河北师范大学软件学院
Software College of Hebei Normal University

# 机器学习入门

## 贝叶斯决策

河北师范大学软件学院　张朝晖

2017.04.16-04.26

**监督式学习**

- 分类
  - A. 懒惰学习--KNN法(K最近邻法)分类模型
  - B. 概率学习 --贝叶斯分类器
  - C. 分而治之--分类树
- 回归
  - D. KNN法回归
  - E. 分而治之--回归树
  - F. 最小二乘回归

**集成学习**

**非监督式学习**

- 聚类 -- G. K-Means Clustering
- 特征提取 -- H. PCA

**模型评价**

# 回顾两种典型的贝叶斯分类模型

# 模型1--最小错误率贝叶斯分类

$$posterior = \frac{likelihood \times prior}{evidence} \qquad \sum_{i=1}^{c} P(\omega_j \mid x) = 1$$

观测空间为连续特征空间

$$P(\omega_j \mid x) = \frac{p(x, \omega_j)}{p(x)} = \frac{p(x \mid \omega_j) P(\omega_j)}{p(x)} = \frac{p(x \mid \omega_j) P(\omega_j)}{\sum_{i=1}^{c} p(x \mid \omega_i) P(\omega_i)}$$

观测空间为离散特征空间

$$P(\omega_j \mid x) = \frac{P(x, \omega_j)}{P(x)} = \frac{P(x \mid \omega_j) P(\omega_j)}{P(x)} = \frac{P(x \mid \omega_j) P(\omega_j)}{\sum_{i=1}^{c} P(x \mid \omega_i) P(\omega_i)}$$

两类情况下，基于最小错误率的判决规则，各等价形式：

1，状态后验概率：

$$\begin{cases} P(\omega_1 \mid x) > P(\omega_2 \mid x), 则 x \in \omega_1 类； \\ P(\omega_1 \mid x) < P(\omega_2 \mid x), 则 x \in \omega_2 类 \end{cases}$$

2，似然值×先验概率

$$\begin{cases} P(\omega_1) p(x \mid \omega_1) > P(\omega_2) p(x \mid \omega_2), 则 x \in \omega_1 类； \\ P(\omega_1) p(x \mid \omega_1) < P(\omega_2) p(x \mid \omega_2), 则 x \in \omega_2 类 \end{cases}$$

3，似然比

$$\begin{cases} 若\ \ l(x) = \dfrac{p(x \mid \omega_1)}{p(x \mid \omega_2)} > \dfrac{P(\omega_2)}{P(\omega_1)}, 则\ \ x \in \omega_1 类； \\ 若\ \ l(x) = \dfrac{p(x \mid \omega_1)}{p(x \mid \omega_2)} < \dfrac{P(\omega_2)}{P(\omega_1)}, 则\ \ x \in \omega_2 类 \end{cases}$$

多类情况下，基于最小错误率的判决规则，各等价形式

(1)后验概率：

若 $\boldsymbol{P}(\omega_i \mid \boldsymbol{x}) = \max\limits_{j=1,2,\ldots,c} \boldsymbol{P}(\omega_j \mid \boldsymbol{x})$

则 $\boldsymbol{x} \in \omega_i$类

(2)似然值×先验概率：

若 $\boldsymbol{p}(\boldsymbol{x} \mid \omega_i)\boldsymbol{P}(\omega_i) = \max\limits_{j=1,2,\ldots,c} \boldsymbol{p}(\boldsymbol{x} \mid \omega_j)\boldsymbol{P}(\omega_j)$

则 $\boldsymbol{x} \in \omega_i$类

对于观测$\boldsymbol{x}$

(1) 计算后验概率：

$$P(\omega_j \mid \boldsymbol{x}) = \frac{P(\omega_j)\,p(\boldsymbol{x} \mid \omega_j)}{p(\boldsymbol{x})} = \frac{P(\omega_j)\,p(\boldsymbol{x} \mid \omega_j)}{\sum_{i=1}^{c} P(\omega_i)\,p(\boldsymbol{x} \mid \omega_i)}$$

$$j = 1, 2, \dots, c$$

(2) 计算$\boldsymbol{x}$的条件风险：

$$R(\alpha_i \mid \boldsymbol{x}) = \sum_{j=1}^{c} \lambda\left(\alpha_i, \omega_j\right) P(\omega_j \mid \boldsymbol{x})$$

$$i = 1, 2, \dots, k$$

(3) 决策：选择关于$\boldsymbol{x}$的条件风险最小的决策$\alpha$，即

$$\alpha = \arg \min_{i=1,2,\dots,k} R(\alpha_i \mid \boldsymbol{x})$$

# 连续特征空间

## 若各类条件概率密度函数为正态分布

# 主要内容

1. 引言
2. 贝叶斯决策模型
   2.1 最小错误率贝叶斯决策
   2.2 最小风险的贝叶斯决策
3. 正态分布的概率密度函数及性质
4. 概率/概率密度函数估计
5. 朴素贝叶斯(Naive Bayes)分类

**问题的引入：**

- 贝叶斯决策中，涉及“连续随机变量或向量的**类条件概率密度函数**”。
- “正态分布”的概率密度函数，特点：

**物理上的合理性。**如果在特征空间中的某一类样本，较多地分布在这一类均值附近，远离均值点的样本比较少，此时用正态分布作为这一类的概率模型是合理的。

**数学上，比较简便。**正态分布概率模型有许多好的性质，有利于作数学分析。

→ 简单、符合一些实际情况

# 主要内容

3.1. 单个随机变量的正态分布

3.2.随机向量的正态分布

连续随机变量$\boldsymbol{X}$概率密度函数定义

$$p_X(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

记 $p_X(x)\sim \boldsymbol{N}\left(\mu,\sigma^2\right)$

标准正态分布：$p_X(x)\sim \boldsymbol{N}\left(0,1\right)$

其中
$$\begin{cases}\mu-\text{随机变量}\boldsymbol{X}\text{的期望，或均值}\\ \qquad \mu\equiv \boldsymbol{E}\left\{\boldsymbol{X}\right\}=\int_{-\infty}^{\infty}xp_X(x)dx\\ \sigma-x\text{的标准差，描述}x\text{的分散程度}\\ \sigma^2-x\text{的方差：}\quad \sigma^2\equiv \boldsymbol{E}\left\{\left(\boldsymbol{X}-\mu\right)^2\right\}=\int_{-\infty}^{\infty}\left(x-\mu\right)^2p_X(x)dx\end{cases}$$

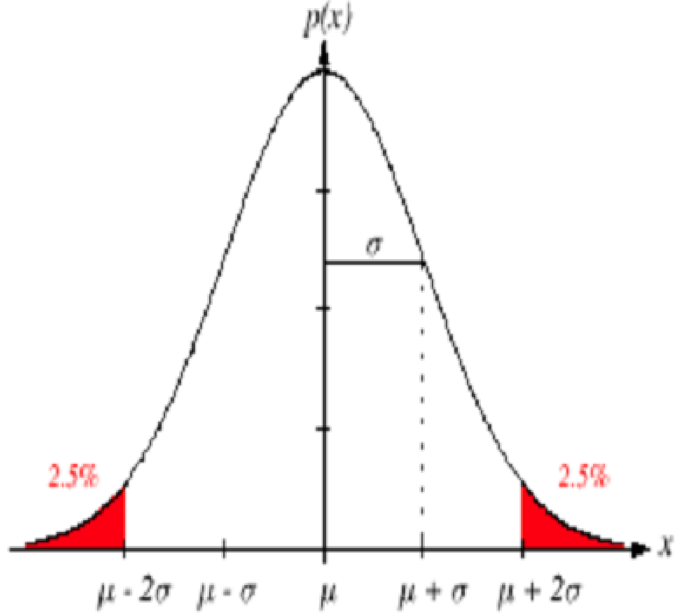
性质： $p_X(x) \ge 0, -\infty < x < +\infty$

$$\int_{-\infty}^{\infty} p_X(x)dx = 1$$

$$\int_{\mu-\sigma}^{\mu+\sigma} p_X(x)dx = 0.683$$

$$\int_{\mu-2\sigma}^{\mu+2\sigma} p_X(x)dx = 0.9544$$

$$\int_{\mu-3\sigma}^{\mu+3\sigma} p_X(x)dx = 0.9974$$




图：单变量正态分布。

其它：熵 $\boldsymbol{H}(\boldsymbol{X}) = -\int p_X(x) \ln p_X(x)dx > 0$

***Mahalanobis***距离 $\boldsymbol{r} = \dfrac{|\boldsymbol{x}-\mu|}{\sigma}$

$$\boldsymbol{E}\{f(\boldsymbol{X})\} = \int_{-\infty}^{\infty} f(x)\, p_X(x)dx$$

$$\boxed{p_X(x=\mu) = \frac{1}{\sqrt{2\pi}\sigma}}$$

# 主要内容

3.1 单个随机变量的正态分布

3.2 随机向量的正态分布

## [1]多元正态分布的概率密度函数--$p_X(x)$

定义：$$p_X(x)=\frac{1}{(2\pi)^{d/2}\left|\Sigma\right|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

记为 $p_X(x)\sim N(\boldsymbol{\mu},\Sigma)$

其中：

- $\boldsymbol{X}$ $-d$维列向量，$\boldsymbol{X}=\left[X^{(1)},X^{(2)},\ldots,X^{(d)}\right]^T$
  $X^{(i)}$为随机变量，$i=1,2,\ldots,\boldsymbol{d}$;
- $\boldsymbol{\mu}$ $-d$维均值向量，$\boldsymbol{\mu}=\left[\mu^{(1)},\mu^{(2)},\ldots,\mu^{(d)}\right]^T$
  $\mu^{(i)}$为第$i$维随机变量$X^{(i)}$的期望，$i=1,2,\ldots,\boldsymbol{d}$;
- $\Sigma-d\times d$维协方差矩阵（*covariance matrix*）
- $\Sigma^{-1}$ –矩阵$\Sigma$的逆矩阵，**精度矩阵**
- $\left|\Sigma\right|$ –矩阵$\Sigma$的行列式

## [1]多元正态分布概率密度函数定义(续) --$\boldsymbol{\mu}$

$\boldsymbol{\mu}$ $-d$维均值向量，$\boldsymbol{\mu}=\left[\mu^{(1)},\mu^{(2)},\ldots,\mu^{(d)}\right]^{T}$

$$\boldsymbol{\mu}\equiv \boldsymbol{E}\left\{\boldsymbol{X}\right\}=\int_{\Re^d}\boldsymbol{x}p_{\boldsymbol{X}}(x)d\boldsymbol{x}$$

$$=\left[\boldsymbol{E}\left\{X^{(1)}\right\},\boldsymbol{E}\left\{X^{(2)}\right\},\ldots,\boldsymbol{E}\left\{X^{(d)}\right\}\right]^{T}=\left[\mu^{(1)},\mu^{(2)},\ldots,\mu^{(d)}\right]^{T}$$

其中：

$$\begin{cases}\mu_i=\boldsymbol{E}\left\{X^{(i)}\right\}=\int_{\Re^d}x^{(i)}p_{\boldsymbol{X}}(x)d\boldsymbol{x}=\int_{-\infty}^{+\infty}x^{(i)}p_{X^{(i)}}(x^{(i)})dx^{(i)}\\ \\ \text{边缘密度函数}p_{X^{(i)}}(x^{(i)})=\int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty}p_{\boldsymbol{X}}(x)dx^{(1)}\ldots dx^{(i-1)}dx^{(i+1)}\ldots dx^{(d)}\end{cases}$$

## [1]多元正态分布概率密度函数定义(续)--$\Sigma$

$\Sigma - d\times d$维协方差矩阵(*covariance matrix*):

$$\Sigma \equiv \boldsymbol{E}\left\{(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T\right\}$$

$$= \boldsymbol{E}\left\{\begin{bmatrix} X^{(1)}-\mu^{(1)} \\ X^{(2)}-\mu^{(2)} \\ \vdots \\ X^{(d)}-\mu^{(d)} \end{bmatrix}\begin{bmatrix} X^{(1)}-\mu^{(1)} & X^{(2)}-\mu^{(2)} & \cdots & X^{(d)}-\mu^{(d)} \end{bmatrix}\right\}$$

$$= \boldsymbol{E}\begin{bmatrix} \left(X^{(1)}-\mu^{(1)}\right)^2 & \left(X^{(1)}-\mu^{(1)}\right)\left(X^{(2)}-\mu^{(2)}\right) & \cdots & \left(X^{(1)}-\mu^{(1)}\right)\left(X^{(d)}-\mu^{(d)}\right) \\ \left(X^{(2)}-\mu^{(2)}\right)\left(X^{(1)}-\mu^{(1)}\right) & \left(X^{(2)}-\mu^{(2)}\right)^2 & \cdots & \left(X^{(2)}-\mu^{(2)}\right)\left(X^{(d)}-\mu^{(d)}{}_d\right) \\ \vdots & \vdots & \vdots & \vdots \\ \left(X^{(d)}-\mu^{(d)}\right)\left(X^{(1)}-\mu^{(1)}\right) & \left(X^{(d)}-\mu^{(d)}\right)\left(X^{(2)}-\mu^{(2)}\right) & \cdots & \left(X^{(d)}-\mu^{(d)}\right)^2 \end{bmatrix}$$

$=$(下页待续)

## [1]多元正态分布概率密度函数定义(续)--$\Sigma$

$\Sigma - d \times d$维协方差矩阵(*covariance matrix*):

$$\Sigma \equiv \boldsymbol{E}\left[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^{T}\right]$$

$$=\begin{bmatrix} \boldsymbol{E}\left\{\left(X^{(1)}-\mu^{(1)}\right)^{2}\right\} & \boldsymbol{E}\left\{\left(X^{(1)}-\mu^{(1)}\right)\left(X^{(2)}-\mu^{(2)}\right)\right\} & \cdots & \boldsymbol{E}\left\{\left(X^{(1)}-\mu^{(1)}\right)\left(X^{(d)}-\mu^{(d)}\right)\right\} \\ \boldsymbol{E}\left\{\left(X^{(2)}-\mu^{(2)}\right)\left(X^{(1)}-\mu^{(1)}\right)\right\} & \boldsymbol{E}\left\{\left(X^{(2)}-\mu^{(2)}\right)^{2}\right\} & \cdots & \boldsymbol{E}\left\{\left(X^{(2)}-\mu^{(2)}\right)\left(X^{(d)}-{\mu^{(d)}}_{d}\right)\right\} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{E}\left\{\left(X^{(d)}-\mu^{(d)}\right)\left(X^{(1)}-\mu^{(1)}\right)\right\} & \boldsymbol{E}\left\{\left(X^{(d)}-\mu^{(d)}\right)\left(X^{(2)}-\mu^{(2)}\right)\right\} & \cdots & \boldsymbol{E}\left\{\left(X^{(d)}-\mu^{(d)}\right)^{2}\right\} \end{bmatrix}$$

$$=\begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots\cdots & \sigma_{1\mathbf{d}} \\ \sigma_{21} & \sigma_{22} & \cdots\cdots & \sigma_{2d} \\ \cdots & \cdots & \cdots\cdots & \cdots \\ \sigma_{d1} & \sigma_{d2} & \cdots\cdots & \sigma_{dd} \end{bmatrix} \qquad \Sigma\text{对称非负定}\begin{cases}\text{对角线元素}\sigma_{ii}=\sigma^{(i)2}, \quad X_i\text{方差} \\ \text{非对角元素}\sigma_{ij}, \ X_i\text{、}X_j\text{协方差}\end{cases}$$

这里：只考虑**$\Sigma$对称正定**

## [2]多元正态分布概率密度函数的几个典型性质

### 性质1 多元正态分布完全由$\mu$、$\Sigma$决定

参数共计$d+\dfrac{d(d+1)}{2}$个

$$\boldsymbol{\mu}=\left[\mu^{(1)},\mu^{(2)},\ldots,\mu^{(d)}\right]^T$$

$$\Sigma=\begin{bmatrix}\sigma_{11} & \sigma_{12} & \cdots\cdots & \sigma_{1\mathbf{d}} \\ \sigma_{21} & \sigma_{22} & \cdots\cdots & \sigma_{2d} \\ \cdots & \cdots & \cdots\cdots & \cdots \\ \sigma_{d1} & \sigma_{d2} & \cdots\cdots & \sigma_{dd}\end{bmatrix}$$ 为对称矩阵

$$p_X(\boldsymbol{x})=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

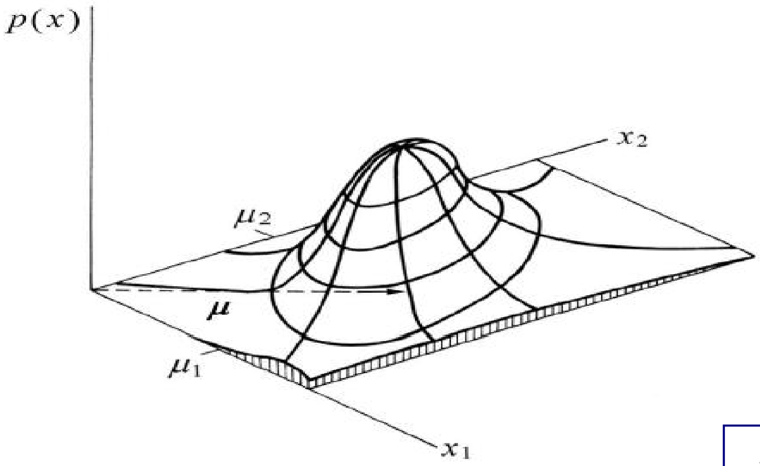
$$p_X(\boldsymbol{x})\sim N(\boldsymbol{\mu},\Sigma)$$

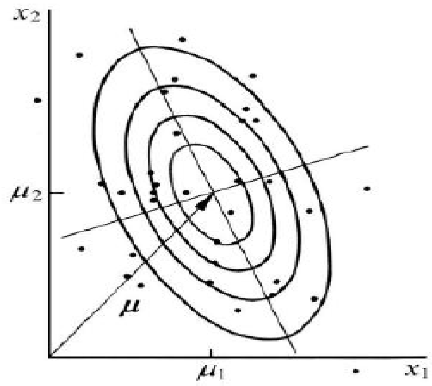
## 性质2 等概率密度点的轨迹为一超椭球面

$$p_X(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

$x$到$\mu$的 *M*ahalanobis**距离**平方 $r^2 = (x-\mu)^T\Sigma^{-1}(x-\mu)$

等概率密度点：$r^2 =$ 常数




二维正态分布，等概率密度点的轨迹为**椭圆周**。

超椭球面中心$\mu$

超椭球面主轴 $\begin{cases}\text{方向取决于}\Sigma\text{本征向量}\\ \text{长度与}\Sigma\text{本征值的平方根成正比。}\end{cases}$

**性质3 不相关性等价于独立性。**

$$\boldsymbol{x}\text{多元正态分布}\begin{cases} ''\boldsymbol{X}\text{任意两分量}X_i,X_j\text{间互不相关}'' \\ \qquad \Leftrightarrow\ ''\text{各分量间相互独立}'' \\ \text{协方差矩阵}\Sigma\text{是对角矩阵} \Rightarrow \boldsymbol{X}\text{各分量相互独立} \end{cases}$$

---

定义：随机变量$X^{(i)},X^{(j)}$之间$\boldsymbol{\mu}=\left[\mu^{(1)},\mu^{(2)},\ldots,\mu^{(d)}\right]^{\boldsymbol{T}}$

$$\begin{cases} X^{(i)},X^{(j)}\text{间不相关：}\quad \boldsymbol{E}\left\{X^{(i)}X^{(j)}\right\}=\boldsymbol{E}\left\{X^{(i)}\right\}\boldsymbol{E}\left\{X^{(j)}\right\} \\ X^{(i)},X^{(j)}\text{相互独立：}\quad p_{X^{(i)}X^{(j)}}(x^{(i)},x^{(j)})=p_{X^{(i)}}(x^{(i)})p_{X^{(j)}}(x^{(j)}) \end{cases}$$

并且

$$\boxed{p_{X^{(i)}X^{(j)}}\left(x^{(i)},x^{(j)}\right)=p_{X^{(i)}}(x^{(i)})p_{X^{(j)}}(x^{(j)})} \Rightarrow \boxed{\boldsymbol{E}\left\{X^{(i)}X^{(j)}\right\}=\boldsymbol{E}\left\{X^{(i)}\right\}\boldsymbol{E}\left\{X^{(j)}\right\}}$$

## 性质3 不相关性等价于独立性。

多元正态分布的任意随机变量间不相关性，等价于独立性。

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots\cdots & \sigma_{1\mathbf{d}} \\ \sigma_{21} & \sigma_{22} & \cdots\cdots & \sigma_{2d} \\ \cdots & \cdots & \cdots\cdots & \cdots \\ \sigma_{d1} & \sigma_{d2} & \cdots\cdots & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} \sigma_{11} & & & \\ & \sigma_{22} & & \\ & & \cdots\cdots & \\ & & & \sigma_{\mathbf{dd}} \end{bmatrix}$$

若$X_i, X_j$间互不相关，则协方差矩阵$\Sigma$是对角的

## 性质3 不相关性等价于独立性。

$$(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})=\sum_{i=1}^{d}\left(\frac{x^{(i)}-\mu^{(i)}}{\sigma^{(i)}}\right)^2$$

$$p_X(\boldsymbol{x})=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

$$=\frac{1}{(2\pi)^{d/2}\prod_{i=1}^{d}\sigma^{(i)}}\exp\left[-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x^{(i)}-\mu^{(i)}}{\sigma^{(i)}}\right)^2\right]=\prod_{i=1}^{d}\frac{1}{\sqrt{2\pi}\sigma^{(i)}}\exp\left[-\frac{1}{2}\left(\frac{x^{(i)}-\mu^{(i)}}{\sigma^{(i)}}\right)^2\right]$$

$$=\prod_{i=1}^{d}p_{X^{(i)}}(x^{(i)})$$

对于多元正态分布$\boldsymbol{X}$，

$$\begin{cases}\text{各分量}X_i,X_j\text{间互不相关}\Leftrightarrow\text{各分量相互独立。}\\ \text{协方差矩阵}\Sigma\text{是对角的}\Rightarrow\boldsymbol{X}\text{各分量相互独立，且正态分布}\end{cases}$$

# 下一节 主题

如何估计类条件概率密度函数?

# 主要内容

1. 引言
2. 贝叶斯决策模型
   2.1 最小错误率贝叶斯决策
   2.2 最小风险的贝叶斯决策
3. 正态分布的概率密度函数及性质
4. <u>概率/概率密度函数估计</u>
5. 朴素贝叶斯(Naive Bayes)分类

# PART1. 问题的引入

## ➢ 如何基于贝叶斯决策解决实际问题?

**基本思路：**

[1] **利用有限规模训练样本,估计**

$$\begin{cases} \text{先验概率} & P(\omega_i) \\ \text{条件概率密度} & p(x|\omega_i) \end{cases} \quad i=1,2,\ldots,c$$

[2] **利用估计的**$\hat{P}(\omega_i)$、$\hat{p}(x|\omega_i)$，**设计贝叶斯分类器**；

**对未知样本**$x$**进行判决**

**上述过程又称基于样本的两步贝叶斯决策**。

问题：如何估计有关概率/概率密度函数?

## ➢ 如何进行有关概率/概率密度函数的估计？

**前提** $\begin{cases}\text{训练样本数量足够多}\\ \text{训练样本分布能代表样本的真实分布(如：}i.i.d\text{)}\end{cases}$

$$\Rightarrow \begin{cases} \hat{p}(x|\omega_i) \xrightarrow{N\to\infty} p(x|\omega_i) \\ \hat{P}(\omega_i) \xrightarrow{N\to\infty} P(\omega_i) \end{cases}$$

**[1] 状态先验概率$P(\omega_i)$，$i=1,2,\ldots,c$ 的估计**

比较容易 $\begin{cases}\textbf{依靠经验}\\ \text{如：异常细胞识别，医生可根据}\\ \quad\text{以往细胞病理检查统计结果，}\\ \quad\text{做出推断}\\ \textbf{利用训练数据中各类出现的频度估计}\end{cases}$

## [2]类条件概率密度$p(x|\omega_i)$，$i=1,2,\ldots,c$ 的估计

**客观情况**
- (1)应包含随机变量的全部信息
- (2)可以是满足如下条件的任何函数

  $p(x)\geq 0$，　$\int p(x)dx=1$

**实际情况**
- (1)训练样本数目不够多
- (2)噪声污染、甚至部分特征丢失
- (3)特征维数对分类器计算复杂度的影响

$\Rightarrow$ **类条件概率密度估计非常困难**

## 概率密度函数估计基本方法



[1] 参数估计（$parametric\ estimation$）

类条件总体概率密度函数

- **形式已知** 根据对问题一般认识，假设随机变量$x$服从某种分布
- **参数未知** 利用训练样本估计分布参数

如：若 $p(\boldsymbol{x}|\omega_i) \sim N(\mu,\Sigma)$，则待估计参数为$\theta=(\mu,\Sigma)$

**两种类型的参数估计法**

- **监督参数估计**($supervised\ parametric\ estimation$) 各训练样本类别状态已知
- **非监督参数估计**($nonsupervised\ parametric\ estimation$) 各训练样本类别状态未知

**概率密度函数的估计 $\Rightarrow$ 参数估计**

## [2] 非参数估计（*nonparametric estimation*）

概率密度函数形式未知；

利用训练样本，直接推断概率密度函数

# PART2. 概率密度函数的参数估计法之一
# ----最大似然估计
# (Maximum Likelihood Estimation, MLE)

# 1.最大似然估计的问题描述



**假设条件**

[1] 参数$\theta$：确定的未知量（$\theta$不是随机变量）

[2] $c$类 $\begin{cases} c\text{个确定类别的样本集}\mathcal{X}_i, i=1,\ldots,c. \\ \text{各样本集的样本个数}N_i, \ i=1,\ldots,c \end{cases}$

[3] 独立同分布（i.i.d : *independent identical distribution*）

样本集$\mathcal{X}_i$中，各样本依$p(x|\omega_i)$独立抽取，$i=1,\ldots,c$

[4] 类条件概率密度$p(x|\omega_i), i=1,\ldots,c$ $\begin{cases} \text{函数形式确定} \\ \text{参数}\theta_i\text{未知} \end{cases}$

$p(x|\omega_i)$依赖于$\theta_i$，记：$p(x|\omega_i;\theta_i)$，或$p(x;\theta_i)$

[5] 各类样本只含本类的分布信息

各参数$\theta_i$在函数上相互独立，$\theta_i$的估计不受$\mathcal{X}_j(j\neq i)$影响。

$\Rightarrow$ **目标** 根据各训练样本集$\mathcal{X}_1,\ldots,\mathcal{X}_c$，分别估计各参数(向量)

$\theta_1,\ldots,\theta_c$的"最可能"的取值，记为 $\hat{\theta}_1,\ldots,\hat{\theta}_c$

## 2.最大似然估计的基本思想

转化为c个独立问题，各问题均可表述为

已知：

**某类**训练样本集 $\mathcal{X}=\{x_1,\ldots,x_N\}$;

各样本按已知整体分布形式$p(x;\theta)$独立抽取$(\mathrm{i.i.d})$;

参数向量$\theta=[\theta_1,\ldots,\theta_S]^T$

目标：

采用最大似然估计法，确定该类参数$\theta$的估计$\hat{\theta}$.

## 2.最大似然估计的基本思想(续1)

### (1)似然函数(*likelihood function*)

$$\boxed{l(\theta)=p(\mathcal{X};\theta)=p(x_1,\ldots,x_N;\theta)=\prod_{i=1}^{N}p(x_i;\theta)}$$

参数$\theta$下，观测到样本集$\mathcal{X}=\{x_1,\ldots,x_N\}$的概率

或：参数$\theta$下，$N$个独立随机样本$x_1,\ldots,x_N$的联合概率。

样本集$\mathcal{X}$固定时，$l(\theta)$随$\theta$的取值而变，是θ的函数。

$\Rightarrow l(\theta)$体现了参数$\theta$下取得样本集$\mathcal{X}$的可能性

是$\theta$关于$\mathcal{X}$的似然函数

## 2.最大似然估计的基本思想(续2)

Software College of Hebei Normal University

(2)对数似然函数

$$H(\theta) = \ln l(\theta) = \ln p(\mathscr{X};\theta) = \sum_{i=1}^{N} \ln p(x_i;\theta)$$

(3)最大似然估计量

若 $\hat{\theta} = \arg\max_{\theta\in\Theta}\left[l(\theta)\right] = \arg\max_{\theta\in\Theta}\left[\ln l(\theta)\right]$

则 $\hat{\theta}$是$\theta$的最大似然估计量。

记：$\hat{\theta} = d(x_1,\ldots,x_N) = d(\mathscr{X})$

或：$\hat{\theta} = \hat{\theta}(x_1,\ldots,x_N) = \hat{\theta}(\mathscr{X})$

最大似然估计的实质：

根据抽取的$N$个样本$x_1,\ldots,x_N$，估计它们"最可能"来自哪一个密度函数.

对数似然函数$H(\theta)$关于$l(\theta)$是单调增加的，所以最大似然估计量$\hat{\theta}$必然使$H(\theta)$、$l(\theta)$同时最大.

# 3.最大似然估计的必要条件

对于密度函数 $p(\boldsymbol{x};\theta)$

待估计的参数向量 $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_S]^T$

似然函数

$$l(\theta) = p(\mathcal{X};\theta) = p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N;\theta) = \prod_{i=1}^{N} p(\boldsymbol{x}_i;\theta)$$

对数似然函数

$$H(\theta) = \ln l(\theta) = \ln p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N;\boldsymbol{\theta})$$

$$= \sum_{k=1}^{N} \ln p(\boldsymbol{x}_k;\boldsymbol{\theta}) = \sum_{k=1}^{N} \ln p(\boldsymbol{x}_k;\theta_1, \ldots, \theta_S)$$

最大似然估计量

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \left[ l(\theta) \right] = \arg\max_{\theta \in \Theta} \left[ H(\theta) \right]$$

记　梯度算子$\nabla_{\theta}=\left[\dfrac{\partial}{\partial\theta_1}\quad\cdots\quad\dfrac{\partial}{\partial\theta_s}\right]^T$

若　似然函数$l(\theta)$, $H(\theta)$关于$\theta$**连续可微**，则$\nabla_{\theta}H(\theta)=0$

$$\nabla_{\theta}H(\theta)=\begin{bmatrix}\dfrac{\partial H(\theta)}{\partial\theta_1}\\ \vdots\\ \dfrac{\partial H(\theta)}{\partial\theta_s}\end{bmatrix}=\begin{bmatrix}\sum_{k=1}^{N}\dfrac{\partial}{\partial\theta_1}\left[\ln p(x_k;\theta)\right]\\ \vdots\\ \sum_{k=1}^{N}\dfrac{\partial}{\partial\theta_s}\left[\ln p(x_k;\theta)\right]\end{bmatrix}$$

$$=\sum_{k=1}^{N}\begin{bmatrix}\dfrac{\partial}{\partial\theta_1}\left[\ln p(x_k;\theta)\right]\\ \vdots\\ \dfrac{\partial}{\partial\theta_s}\left[\ln p(x_k;\theta)\right]\end{bmatrix}=\sum_{k=1}^{N}\nabla_{\theta}\left[\ln p(x_k;\theta)\right]$$

可得最大似然估计的**必要条件**：

最大似然估计量$\hat{\theta}$必满足

$$\nabla_{\theta} H(\theta) = \sum_{k=1}^{N} \nabla_{\theta}\left[\ln p(x_k;\theta)\right] = 0$$

或 $\nabla_{\theta} l(\theta) = 0$

$$\nabla_{\theta} H(\theta) = 0, \quad \text{即方程组} \begin{cases} \sum_{k=1}^{N} \dfrac{\partial}{\partial \theta_1}\left[\ln p(x_k;\theta)\right] = 0 \\ \vdots \\ \sum_{k=1}^{N} \dfrac{\partial}{\partial \theta_s}\left[\ln p(x_k;\theta)\right] = 0 \end{cases}$$

若上述方程组的某个解$\hat{\theta}$能使$H(\theta)$最大，则$\hat{\theta}$为$\theta$的最大似然估计。

# PART3. 单变量正态分布概率密度函数的最大似然估计

某类类条件概率密度函数 $p(x \mid \omega_i;\boldsymbol{\theta}) = p(x;\boldsymbol{\theta})$

$$p(x;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \frac{1}{\sqrt{2\pi\theta_2}}\exp\left[-\frac{(x-\theta_1)^2}{2\theta_2}\right]$$

待估计参数 $\boldsymbol{\theta} = [\theta_1,\theta_2]^{\boldsymbol{T}} = [\mu,\sigma^2]^{\boldsymbol{T}}$

该类样本集 $\mathcal{X} = \{x_1,\ldots,x_N\}$ 各样本独立抽取($\boldsymbol{i.i.d}$)

似然函数

$$\boldsymbol{l}(\boldsymbol{\theta}) = \boldsymbol{p}(\mathcal{X};\boldsymbol{\theta}) = \prod_{\boldsymbol{i}=1}^{\boldsymbol{N}} \boldsymbol{p}(\boldsymbol{x_i};\boldsymbol{\theta}) = \frac{1}{(2\pi\theta_2)^{\frac{N}{2}}}\exp\left[-\frac{\sum_{i=1}^{N}(x_i-\theta_1)^2}{2\theta_2}\right]$$

对数似然函数 $\boldsymbol{H}(\boldsymbol{\theta}) = \ln \boldsymbol{l}(\boldsymbol{\theta}) = -\frac{\sum_{i=1}^{N}(x_i-\theta_1)^2}{2\theta_2} - \frac{\boldsymbol{N}}{2}\ln(2\pi\theta_2)$

由于： $\boldsymbol{H}(\theta)=\ln \boldsymbol{l}(\theta)=-\dfrac{\sum_{i=1}^{N}(x_i-\theta_1)^2}{2\theta_2}-\dfrac{N}{2}\ln(2\pi\theta_2)$

所以 $\nabla_{\theta}\boldsymbol{H}(\theta)=\begin{bmatrix}\dfrac{\partial}{\partial\theta_1}\boldsymbol{H}(\theta)\\ \dfrac{\partial}{\partial\theta_2}\boldsymbol{H}(\theta)\end{bmatrix}=\begin{bmatrix}\dfrac{1}{\theta_2}\sum_{i=1}^{N}(x_i-\theta_1)\\ -\dfrac{N}{2\theta_2}+\dfrac{\sum_{i=1}^{N}(x_i-\theta_1)^2}{2\theta_2^{\ 2}}\end{bmatrix}$

最大似然估计量$\hat{\boldsymbol{\theta}}$满足方程：$\nabla_{\theta}\boldsymbol{H}(\boldsymbol{\theta})=\boldsymbol{0}$

即 $\begin{cases}\sum_{i=1}^{N}\dfrac{1}{\hat{\theta}_2}(x_i-\hat{\theta}_1)=0\\ \sum_{i=1}^{N}\left[-\dfrac{1}{2\hat{\theta}_2}+\dfrac{(x_i-\hat{\theta}_1)^2}{2\hat{\theta}_2^{\ 2}}\right]=0\end{cases}$

解得 $\begin{cases}\hat{\mu}=\hat{\theta}_1=\dfrac{1}{N}\sum_{i=1}^{N}x_i\\ \widehat{\sigma^2}=\hat{\theta}_2=\dfrac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu})^2\end{cases}$

样本集$\mathcal{X}=\{x_1,\ldots,x_N\}$各样本依$p(x;\theta)$独立抽取

$$\begin{cases}\text{样本均值} & \bar{x}=\dfrac{1}{N}\sum_{i=1}^{N}x_i \\ \text{样本方差} & s^2=\dfrac{1}{N-1}\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\end{cases}$$

参数$\mu$、$\sigma^2$的最大似然估计

$$\begin{cases}\hat{\mu}=\bar{x} \\ \widehat{\sigma^2}=\dfrac{1}{N}\sum_{k=1}^{N}\left(x_k-\hat{\mu}\right)^2\end{cases}$$

# PART4. 多元正态分布概率密度函数的最大似然估计

- 一般情况
- 各特征相互独立

**对于多变量正态分布** $\mu$、$\Sigma$ 均未知

$d$维特征空间

样本集 $\mathcal{X}=\{x_1,\dots,x_N\}$ $x_i=\left[x_i^{(1)},\dots,x_i^{(d)}\right]^T, i=1,\dots,N$

待估计的参数 $\theta$

均值向量 $\boldsymbol{\mu}=\left[\mu^{(1)},\mu^{(2)},\dots,\mu^{(d)}\right]^T$

协方差矩阵 $\Sigma=\begin{bmatrix}\sigma_{11} & \sigma_{12} & \cdots\cdots & \sigma_{1d}\\ \sigma_{21} & \sigma_{22} & \cdots\cdots & \sigma_{2d}\\ \cdots & \cdots & \cdots\cdots & \cdots\\ \sigma_{d1} & \sigma_{d2} & \cdots\cdots & \sigma_{dd}\end{bmatrix}$

其中：$\sigma_{ij}=\sigma_{ji}$

概率密度函数

$$p(\boldsymbol{x};\boldsymbol{\theta})=p(\boldsymbol{x};\mu,\Sigma)=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

样本集 $\mathcal{X}=\{x_1,\ldots,x_N\}$ 各样本依 $p(x\mid\theta)$ 独立抽取

**样本均值** $$\bar{x}=\frac{1}{N}\sum_{k=1}^{N}\boldsymbol{x}_k$$

**样本协方差矩阵** $$\boldsymbol{C}=\frac{1}{N-1}\sum_{k=1}^{N}(\boldsymbol{x}_k-\bar{x})(\boldsymbol{x}_k-\bar{x})^T$$

**最大似然参数估计**

**期望** $$\hat{\mu}=\frac{1}{N}\sum_{k=1}^{N}\boldsymbol{x}_k=\bar{\boldsymbol{x}}$$

**协方差矩阵** $$\hat{\Sigma}=\frac{1}{N}\sum_{k=1}^{N}(\boldsymbol{x}_k-\hat{\boldsymbol{\mu}})(\boldsymbol{x}_k-\hat{\boldsymbol{\mu}})^T$$

# PART4. 多元正态分布概率密度函数的最大似然估计

- 一般情况
- <u>各特征相互独立</u>

## 对于多变量正态分布 $\mu$、$\Sigma$ 均未知

$$\Sigma = \begin{bmatrix} \sigma_{11} & & & \\ & \sigma_{22} & & \\ & & \cdots\cdots & \\ & & & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} \sigma^{(1)2} & & & \\ & \sigma^{(2)2} & & \\ & & \cdots\cdots & \\ & & & \sigma^{(d)2} \end{bmatrix}$$

概率密度函数

$$p_X(\boldsymbol{x};\boldsymbol{\theta}) = p_X(\boldsymbol{x};\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

$$= \prod_{i=1}^{d}\frac{1}{\sqrt{2\pi}\sigma^{(i)}}\exp\left[-\frac{1}{2}\left(\frac{x^{(i)}-\mu^{(i)}}{\sigma^{(i)}}\right)^2\right] = \prod_{i=1}^{d} \boldsymbol{p}_{X^{(i)}}(x^{(i)};\mu^{(i)},\sigma^{(i)2})$$

$$\boldsymbol{p}_{X^{(i)}}(x^{(i)};\mu^{(i)},\sigma^{(i)2}) = \frac{1}{\sqrt{2\pi}\sigma^{(i)}}\exp\left[-\frac{1}{2}\left(\frac{x^{(i)}-\mu^{(i)}}{\sigma^{(i)}}\right)^2\right]$$

随机变量$\boldsymbol{X}^{(i)}$的边缘密度

$$\boldsymbol{p}_{X^{(i)}}(x^{(i)};\mu^{(i)},\sigma^{(i)2})=\frac{1}{\sqrt{2\pi}\sigma^{(i)}}\exp\left[-\frac{1}{2}\left(\frac{x^{(i)}-\mu^{(i)}}{\sigma^{(i)}}\right)^2\right]$$

随机向量$\boldsymbol{X}$的概率密度函数：$p_X(\boldsymbol{x};\mu,\Sigma)=\prod_{i=1}^{d}\boldsymbol{p}_{X^{(i)}}(x^{(i)};\mu^{(i)},\sigma^{(i)2})$

**边缘密度$p_{X^{(i)}}(x^{(i)};\mu^{(i)},\sigma^{(i)2})$最大似然参数估计**

$$\begin{cases}\textbf{期望} & \hat{\mu}^{(i)}=\dfrac{1}{N}\sum_{k=1}^{N}\boldsymbol{x}_k^{(i)}\\ \textbf{方差} & \widehat{\sigma^{(i)}}^{2}=\dfrac{1}{N}\sum_{k=1}^{N}\left(x_k^{(i)}-\hat{\mu}^{(i)}\right)^2\end{cases}$$

$$\Rightarrow\begin{cases}\textbf{期望向量}\mu & \hat{\boldsymbol{\mu}}=\left[\hat{\mu}^{(1)},\hat{\mu}^{(2)},\ldots,\hat{\mu}^{(d)}\right]^{T}\\ \textbf{协方差矩阵} & \hat{\Sigma}=diag\left(\widehat{\sigma^{(1)}}^{2},\widehat{\sigma^{(2)}}^{2},\ldots,\widehat{\sigma^{(d)}}^{2}\right)\end{cases}$$

**基于样本的两步贝叶斯决策**

[1] **利用有限规模训练样本** $\left\{\left(x_j, y_j\right), j=1,\ldots,N\right\}$,

**设计贝叶斯分类器. 估计**

$$\begin{cases} \text{先验概率} & P(\omega_i) \\ \text{条件概率密度} & p(x|\omega_i)\text{或条件概率}P(x|\omega_i) \end{cases}$$

$$i=1,2,\ldots,c$$

[2] $\begin{cases}\text{利用估计的}\hat{P}(\omega_i)\text{、}\hat{p}(x|\omega_i)\text{或}\hat{P}(x|\omega_i) \\ \text{对未知样本}x\text{进行判决}\end{cases}$

若观测样本$x$所在特征空间为连续的：

$$P(\omega_j \mid x) = \frac{\hat{P}(\omega_j)\hat{p}(x \mid \omega_j)}{p(x)} = \frac{\hat{P}(\omega_j)\hat{p}(x \mid \omega_j)}{\sum_{i=1}^{c}\hat{P}(\omega_i)\hat{p}(x \mid \omega_i)}$$

$$j = 1, 2, \ldots, c$$

若观测样本$x$所在特征空间为离散的：

$$P(\omega_j \mid x) = \frac{P(\omega_j)P(x \mid \omega_j)}{P(x)} = \frac{P(\omega_j)P(x \mid \omega_j)}{\sum_{i=1}^{c}P(\omega_i)P(x \mid \omega_i)}$$

$$j = 1, 2, \ldots, c$$

# 主要内容

1. 引言
2. 贝叶斯决策模型
    2.1 最小错误率贝叶斯决策
        连续情况；离散情况
    2.2 最小风险的贝叶斯决策
        连续情况；离散情况
3. 正态分布的概率密度函数及性质
4. 概率/概率密度函数估计
5. 朴素贝叶斯(Naive Bayes)分类

**朴素贝叶斯法**，基于贝叶斯公式、以及各特征条件独立的假设，实现贝叶斯决策。

连续特征空间：$p(\mathbf{x}\,|\,\omega_j)=\prod_{k=1}^{d}p(x^{(k)}\,|\,\omega_j)$

离散特征空间：$P(\mathbf{x}\,|\,\omega_j)=\prod_{k=1}^{d}P(x^{(k)}\,|\,\omega_j)$

# PART1. 离散特征空间的
# 朴素贝叶斯(Naive Bayes)分类
# —算法描述

**输入：**

(1)训练样本集$\left\{\left(\boldsymbol{x}_j, \mathrm{y}_j\right), \boldsymbol{j}=1,\ldots,\boldsymbol{N}\right\}$，

其中$\boldsymbol{x}_j=\left[x_j^{(1)},\ldots,x_j^{(\boldsymbol{d})}\right]^{\boldsymbol{T}}$

$x_j^{(\boldsymbol{k})}$是第$\boldsymbol{j}$个样本的第$\boldsymbol{k}$个特征；

并且$x_j^{(\boldsymbol{k})}\in\left\{a_{k1},a_{k2},\ldots,a_{kS_k}\right\},k\in\{1,2,\ldots,d\}$

$\mathrm{y}_j\in\left\{\omega_1,\omega_2,\ldots,\omega_C\right\}$

(2)待决策的观测样本$\boldsymbol{x}=\left[x^{(1)},\ldots,x^{(\boldsymbol{d})}\right]^{\boldsymbol{T}}$

**输出：**观测样本$\boldsymbol{x}$的类别

**实现步骤：**

(1)利用训练样本集估计先验概率及条件概率

## 估计方式1--最大似然估计法

$$\boldsymbol{X}=\left[X^{(1)},\ldots,X^{(d)}\right]^{\boldsymbol{T}}$$

$$\hat{\boldsymbol{P}}(\omega=\omega_i)=\frac{\sum_{j=1}^{N} I\left(y_j=\omega_i\right)}{\boldsymbol{N}},\quad i\in\{1,2,\ldots,C\}$$

$$\hat{\boldsymbol{P}}(X^{(k)}=a_{kl}|\omega_i)=\frac{\hat{\boldsymbol{P}}(X^{(k)}=a_{kl},\omega=\omega_i)}{\hat{\boldsymbol{P}}(\omega=\omega_i)}=\frac{\sum_{j=1}^{N} I\left(x_j^{(k)}=a_{kl},y_j=\omega_i\right)\Big/\boldsymbol{N}}{\sum_{j=1}^{N} I\left(y_j=\omega_i\right)\Big/\boldsymbol{N}}$$

$$=\frac{\sum_{j=1}^{N} I\left(x_j^{(k)}=a_{kl},y_j=\omega_i\right)}{\sum_{j=1}^{N} I\left(y_j=\omega_i\right)}\qquad \begin{cases} k\in\{1,2,\ldots,d\} \\ l\in\{1,2,\ldots,S_k\} \\ i\in\{1,2,\ldots,C\} \end{cases}$$

可能会导致某个估计结果为0，影响后续的后验概率计算，以及决策。

**实现步骤：**(1)利用训练样本集估计先验概率及条件边缘概率

**估计方式2--贝叶斯估计法**

**($\lambda$=1，称LAPLACE平滑; 0<$\lambda$<1,Lidstone平滑)**

MultinomialNB

$$\hat{P}(\omega=\omega_i)=\frac{\lambda+\sum_{j=1}^{N}I\left(y_j=\omega_i\right)}{C\lambda+N},\quad i\in\{1,2,\ldots,C\}$$

$$\hat{P}(X^{(k)}=a_{kl}|\omega_i)=\frac{\lambda+\sum_{j=1}^{N}I\left(x_j^{(k)}=a_{kl},y_j=\omega_i\right)}{S_k\lambda+\sum_{j=1}^{N}I\left(y_j=\omega_i\right)}\quad \begin{cases} k\in\{1,2,\ldots,d\} \\ l\in\{1,2,\ldots,S_k\} \\ i\in\{1,2,\ldots,C\} \end{cases}$$

**(多项朴素贝叶斯,Multinomial Naive Bayes)**

Multinomial Naive Bayes 基于各类多项分布假设(各特征只能取有限个离散值之一)，是面向文档分类中的两种经典朴素贝叶斯模型之一(另一种为Bernoulli Naive Bayes)。

(2)对于给定的观测样本$\boldsymbol{x}=\left[x_1,\ldots,x_d\right]^T$，计算

$$\hat{P}(\omega=\omega_j)\hat{P}(X=\boldsymbol{x}|\omega=\omega_j) \qquad j\in\{1,2,\ldots,C\}$$

$$\hat{P}(\omega=\omega_j)\hat{P}(X=\boldsymbol{x}|\omega=\omega_j)$$
$$=\hat{P}(\omega=\omega_j)\prod_{k=1}^{d}\hat{P}(X^{(k)}=x^{(k)}|\omega=\omega_j)$$

(3)确定观测样本$\boldsymbol{x}$的预测类别$y$

$$y=\arg\max_{\omega_j}\left[\hat{P}(\omega=\omega_j)\hat{P}(X=\boldsymbol{x}|\omega=\omega_j)\right]$$

试由表 的训练数据学习一个朴素贝叶斯分类器并确定 $x=(2,S)^T$ 类标记 $y$。

表中 $X^{(1)}$，$X^{(2)}$ 为特征，取值的集合分别为 $A_1=\{1,2,3\}$，$A_2=\{S,M,L\}$，

$Y$ 为类标记，$Y\in C=\{1,-1\}$

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X^{(1)}$ | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 |
| $X^{(2)}$ | $S$ | $M$ | $M$ | $S$ | $S$ | $S$ | $M$ | $M$ | $L$ | $L$ | $L$ | $M$ | $M$ | $L$ | $L$ |
| $Y$ | −1 | −1 | 1 | 1 | −1 | −1 | −1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | −1 |

先验概率估计$(\lambda=1$，类别数$=2)$：$\hat{P}(Y=c_i)=\dfrac{\lambda+\sum_{j=1}^{15} I(y_j=c_i)}{2\lambda+15}$ $\quad\{c_1=1,c_2=-1\}$

条件边缘概率估计 $\hat{P}\left(X^{(k)}=a_{kl}\mid Y=c_i\right)=\dfrac{\lambda+\sum_{j=1}^{15} I\left(x_j^{(k)}=a_{kl},y_j=c_i\right)}{s_k\lambda+\sum_{j=1}^{15} I(y_j=c_i)}$ $\quad l\in\{1,\dots,s_k\}$

$X^{(1)}\in A_1=\{1,2,3\},s_1=3 \qquad X^{(2)}\in A_2=\{S,M,L\},s_2=3$

## 解：首先进行先验概率、以及各类条件概率估计

**方式1：最大似然估计**

$$P(Y=1)=\frac{9}{15},\quad P(Y=-1)=\frac{6}{15}$$

$$P(X^{(1)}=1|Y=1)=\frac{2}{9},\quad P(X^{(1)}=2|Y=1)=\frac{3}{9},\quad P(X^{(1)}=3|Y=1)=\frac{4}{9}$$

$$P(X^{(2)}=S|Y=1)=\frac{1}{9},\quad P(X^{(2)}=M|Y=1)=\frac{4}{9},\quad P(X^{(2)}=L|Y=1)=\frac{4}{9}$$

$$P(X^{(1)}=1|Y=-1)=\frac{3}{6},\quad P(X^{(1)}=2|Y=-1)=\frac{2}{6},\quad P(X^{(1)}=3|Y=-1)=\frac{1}{6}$$

$$P(X^{(2)}=S|Y=-1)=\frac{3}{6},\quad P(X^{(2)}=M|Y=-1)=\frac{2}{6},\quad P(X^{(2)}=L|Y=-1)=\frac{1}{6}$$

基于上述概率信息，对观测样本x进行类别决策：

对于给定的 $x=(2,S)^{\mathrm{T}}$ 计算：

$$P(Y=1)P(X^{(1)}=2|Y=1)P(X^{(2)}=S|Y=1)=\frac{9}{15}\cdot\frac{3}{9}\cdot\frac{1}{9}=\frac{1}{45}$$

$$P(Y=-1)P(X^{(1)}=2|Y=-1)P(X^{(2)}=S|Y=-1)=\frac{6}{15}\cdot\frac{2}{6}\cdot\frac{3}{6}=\frac{1}{15}$$

因为 $P(Y=-1)P(X^{(1)}=2|Y=-1)P(X^{(2)}=S|Y=-1)$ 最大，所以 $y=-1$ .

解：首先进行先验概率、以及各类条件概率估计

方式2：LAPLACE平滑

$$P(Y=1)=\frac{10}{17},\quad P(Y=-1)=\frac{7}{17}$$

$$P(X^{(1)}=1|Y=1)=\frac{3}{12},\quad P(X^{(1)}=2|Y=1)=\frac{4}{12},\quad P(X^{(1)}=3|Y=1)=\frac{5}{12}$$

$$P(X^{(2)}=S|Y=1)=\frac{2}{12},\quad P(X^{(2)}=M|Y=1)=\frac{5}{12},\quad P(X^{(2)}=L|Y=1)=\frac{5}{12}$$

$$P(X^{(1)}=1|Y=-1)=\frac{4}{9},\quad P(X^{(1)}=2|Y=-1)=\frac{3}{9},\quad P(X^{(1)}=3|Y=-1)=\frac{2}{9}$$

$$P(X^{(2)}=S|Y=-1)=\frac{4}{9},\quad P(X^{(2)}=M|Y=-1)=\frac{3}{9},\quad P(X^{(2)}=L|Y=-1)=\frac{2}{9}$$

基于上述概率信息，对观测样本X进行类别决策：

对于给定的 $x=(2,S)^{\mathrm{T}}$ 计算：

$$P(Y=1)P(X^{(1)}=2|Y=1)P(X^{(2)}=S|Y=1)=\frac{10}{17}\cdot\frac{4}{12}\cdot\frac{2}{12}=\frac{5}{153}=0.0327$$

$$P(Y=-1)P(X^{(1)}=2|Y=-1)P(X^{(2)}=S|Y=-1)=\frac{7}{17}\cdot\frac{3}{9}\cdot\frac{4}{9}=\frac{28}{459}=0.0610$$

由于 $P(Y=-1)P(X^{(1)}=2|Y=-1)P(X^{(2)}=S|Y=-1)$ 最大，所以 $y=-1$ .

# PART2. 连续特征空间的
# 朴素贝叶斯(Naive Bayes)分类
# ——算法描述

**输入：**

(1)训练样本集$\boldsymbol{D}=\left\{\left(\boldsymbol{x}_j, y_j\right), \boldsymbol{j}=1,\ldots,\boldsymbol{N}\right\}$,

其中$\boldsymbol{x}_j=\left[x_j^{(1)},\ldots,x_j^{(\boldsymbol{d})}\right]^{\boldsymbol{T}}$

$x_j^{(k)}$是第$\boldsymbol{j}$个样本的第$\boldsymbol{k}$个特征；

$k\in\{1,2,\ldots,d\}$,　　$y_j\in\{\omega_1,\omega_2,\ldots,\omega_C\}$

(2)待决策的观测样本$\boldsymbol{x}=\left[x^{(1)},\ldots,x^{(\boldsymbol{d})}\right]^{\boldsymbol{T}}$

**输出**：观测样本$\boldsymbol{x}$的类别

**实现步骤：**

(1)利用训练样本集估计**先验概率**及**条件概率密度函数**

估计**先验概率**：$\hat{P}(\omega=\omega_i)=\frac{\sum_{j=1}^{N} I\left(y_j=\omega_i\right)}{N}, \quad i\in\{1,2,\ldots,C\}$

估计**条件密度**：$\hat{p}(\boldsymbol{X}=\boldsymbol{x}|\omega=\omega_i)=\prod_{k=1}^{d}\hat{p}(X^{(k)}=x^{(k)}|\omega=\omega_i)$

分别利用训练样本集内第$\omega_i$类所有样本的第$k$个特征取值确定估计的**条件边缘密度**:$\hat{p}(X^{(k)}=x^{(k)}|\omega=\omega_i)$

$k\in\{1,2,\ldots,d\}$；$i\in\{1,2,\ldots,C\}$

特别地，若为**Gaussian Naive Bayes**分类模型 **GaussianNB**

$$p(\boldsymbol{X}=\boldsymbol{x} \mid \omega=\omega_i)=\prod_{k=1}^{d} p(X^{(k)}=x^{(k)} \mid \omega=\omega_i)$$

并且 $p(X^{(k)}=x^{(k)} \mid \omega=\omega_i)=\dfrac{1}{\sqrt{2\pi}\sigma_i^{(k)}}\exp\left[-\dfrac{1}{2}\left(\dfrac{x^{(k)}-\mu_i^{(k)}}{\sigma_i^{(k)}}\right)^2\right]$

所以 $\hat{p}(X^{(k)}=x^{(k)} \mid \omega=\omega_i)=\dfrac{1}{\sqrt{2\pi}\hat{\sigma}_i^{(k)}}\exp\left[-\dfrac{1}{2}\left(\dfrac{x^{(k)}-\hat{\mu}_i^{(k)}}{\hat{\sigma}_i^{(k)}}\right)^2\right]$

**期望** $\hat{\mu}_i^{(k)}=\dfrac{1}{N_i}\sum_{\substack{(\boldsymbol{x}_j,\mathrm{y}_j)\in \boldsymbol{D} \\ \text{并且}\,\mathrm{y}_j=\omega_i}} \boldsymbol{x}_j^{(k)}$

**方差** $\widehat{\sigma_i^{(k)}}^2=\dfrac{1}{N_i}\sum_{\substack{(\boldsymbol{x}_j,\mathrm{y}_j)\in \boldsymbol{D} \\ \text{并且}\,\mathrm{y}_j=\omega_i}} \left(\boldsymbol{x}_j^{(k)}-\hat{\mu}_i^{(k)}\right)^2$

$k \in \{1,2,\ldots,d\}$

$i \in \{1,2,\ldots,C\}$

(2)对于给定的观测样本$\boldsymbol{x}=\left[x_1,\ldots,x_{\boldsymbol{d}}\right]^{\boldsymbol{T}}$，计算

$$\widehat{\boldsymbol{P}}(\omega=\omega_j)\widehat{\boldsymbol{p}}(\boldsymbol{X}=\boldsymbol{x}|\omega=\omega_j) \qquad j\in\{1,2,\ldots,C\}$$

$$\widehat{\boldsymbol{P}}(\omega=\omega_j)\widehat{\boldsymbol{p}}(\boldsymbol{X}=\boldsymbol{x}|\omega=\omega_j)$$

$$=\widehat{\boldsymbol{P}}(\omega=\omega_j)\prod_{k=1}^{d}\widehat{p}(X^{(k)}=x^{(k)}|\omega=\omega_i)$$

(3)确定观测样本$\boldsymbol{x}$的预测类别$y$

$$y=\arg\max_{\omega_j}\left[\widehat{\boldsymbol{P}}(\omega=\omega_j)\widehat{\boldsymbol{p}}(\boldsymbol{X}=\boldsymbol{x}|\omega=\omega_j)\right]$$

# 思考题

河北师范大学软件学院 Software College of Hebei Normal University

1. 面向两类别/多类别分类问题的两种贝叶斯分类模型决策规则是什么？

(1)基于最小错误率的贝叶斯分类，决策规则；

(2)基于最小风险的贝叶斯分类，决策规则。

2. 正确写出单变量/多元正态分布的概率密度函数表达式。

3. (1)若d维随机向量正态分布，并且N个观测样本按照正态分布独立抽取得到，请写出N个观测样本组成的样本集的似然值；并写出概率密度函数的各参数最大似然估计结果.

(2)若各特正确写出征相互独立，请写出正态分布概率密度函数的最大似然估计结果

# 思考题

河北师范大学软件学院 Software College of Hebei Normal University

4. 掌握如下算法：

(1) 离散特征空间，基于最小错误率的朴素贝叶斯分类模型；

(2) 连续特征空间，各类别条件概率密度函数正态分布情况下，基于最小错误率的朴素贝叶斯分类模型。