y_i = c)\}_{i=1}^{m}$ ，如果是离散属性

$$P(c) = \frac{|D_c|}{|D|}, \quad P(x_i \mid c) = \frac{|D_{c,x_i}|}{|D_c|}$$

D_c中第i个属性取值为x_i的样本数

如果第i个属性是连续属性，则利用最大似然估计每个属性的类概率密度函数 $p(x_i \mid c)$ ，然后可利用朴素贝叶斯分类器

单变量概率密度，容易许多

朴素贝叶斯分类器 – 举个例子

- 西瓜数据集3.0

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜	
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.46	是	
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是	
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是	
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是	
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是	
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是	训练
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是	
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是	
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否	
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否	
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否	
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否	
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否	
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否	
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.36	0.37	否	
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否	
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否	测试

朴素贝叶斯分类器 – 举个例子

- 训练过程就是计算各类先验概率和各属性类别分布概率的过程
- 包含两类：c_1=是， c_2=否。先验概率：

$$P(c_1)=\frac{8}{16}=0.5,\quad P(c_2)=\frac{8}{16}=0.5$$

- 类别分布概率 $p(x_i\,|\,c)$： x_1=色泽

$$P(色泽=青绿|\,y=c_1)=\frac{3}{8}\qquad P(色泽=青绿|\,y=c_2)=\frac{2}{8}$$

$$P(色泽=乌黑|\,y=c_1)=\frac{4}{8}\qquad P(色泽=乌黑|\,y=c_2)=\frac{2}{8}$$

$$P(色泽=浅白|\,y=c_1)=\frac{1}{8}\qquad P(色泽=浅白|\,y=c_2)=\frac{4}{8}$$

编号	色泽	好瓜
1	青绿	是
2	乌黑	是
3	乌黑	是
4	青绿	是
5	浅白	是
6	青绿	是
7	乌黑	是
8	乌黑	是
9	乌黑	否
10	青绿	否
11	浅白	否
12	浅白	否
13	青绿	否
14	浅白	否
15	乌黑	否
16	浅白	否
17	青绿	否

朴素贝叶斯分类器 – 举个例子

- 类似第可以计算其他**离散属性**类别分布概率 $p(x_i \mid c)$

  x_i = 根蒂，敲声，纹理，脐部，触感

  $P(x_i\text{=取值} \mid y = c_1)$ $P(x_i\text{=取值} \mid y = c_2)$

- 密度和含糖量是**连续属性**，如何计算 $p(x_i \mid c)$？
  - 假设：概率密度函数(以高斯分布为例)

$$p(x_i \mid c) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

  - 极大似然法估计**每类中每个属性**参数 μ, σ

	密度	含糖量
c_1	$\mu: 0.57;\ \sigma: 0.13$	$\mu: 0.28;\ \sigma: 0.10$
c_2	$\mu: 0.47;\ \sigma: 0.19$	$\mu: 0.16;\ \sigma: 0.11$

密度	含糖率	好瓜
0.697	0.46	是
0.774	0.376	是
0.634	0.264	是
0.608	0.318	是
0.556	0.215	是
0.403	0.237	是
0.481	0.149	是
0.437	0.211	是
0.666	0.091	否
0.243	0.267	否
0.245	0.057	否
0.343	0.099	否
0.639	0.161	否
0.657	0.198	否
0.36	0.37	否
0.593	0.042	否
0.719	0.103	否

朴素贝叶斯分类器 – 举个例子

- 测试样本

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103

判决准则：$$c^* = \arg\max_{c\in\mathcal{Y}} P(c)\prod_{i=1}^{d} p(x_i \mid c)$$

- 是好瓜的后验概率 c_1=是

$$P(\text{色泽=青绿}| \, y=c_1) \times P(\text{根蒂=蜷缩}| \, y=c_1) \times \cdots \times P(\text{触感=硬滑}| \, y=c_1) \times$$

$$p(\text{密度}=0.719| \, y=c_1) \times p(\text{含糖率}=0.103| \, y=c_1) = 0.0014$$

- 不是好瓜的后验概率 c_2=否

$$P(\text{色泽=青绿}| \, y=c_2) \times P(\text{根蒂=蜷缩}| \, y=c_2) \times \cdots \times P(\text{触感=硬滑}| \, y=c_2) \times$$

$$p(\text{密度}=0.719| \, y=c_2) \times p(\text{含糖率}=0.103| \, y=c_2) = 0.0078$$

数据缺失与隐含属性

- 现实应用中常常遇到两种情况：
  - “不完整”的样本：西瓜已经脱落的根蒂，无法看出是“蜷缩”还是“坚挺”,则训练样本的“根蒂”属性变量值未知，如何计算？
  - 无法直接测量属性：要测量西瓜含糖量就要打开西瓜，这样就破坏了西瓜的完整性，怎么估计？
- 第一种属于**部分样本**属性缺失，第二种属于**所有样本**的共同未知属性

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

无法直接使用极大似然法进行类条件概率估计

- 期望最大化EM (Expectation-Maximization)算法是常用的估计数据缺失和隐含属性的利器。

EM算法

- 缺失属性或隐含属性统称为隐变量，记为$\mathbf{z}$
  - 隐变量 $\mathbf{z}$ 是待确定的**数据参数**，与类条件概率或模型无关
  - 模型参数$\boldsymbol{\theta}_c$是待估计的**模型参数** (类概率参数)，与数据无关
- 此时类条件概率密度记为 $p(\mathbf{x},\mathbf{z}\,|\,\boldsymbol{\theta}_c)$ ，**是关于x, z的联合概率分布，具体形式则由参数$\boldsymbol{\theta}_c$决定**
- 而似然函数是关于 $\mathbf{z}$ 和 $\boldsymbol{\theta}_c$ 的函数

$$LL(\boldsymbol{\theta}_c,\mathbf{z}) = \sum_{i=1}^{m} \log p(\mathbf{x}_i,\mathbf{z}\,|\,\boldsymbol{\theta}_c)$$

对于训练数据，$(\mathbf{x}_i, y_i)$已知

EM算法

- 给定训练数据 $D_c=\{(\mathbf{x}_i, y_i=c_i)\}_{i=1}^m$ ，EM算法包括E步和M步：
  - **E 步**(Expectation)：若模型参数$\boldsymbol{\theta}_c$已知，则**z**的概率分布 $p(\mathbf{x},\mathbf{z}\,|\,\boldsymbol{\theta}_c)$ 可知。对**z**求期望可消去隐变量**z** 的影响，得到关于**x**的类条件概率

$$p(\mathbf{x}\,|\,\boldsymbol{\theta}_c)=\int_{\mathbf{z}}\mathbf{z}\,p(\mathbf{x},\mathbf{z}\,|\,\boldsymbol{\theta}_c)d\mathbf{z}$$

  - **M 步**(Maximization)： 若隐变量**z**已知，则利用极大似然法估计模型参数$\boldsymbol{\theta}_c$

$$\boldsymbol{\theta}_c^*=\arg\max_{\boldsymbol{\theta}_c} LL(\boldsymbol{\theta})=\arg\max_{\boldsymbol{\theta}_c}\sum_{i=1}^m \log P(\mathbf{x}_i\,|\,\boldsymbol{\theta}_c)$$

- 随机初始化参数$\boldsymbol{\theta}_c$和 **z**， 然后 E 步和 M 步交替迭代直到收敛
- (例子间后面GMM聚类)

总结

- 贝叶斯决策论

$$h^*(\mathbf{x}) = \arg\max_{y\in\mathcal{Y}} P(y\,|\,\mathbf{x})$$

关键问题如何求 $P(y=c_i\,|\,\mathbf{x}) = \dfrac{p(\mathbf{x}\,|\,y=c_i)P(y=c_i)}{p(\mathbf{x})}$

- 极大似然：假设类条件概率 $P(\mathbf{x}\,|\,y=c,\boldsymbol{\theta}_c)$ 然后构建似然函数并估计 $\boldsymbol{\theta}_c$

$$\boldsymbol{\theta}_c^* = \arg\max L(\boldsymbol{\theta}_c) = \arg\max \prod_{i=1}^{m} p(\mathbf{x}_i\,|\,\boldsymbol{\theta}_c)$$

- 朴素贝叶斯：给属性相互独立，因此类条件概率可拆分

$$P(c\,|\,\mathbf{x}) = \frac{p(\mathbf{x}\,|\,c)P(c)}{p(\mathbf{x})} = \frac{P(c)}{p(\mathbf{x})}\prod_{i=1}^{d} p(x_i\,|\,c)$$

- EM算法处理缺失属性或隐含属性**z**：E步求**z**期望，M步极大似然估计参数