第3章:贝叶斯决策理论

Chapter 3: Bayesian Decision Theory

张永飞

2023年9月19日

内容提要

- 概率统计理论基础
- 贝叶斯决策理论
  - 基本概念
  - 最小错误率贝叶斯决策
  - 最小风险贝叶斯决策
  - 朴素贝叶斯决策

概率统计理论基础

- **概率 (Probability)**

  对随机事件发生可能性大小的度量

- **联合概率 (Joint Probability)**

  A和B共同发生的概率，称事件A和B的联合概率，记作P(A, B)或P(A∩B)

- **条件概率 (Conditional Probability)**

  事件A已发生的条件下，事件B发生的概率，记作P(B|A)

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

概率统计理论基础

- **独立事件 (Independent Events)**

  事件A(或B)是否发生对事件B(或A)的发生概率没有影响，则称A和B为相互独立事件

- **条件独立 (Conditional Independence)**

  在给定C的条件下，若事件A和B满足：

$$P(A, B|C) = P(A|C) \cdot P(B|C)$$

或：

$$P(A|B, C) = P(A|C)$$

则称在给定C的情况下A和B独立

概率统计理论基础

- **乘法公式(Multiplication Theorem/Formula)**
  - **设A，B为任意事件**

$$\begin{aligned} P(A,B) &= P(A|B)\cdot P(B) \\ &= P(B|A)\cdot P(A) \end{aligned}$$

  - **推广到n个事件的情况：**

$$P(A_1A_2\ldots A_n) = P(A_1)\cdot P(A_2|A_1)\cdot P(A_3|A_1A_2)\cdot\ldots\cdot P(A_n|A_1A_2\ldots A_{n-1})$$

$$= \prod_{i=1}^{n} P(A_i|A_1A_2\ldots A_{i-1})$$

概率统计理论基础

- **全概率公式 (Law of Total Probability)**

  - 设A1,A2,⋯,An两两互不相容，且

$$B \subset A_1 + A_2 + \ldots + A_n$$

即B的发生总是与A1, A2,⋯,An之一同时发生，则对于事件B，有

$$P(B) = \sum_{k=1}^{n} P(A_k) P(B \mid A_k)$$

知因求果

概率统计理论基础

- **贝叶斯公式 (Bayes' Theorem/Formula)** 知果求因
  - 设$A_1,A_2,\ldots,A_n$两两互不相容，且 $B \subset A_1 + A_2 + \ldots + A_n$

即B的发生总是与$A_1,A_2,\ldots,A_n$之一同时发生，则在B已经发生的条件下，A_k的条件概率为

乘法公式

$$\boxed{P(A_k \mid B) = \frac{\boxed{P(A_k B)}}{P(B)}} = \frac{\boxed{P(A_k)P(B \mid A_k)}}{\boxed{\sum_{i=1}^{n} P(A_i)P(B \mid A_i)}}$$

乘法公式

全概率公式

其中$P(A_k)$为先验概率； $P(A_k|B)$ 为后验概率

贝叶斯公式给出了“结果”事件B已经发生的条件下，“原因”事件A的条件概率，对结果的任何观测都将增加我们对原因事件A的真正分布的知识

贝叶斯决策理论

- 统计决策理论
  - 是机器学习/模式分类问题的基本理论之一
  - 用概率统计的观点和方法（基于贝叶斯公式）来解决模式识别问题
- 贝叶斯决策理论
  - 是统计决策理论中的一个基本方法和基础
  - 是“最优分类器”：使平均错误率最小
  - 最小错误率贝叶斯决策
  - 最小风险贝叶斯决策
  - 朴素贝叶斯决策

基本概念

- **样本（sample）** $\mathbf{x} \in R^d$
- **类别/状态（class/state）** ω_i
- **先验概率（a priori probability or prior）** $P(\omega_i)$
- **样本分布密度（sample distribution density）** $p(\mathbf{x})$
- **类条件概率密度（class-conditional probability density）**

$$P(\mathbf{x}|\omega_i)$$

基本概念

- **后验概率（a posteriori probability or posterior ）**

$$P(\omega_i|\mathbf{x})$$

- **错误概率（probability of error）：**

$$P(e|\mathbf{x}) = \begin{cases} P(w_2|\mathbf{x}) \text{ if } \mathbf{x} \text{ is assigned to } w_1 \\ P(w_1|\mathbf{x}) \text{ if } \mathbf{x} \text{ is assigned to } w_2 \end{cases}$$

- **平均错误率（average probability of error）**

$$P(e) = \int P(e|\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

- **正确率（probability of correctness）** $P(c)$

基本概念

- 贝叶斯公式

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

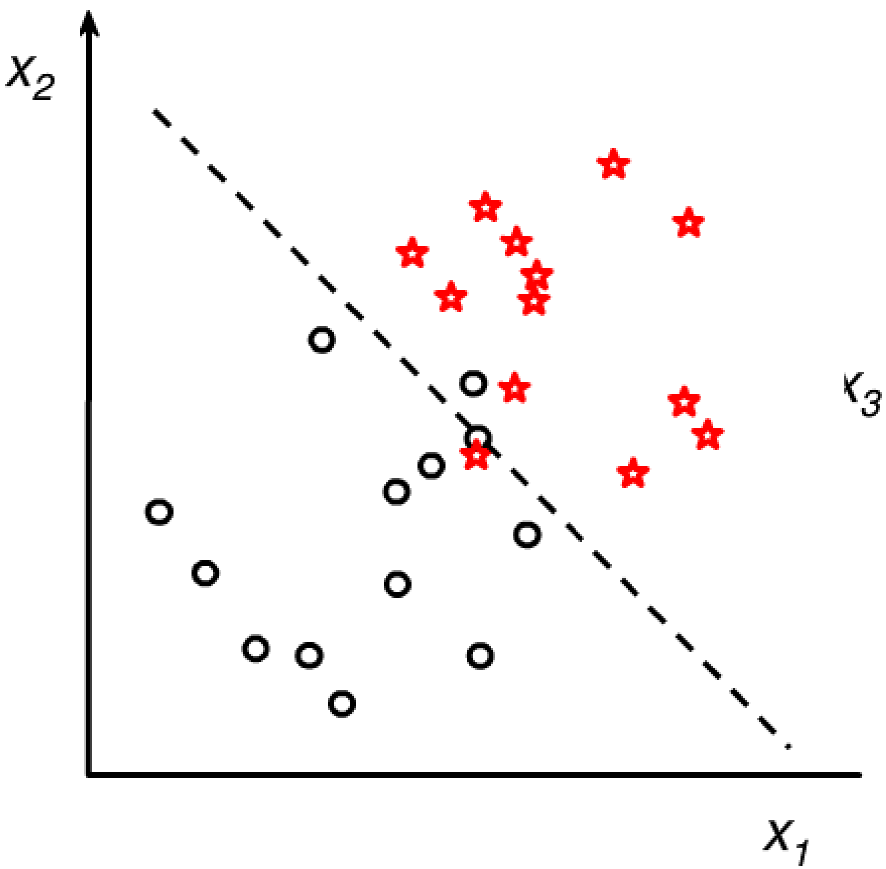
- 先验概率vs.后验概率
  - 先验概率：由以往历史数据得到的概率
  - 后验概率：利用最新输入数据对先验概率加以修正后的概率
  - 示例：性别比例

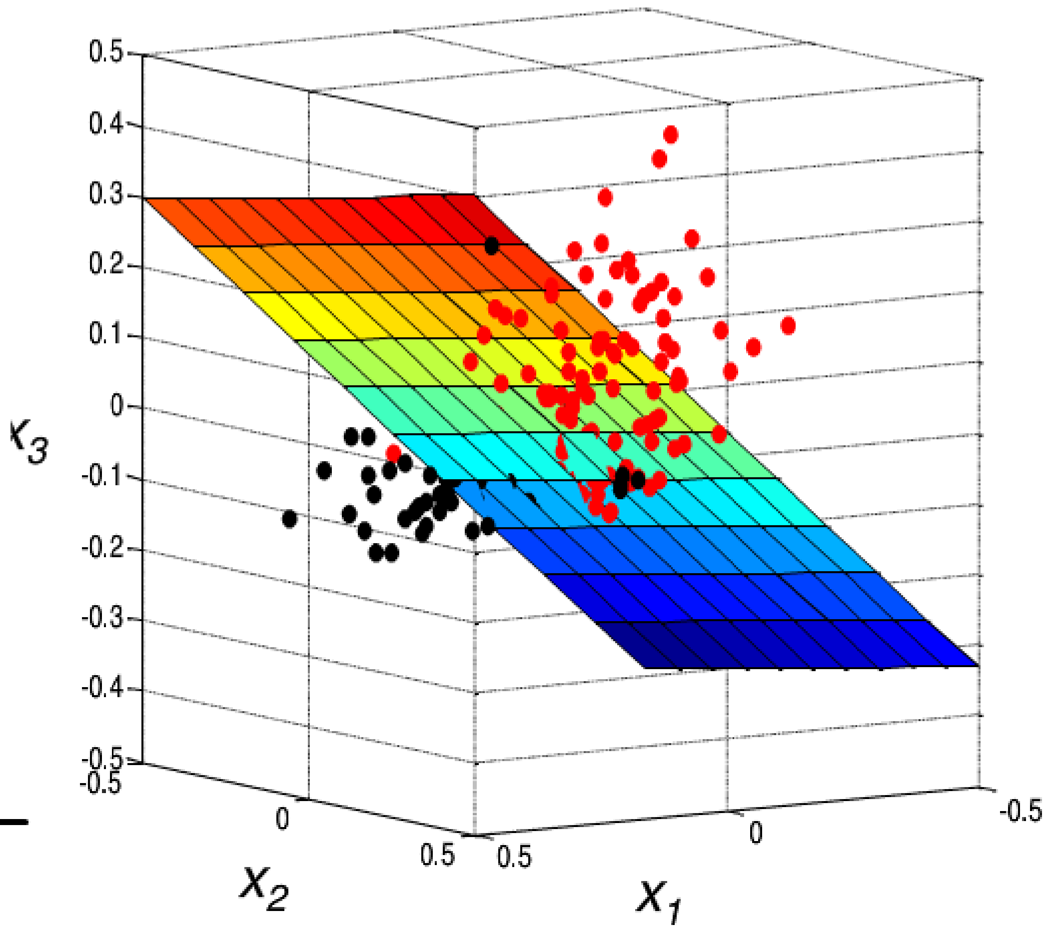
分类问题描述

- **给定**：m个类、已知类别属性的训练样本和未知类别属性的输入数据
- **目标**：确定每一个输入数据的类别属性
- **两个阶段**：
  - 建模/学习：基于训练样本 学习 分类规则
  - 分类/测试：对于出入数据 应用 分类规则

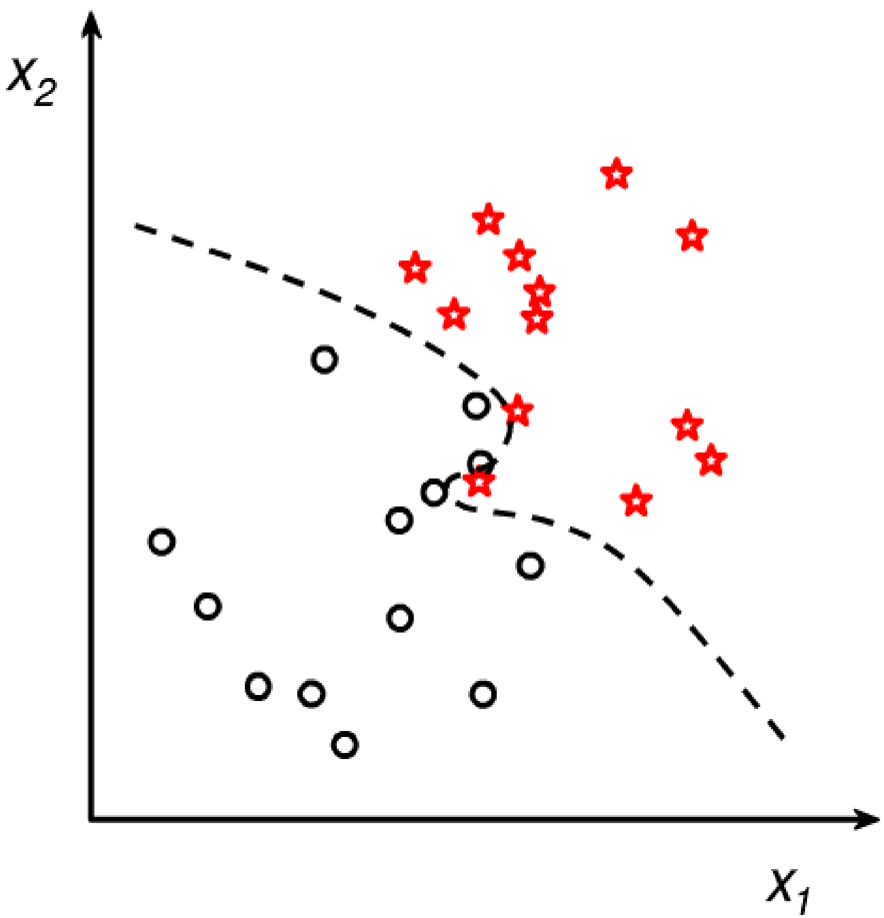
线性决策边界

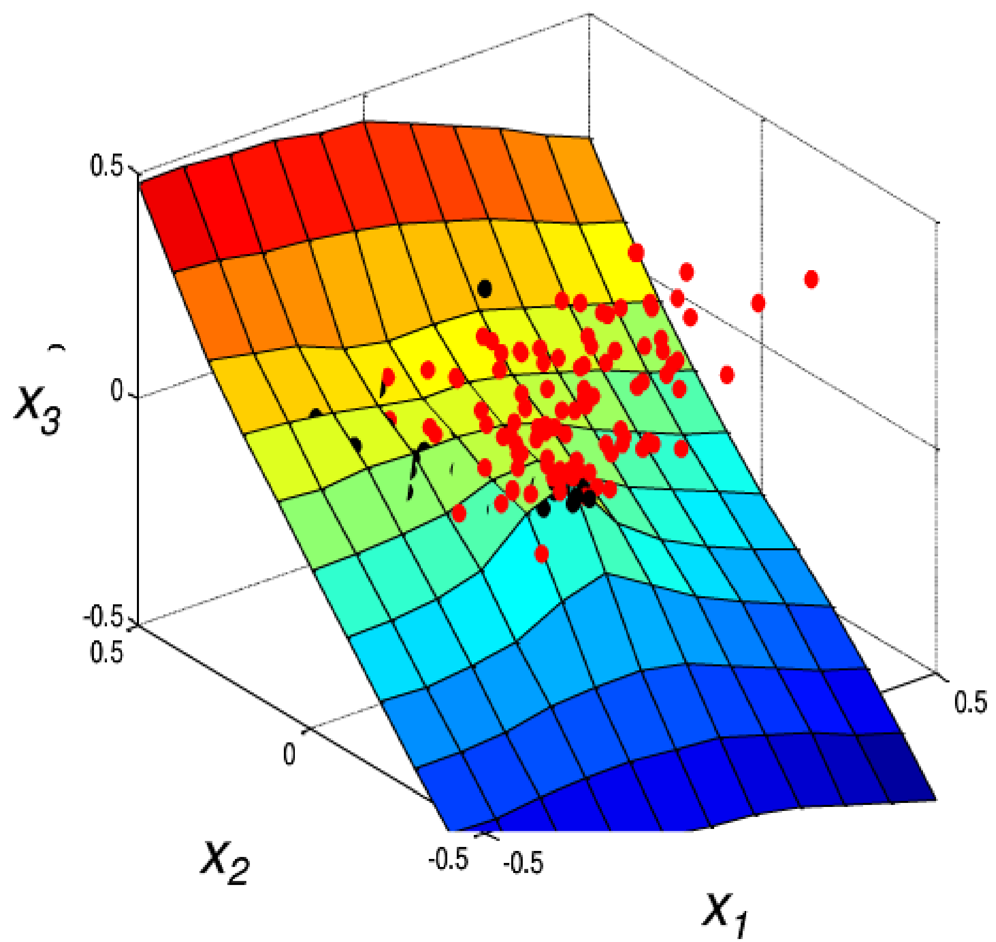
hyperplane





非线性决策边界





贝叶斯决策

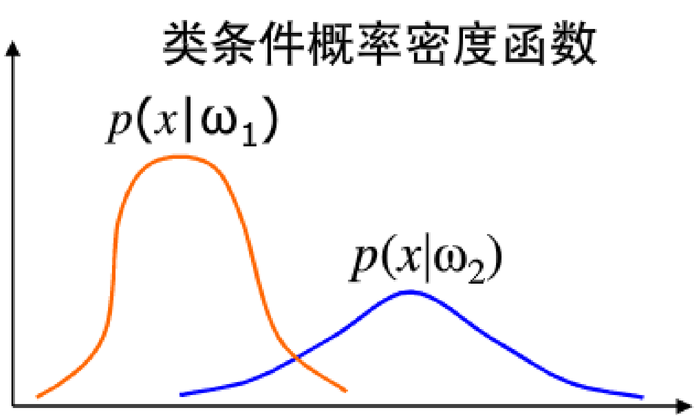
- **已知条件**
  - 类别数一定（决策论中把类别也称为状态）：
    - ω_{i}, $i = 1,2,\ldots,c$
  - 已知各类在这d维特征空间的统计分布
    - 各类别ω_i $i = 1,2,\ldots,c$的先验概率 $P(\omega_i)$， $i = 1,2,\ldots,c$
    - 类条件概率密度函数 $P(\mathrm{x}|\omega_i)$， $i = 1,2,\ldots,c$
- **决策**：根据贝叶斯公式计算后验概率 $P(\omega_i|\mathrm{x})$，基于最大后验概率进行判决

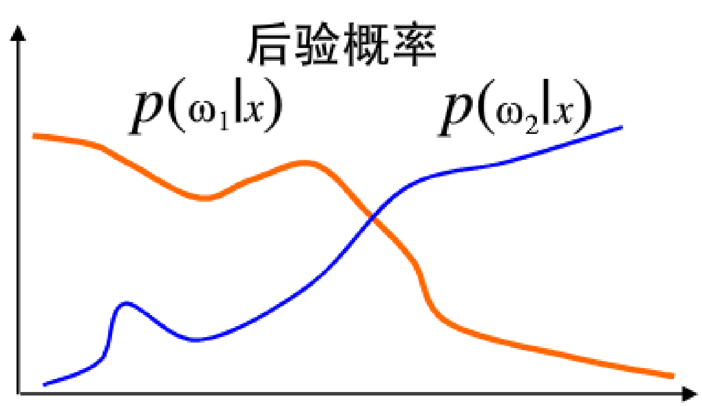
贝叶斯决策

- 以**最大后验概率**为判决函数

$$x \in \omega_k \quad iff \quad k = \arg\max_i \{P(\omega_i|\mathbf{x})\}$$

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





最小错误率贝叶斯决策

- **目标** $\min \quad P(e) = \int P(e|x)p(x)dx$

- 因为 $P(e|\mathbf{x}) \geq 0, \; p(x) \geq 0$

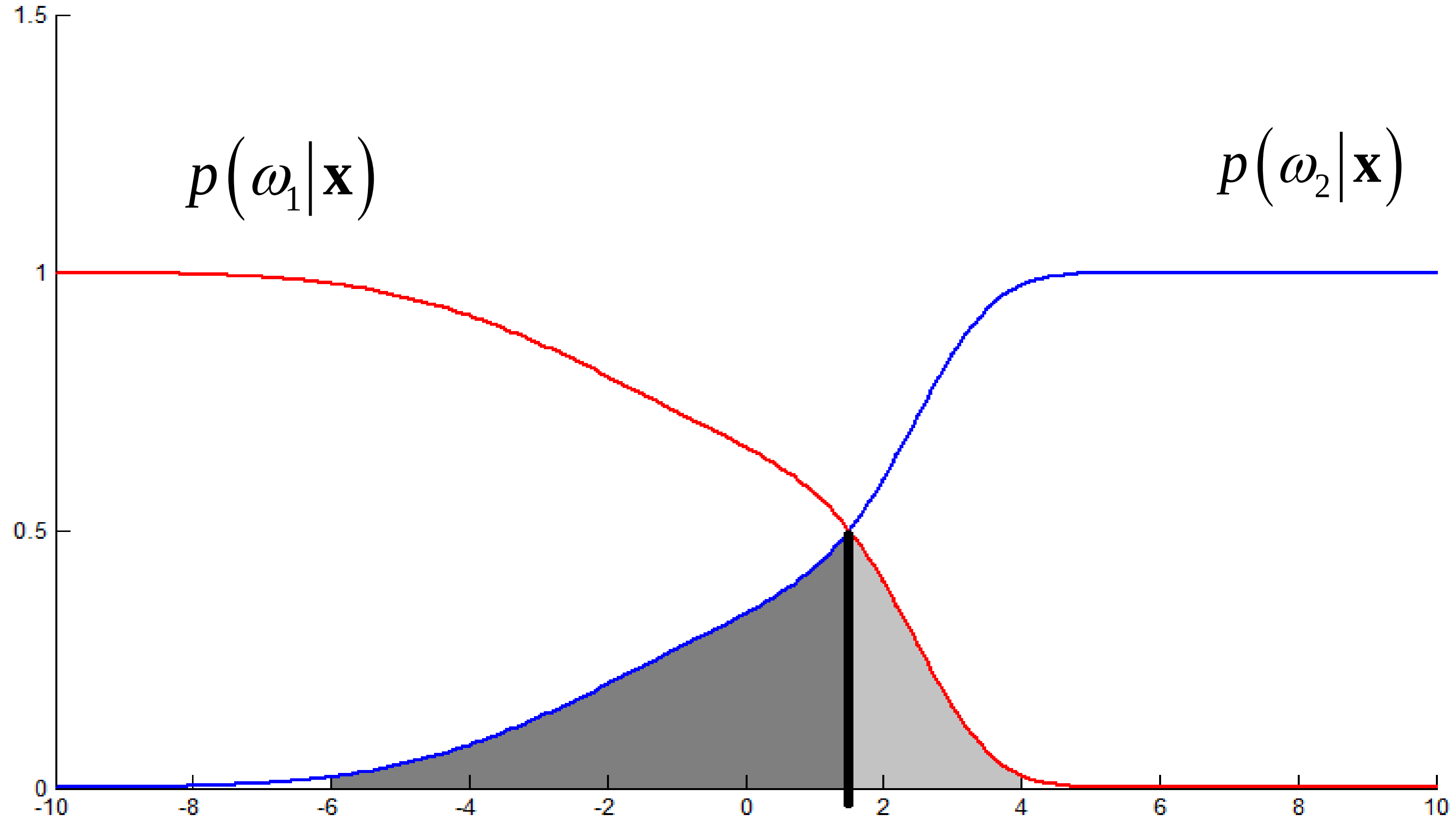
$\Longrightarrow$ $\min P(e|\mathbf{x}) \text{ for all } x$

而 $P(e|\mathbf{x}) = \begin{cases} P(w_2|\mathbf{x}), \; \text{if } P(w_1|\mathbf{x}) > P(w_2|\mathbf{x}) \\ P(w_1|\mathbf{x}), \; \text{if } P(w_2|\mathbf{x}) > P(w_1|\mathbf{x}) \end{cases}$

$$\text{if } P(w_1|\mathbf{x}) \gtrless P(w_2|\mathbf{x}) \text{ assign } \begin{matrix} x \in w_1 \\ x \in w_2 \end{matrix}$$

$$P(w|\mathbf{x}) = \max_{j=1,\ldots,c} P(w_j|\mathbf{x})$$

最小错误率贝叶斯决策



最小错误率贝叶斯决策

- 等价表达形式
  - (1) $if\ \ P(\omega_i|\mathbf{x}) = \max_{j=1,\dots,c} P(\omega_j|\mathbf{x}) \quad then\ x \in \omega_i$
  - (2) $if\ \ P(\mathbf{x}|\omega_i)P(\omega_i) = \max_{j=1,\dots,c} P(\mathbf{x}|\omega_i)P(\omega_i) \quad then\ x \in \omega_i$
  - (3) $if\ \ l(x) = \dfrac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)} \begin{matrix} > \\ < \end{matrix} \dfrac{P(\omega_2)}{P(\omega_1)} \quad then\ x \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$
  - (4) 定义 $h(x) = \ln[l(x)] = \ln P(\mathbf{x}|\omega_1) - \ln P(\mathbf{x}|\omega_2)$

$$if\ \ h(x) \begin{matrix} > \\ < \end{matrix} \ln\left(\frac{P(\omega_2)}{P(\omega_1)}\right), \quad then\ x \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

其中,$l(x)$为似然比;$h(x)$为对数似然比;$\dfrac{P(\omega_2)}{P(\omega_1)}$ 为似然比阈值

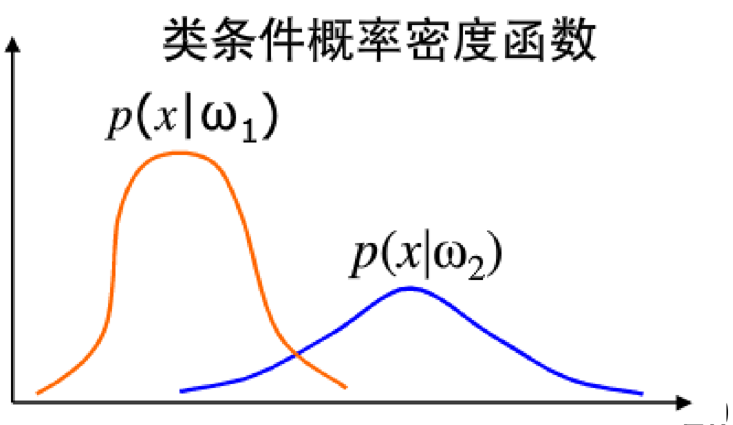
示例-最小错误率贝叶斯决策

- **细胞分类诊断：**
  - 假设在某个局部区域细胞中正常(ω_1)和异常(ω_2)两类的先验概率分别为：
  - 正常状态　　$P(\omega_1)=0.9$
  - 异常状态　　$P(\omega_2)=0.1$

  现有一待识别细胞，其观察值为x，从类条件概率密度曲线上查得

  - $p(x|\omega_1)=0.2$
  - $p(x|\omega_2)=0.4$

  试对该细胞x进行分类



示例-最小错误率贝叶斯决策

解：利用贝叶斯公式计算ω_1和ω_2的后验概率

$$P(w_1|x) = \frac{p(x|w_1)P(w_1)}{\sum_{j=1}^{2} p(x|w_j)P(w_j)} = \frac{0.2\times0.9}{0.2\times0.9+0.4\times0.1} = 0.818$$

$$P(w_2|x) = 1 - P(w_1|x) = 0.182$$

根据贝叶斯决策规则有

$$P(w_1|x) = 0.818 > P(w_2|x) = 0.182$$

因此，把x归类于正常细胞

最小错误率贝叶斯决策的问题

- 决策的风险
  - 不同的决策具有不同的风险或损失
  - 医疗诊断为例：
    - 没病判为有病：精神负担、可进一步检查，损失不大
    - 有病判为没病：贻误病情，后果严重
- 最小错误率贝叶斯决策以错误率最小为准则，未考虑决策的风险

最小风险贝叶斯决策

- **损失函数**：对于特定的x采取决策 α_i 的期望损失：$\lambda(\alpha_i, \omega_j)$

- **条件期望损失**：

$$R(\alpha_i|\mathbf{x}) = E[\lambda(\alpha_i, \omega_j)] = \sum_{j=1}^{c} \lambda(\alpha_i, \omega_j) P(\omega_j|\mathbf{x}), i = 1, 2, \ldots, a$$

- **期望风险**：对所有可能的x采取决策 $\alpha(x)$ 所造成的期望损失之和

$$R(\alpha) = \int R(\alpha(x) \mid x) p(x) \mathrm{d}x$$

最小风险贝叶斯决策

- **目标：期望风险最小**

$$\min \quad R(\alpha)=\int R(\alpha(x)\,|\,x)\,p(x)\mathrm{d}x$$

若对每一个决策，都使其条件风险 $R\left(\alpha_i \middle| \mathbf{x}\right)$ 最小，则对所有 x 做出决策时，其期望风险 R 也最小

- **决策：**

如果 $R\left(\alpha_k \middle| \mathbf{x}\right)=\min\limits_{i=1,2\ldots,a} R\left(\alpha_i \middle| \mathbf{x}\right)$ **，则** $\alpha=\alpha_k$

最小风险贝叶斯决策

- **算法步骤:**
  - 已知先验概率P(ω_i)， i= 1,2,…,c ，类条件概率p(x| ω_i)，i= 1,2,…,c ，以及待分类输入数据x
  - 根据贝叶斯公式计算后验概率

$$P\left(\omega_i \middle| \mathbf{x}\right) = \frac{P\left(\mathbf{x} \middle| \omega_i\right) P\left(\omega_i\right)}{\sum_{i=1}^{c} P\left(\mathbf{x} \middle| \omega_i\right) P\left(\omega_i\right)}$$

  - 利用后验概率与损失函数，计算条件风险

$$R\left(\alpha_i \middle| \mathbf{x}\right) = \sum_{j=1}^{c} \lambda(\alpha_i, \omega_j) P\left(\omega_j \middle| \mathbf{x}\right), i = 1, 2, \ldots, a$$

  - 决策:

$$R\left(\alpha_k \middle| \mathbf{x}\right) = \min_{i=1,2\ldots,a} R\left(\alpha_i \middle| \mathbf{x}\right)$$

示例-最小风险贝叶斯决策

- **细胞分类诊断**
  - 决策表

损失 状态 / 决策	ω_1	ω_2
α_1	0	6
α_2	1	0

  - 即 $\lambda_{11}=0;\ \lambda_{12}=6;\ \lambda_{21}=1;\ \lambda_{22}=0$

示例-最小风险贝叶斯决策

- **解:**
  - **损失函数** $\lambda_{11}=0;\ \lambda_{12}=6;\ \lambda_{21}=1;\ \lambda_{22}=0$
  - **后验概率:** $P(w_1|x)=0.818\quad P(w_2|x)=0.182$
  - **计算条件风险:** $R(\alpha_1|x)=\sum_{j=1}^{2}\lambda_{1j}P(w_j|x)=\lambda_{12}P(w_2|x)=1.092$

    $R(\alpha_2|x)=\lambda_{21}P(w_1|x)=0.818$
  - **决策：由于** $R(\alpha_1|x)>R(\alpha_2|x)$

    因此把x决策/归类为异常细胞

两种贝叶斯决策的关系

- **最小错误率贝叶斯决策** $\omega_k = \arg\max_{j=1,\ldots,c} P(\omega_j|\mathbf{x})$
- **最小风险贝叶斯决策**

$$\omega_k = \arg\min_i R(\alpha_i|\mathbf{x}) = \arg\min_i \sum_{j=1}^{c} \lambda(\alpha_i,\omega_j) P\left(\omega_j|\mathbf{x}\right)$$

设损失函数为 $\lambda(\alpha_i,\omega_j) = \begin{cases} 0, i=j \\ 1, i \neq j \end{cases}, i, j = 1, 2, \ldots, c$

则: $\omega_k = \arg\min_i \sum_{\substack{j=1 \\ j \neq i}}^{c} P\left(\omega_j|\mathbf{x}\right)$

最小化 $j! = i$时的后验概率 即 最大化$j = i$时的后验概率！

因此，最小错误率贝叶斯决策就是在0-1损失函数条件下的最小风险贝叶斯决策

朴素贝叶斯决策(Naïve Bayes)

- **贝叶斯决策的问题:**
  - 类条件概率 $P(\mathbf{x}|\omega_i)$ 是所有属性上的联合概率，难以从有限的训练样本直接估计得到
- **解决方法：朴素贝叶斯决策**
  - 属性条件独立性假设：对于已知类别，假设所有属性相互独立；即假设各属性独立地对分类结果发生影响，即

$$P(\mathbf{x}\,|\,\omega) = P(x_1 x_2, \ldots, x_i, \ldots x_d\,|\,\omega) = \prod_{i=1}^{d} P(x_i\,|\,\omega)$$

  - 好处：降低样本集大小需求；降低复杂度

示例-朴素贝叶斯决策(Naïve Bayes)

- 西瓜分类(见教材P151-154)

表 4.3 西瓜数据集 3.0

  - 训练样本

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	0.460	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
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	0.360	0.370	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

  - 测试数据

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

朴素贝叶斯决策(Naïve Bayes)

- **贝叶斯公式+属性独立性条件**

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

- **朴素贝叶斯决策**

$$\omega_k=\arg\max_j P(\omega_j)\prod_{i=1}^{d}P(x_i|\omega_j)$$

朴素贝叶斯方法的训练过程：基于训练样本数据D 来估计类先验概率 P(ω_i)，并为每个属性估计条件概率P(x|ω_i)，也就是它们在训练数据上的频率，然后再用上式分类新样本数据

朴素贝叶斯决策(Naïve Bayes)

- **先验概率估计**:

$$P(\omega_j) = \frac{|D_{\omega_j}|}{|D|} \qquad P(\omega_j) = \frac{|D_{\omega_j}|+1}{|D|+N}$$

- **类条件概率估计-离散属性**

$$P(x_i \mid \omega_j) = \frac{|D_{\omega_j,x_i}|}{|D_{\omega_j}|} \qquad P(x_i \mid \omega_j) = \frac{|D_{\omega_j,x_i}|+1}{|D_{\omega_j}|+N_i}$$

- **类条件概率估计-连续属性**
  - 假设 $P(x_i \mid \omega_j) \sim N(\mu_{\omega_j,i}, \sigma^2_{\omega_j,i})$

概率密度估计！

$$P(x_i \mid \omega_j) = \frac{1}{\sqrt{2p}\sigma_{\omega_j,i}} exp\left(- \frac{(x_i - \mu_{\omega_j,i})^2}{2\sigma^2_{\omega_j,i}} \right)$$

示例-朴素贝叶斯决策(Naïve Bayes)

- 西瓜分类(见教材P151-154)

表 4.3 西瓜数据集 3.0

- 训练样本

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	0.460	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
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	0.360	0.370	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

- 测试数据

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

示例-朴素贝叶斯决策

- 西瓜分类(见教材P151-154)
  - 计算先验概率： $P(\text{好瓜}=\text{是})=\frac{8}{17}\approx 0.471$， $P(\text{好瓜}=\text{否})=\frac{9}{17}\approx 0.529$.
  - 计算类条件概率：

$$P_{\text{青绿}|\text{是}}=P(\text{色泽}=\text{青绿}\mid\text{好瓜}=\text{是})=\frac{3}{8}=0.375,$$

$$P_{\text{青绿}|\text{否}}=P(\text{色泽}=\text{青绿}\mid\text{好瓜}=\text{否})=\frac{3}{9}\approx 0.333,$$

$$P_{\text{蜷缩}|\text{是}}=P(\text{根蒂}=\text{蜷缩}\mid\text{好瓜}=\text{是})=\frac{5}{8}=0.375,$$

……

  - 计算后验概率：

$$P(\text{好瓜}=\text{是})\times P_{\text{青绿}|\text{是}}\times P_{\text{蜷缩}|\text{是}}\times P_{\text{浊响}|\text{是}}\times P_{\text{清晰}|\text{是}}\times P_{\text{凹陷}|\text{是}}$$
$$\times P_{\text{硬滑}|\text{是}}\times p_{\text{密度: 0.697}|\text{是}}\times p_{\text{含糖: 0.460}|\text{是}}\approx 0.038,$$

$$\omega_k=\arg\max_j P(\omega_j)\prod_{i=1}^{d}P(x_i\mid\omega_j)$$

$$P(\text{好瓜}=\text{否})\times P_{\text{青绿}|\text{否}}\times P_{\text{蜷缩}|\text{否}}\times P_{\text{浊响}|\text{否}}\times P_{\text{清晰}|\text{否}}\times P_{\text{凹陷}|\text{否}}$$
$$\times P_{\text{硬滑}|\text{否}}\times p_{\text{密度: 0.697}|\text{否}}\times p_{\text{含糖: 0.460}|\text{否}}\approx 6.80\times 10^{-5}.$$

  - 决策： $0.038>6.80*10^{-5}$ 所以：好瓜

贝叶斯决策小结

- **前提假设**：已知类条件概率密度 $P(x|\omega_i)$ 和先验概率 $P(\omega_i)$，计算后验概率$P(\omega_i|x)$进行决策
- **实际问题**：只有一定数目的样本；类条件概率密度 $P(x|\omega_i)$ 和先验概率 $P(\omega_i)$并不一定知道，且往往不好求
- **解决方法**：基于样本的两步贝叶斯决策
  - 概率密度估计：先根据样本数据估计 $P(x|\omega_i)$ 和 $P(\omega_i)$
  - 贝叶斯决策：再根据估计的概率密度进行贝叶斯决策

概率密度函数估计

Probability Density Estimation

概率密度估计

- **问题与任务**
  - 根据样本数据估计类条件概率密度 $P(x|\omega_i)$ 和 先验概率 $P(\omega_i)$

- **方法分类**
  - 参数化方法 Parametric (Density) Methods*
  - 非参数化方法 Nonparametric (Density) Methods

参考：《模式识别》-边肇祺张学工P65-72

参数化方法*

- **问题：**
  - **前提**：已知概率密度函数的形式，只是其中几个参数未知（可以写成某些参数的函数，如典型分布）
  - **目标**：依据样本估计这些未知参数的值
- **典型方法**
  - 最大似然估计*
  - 贝叶斯估计

非参数化方法

- **问题：**
  - **前提**：概率密度函数的形式非已知（不能写成某些参数的函数，非典型分布）
  - **目标**：直接依据样本估计总体分布
- **典型方法**
  - Parzen窗法
  - k_n近邻法

参数化方法

- 极大似然估计
  - 把待估计参数看做是确定的量，只是其取值未知
  - 最佳估计就是使产生已观测到样本的概率最大的那个值
- 贝叶斯估计
  - 把待估计参数看做是符合某种先验概率分布的随机变量
  - 对样本进行观测的过程，就是把先验概率密度转化为后验概率密度，从而利用样本信息修正参数的初始估计值

极大似然估计
（Maximum Likelihood Estimation, MLE）

- **假设条件：**
  - $P(x|\omega_i)$具有某种确定的函数形式，只其参数θ未知；参数θ通常为向量，如一维正态分布$N(\mu_i,\sigma_i^2)$，$\theta_i=[\mu_i,\sigma_i^2]^T$
  - 参数θ是确定的未知量（不是随机量）
  - 各类样本集x_i，i=1,2,…,c满足独立同分布(i.i.d.)，即x_i均为从密度为$P(x|\omega_i)$的总体中独立抽取出来的
  - 各类样本只包含本类分布的信息；因此，$P(x|\omega_i)$可记为$P(x|\omega_i;\theta_i)$或$P(x;\theta_i)$
- 基于上述假设，各类条件概率密度可根据各类样本分别估计

极大似然估计

- **似然函数（Likelihood function）**：
  - 针对一类已知样本 $X=\{x_i, i=1,2,\ldots, N\}$，定义参数$\theta$下观测到样本集$X$的(联合分布)概率密度，称为相对于样本集$X$的$\theta$的**似然函数**

$$l(\theta)=p(x_1,x_2,\ldots,x_n;\theta)=\prod_{i=1}^{n}p(x_i;\theta)$$

  - 给出了从总体中抽取$x_1, x_2,\ldots,x_N$这N个样本的概率
  - 将样本观测值的“概率密度/似然程度”看成是未知参数θ的函数，因此可根据已知样本估计未知参数θ!

极大似然估计

- **基本思想**：就是在θ可能的取值范围内选择使似然函数达到最大的参数值 $\hat{\theta}$ 作为参数θ的估计值，即求 $\hat{\theta}$ ，使得

$$l(\hat{\theta}) = \max_{\theta} l(\theta)$$

- 即如果参数 $\theta = \hat{\theta}$ 时，$l(\theta)$ 最大，则$\hat{\theta}$应该是“最可能”的参数值。它是样本集的函数，记作：$\hat{\theta} = d(x_1, x_2, \ldots, x_N) = d(X)$ 称为**极大似然估计量**
- 为便于分析，也可定义**对数似然函数** $H(\theta) = \ln l(\theta)$

极大似然估计

- **求解：**
  - 若似然函数连续可微，最大似然函数估计量就是方程

$$dl(\theta)/d\theta = 0 \quad \text{或} \quad dH(\theta)/d\theta = 0$$

的解(必要条件)

  - 多参数情况：若未知参数不止一个，即 $\theta = [\theta_1, \theta_1, ..., \theta_s]^T$，则需求解以下$s$个方程组即可

$$dH(\theta)/d\theta_i = 0, \quad i = 1, 2, ..., s$$

极大似然估计

- **讨论：**
  - 若似然函数连续可导且存在最大值，且必要条件方程有唯一解，则其解就是极大似然估计量（如正态分布）
  - 如果必要条件有多个解(多个极值)，则是似然函数值最大者为极大似然估计量
  - 若不满足连续可导，不能通过似然方程求极大似然估计（方程无解）；若似然函数单调时，可根据极大似然思想，将似然函数最大值点作为参数的极大似然估计值（如均匀分布）

极大似然估计

- **示例-单变量正态分布**
- **已知：**
  - 参数 $\theta = [\theta_1, \theta_2], \quad \theta_1 = \mu, \theta_2 = \sigma^2$
  - 密度函数 $p(x|\theta) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right]$
  - 样本集： $X = \{x_1, x_2, ..., x_N\}$

极大似然估计

- **示例-单变量正态分布**
- **求解：**
  - 似然函数：$$l(\theta)=p(x_1,x_2,\ldots,x_N;\theta)=\prod_{i=1}^{N}p(x_i;\theta)$$
  - 对数似然函数：$$H(\theta)=\ln l(\theta)=\sum_{i=1}^{N}\ln p(x_i\,|\,\theta)$$
  - 求解：$$\begin{cases}\dfrac{\partial H}{\partial \mu}=0\\[2mm]\dfrac{\partial H}{\partial \sigma^2}=0\end{cases}\qquad\begin{cases}\dfrac{1}{\sigma^2}[\sum_{i=1}^{N}x_i-N\mu]=0\\[2mm]-\dfrac{n}{2\sigma^2}+\dfrac{1}{2\sigma^4}\sum_{i=1}^{N}(x_i-\mu)^2=0\end{cases}$$

得：

$$\hat{\mu}=\frac{1}{N}\sum_{i=1}^{N}x_i\qquad \hat{\sigma}^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu})^2$$

极大似然估计

- **示例-单变量均匀分布**
- **已知：**
  - **参数：** $$\theta = [\theta_1, \theta_2]$$
  - **密度函数：** $$p(x \mid \theta) = \begin{cases} \dfrac{1}{\theta_2 - \theta_1} & \theta_1 < x < \theta_2 \\ 0 & \text{otherwise} \end{cases}$$
  - **样本集：** $$X = \{x_1, x_2, \ldots, x_N\}$$

极大似然估计

- **示例-单变量均匀分布**
- **求解：**
  - **似然函数：**
  $$l(\theta)=\begin{cases}\dfrac{1}{(\theta_2-\theta_1)^N}\\ 0\end{cases}$$
  - **对数似然函数：**
  $$H(\theta)=-N\ln(\theta_2-\theta_1)$$
  - **求解：**
  $$\partial H/\partial\theta_1=N\cdot\frac{1}{\theta_2-\theta_1},\partial H/\partial\theta_2=-N\cdot\frac{1}{\theta_2-\theta_1}$$
  - **得：**
  $$\hat{\theta}_1=\min\{x\},\hat{\theta}_2=\max\{x\}$$

极大似然估计

- **课后练习**-试针对给定样本数据，求解以下常用分布的极大似然估计量
  - 0-1分布
  - 二项分布
  - 泊松分布
  - 指数分布
  - …

贝叶斯估计

- **思路：**
  - 与贝叶斯决策类似，只是离散的状态决策(ω_i)变成了连续的估计(θ)
- **基本思想：**
  - 把待估计参数 θ 看作是具有先验分布 $p(\theta)$ 的随机变量，其取值与样本集 X 有关，根据样本集 X 估计 θ（利用样本将先验概率修正为后验概率）

回顾-最小风险贝叶斯决策

- **损失函数**：对于特定的x采取决策 α_i 的期望损失：$\lambda(\alpha_i, \omega_j)$
- **条件期望损失**：

$$R\left(\alpha_i \middle| \mathbf{x}\right) = E[\lambda(\alpha_i, \omega_j)] = \sum_{j=1}^{c} \lambda(\alpha_i, \omega_j) P\left(\omega_j \middle| \mathbf{x}\right), i = 1, 2, \ldots, a$$

- **期望风险**：对所有可能的x采取决策 $\alpha(x)$ 所造成的期望损失之和

$$R\left(\alpha\right) = \int R(\alpha(x) \mid x) p(x) \mathrm{d}x$$

回顾-最小风险贝叶斯决策

- **目标：风险最小**

若对每一个决策，都使其条件风险 $R(\alpha_i|\mathbf{x})$ 最小，则对所有x做出决策时，其期望风险 R 也最小

- **决策：**

如果 $R(\alpha_k|\mathbf{x}) = \min_{i=1,2\ldots,a} R(\alpha_i|\mathbf{x})$，则 $\alpha = \alpha_k$

贝叶斯估计

- **损失函数**：把 θ 估计为 $\hat{\theta}$ 所造成的损失，记为 $\lambda(\hat{\theta},\theta)$
  - 离散情况：损失函数表（决策表）
  - 连续情况：损失函数
    - 常用损失函数：平方误差损失函数 $\lambda(\hat{\theta},\theta)=(\theta-\hat{\theta})^2$
- **期望风险**：

$$R=\int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)p(x;\theta)d\theta dx$$

$$=\int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\,|\,x)p(x)d\theta dx$$

$$=\int_{E^d}R(\hat{\theta}\,|\,x)p(x)dx$$

贝叶斯估计

- **期望风险：** $R=\int_{E^d} R(\hat{\theta}\,|\,x)p(x)dx$
- **条件风险：** $R(\hat{\theta}\,|\,x)=\int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\,|\,x)d\theta$

其中： $x\in E^d$， $\theta\in\Theta$

- **最小化期望风险 $\Rightarrow$ 最小化条件风险（对所有可能的*x*）**
- **贝叶斯估计量：(在样本集*X*下）使条件风险（经验风险）最小的估计量 $\hat{\theta}$ ，即**

$$\hat{\theta}=\arg\min\left(R(\hat{\theta}\,|\,x)\right)=\arg\min\left(\int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\,|\,x)d\theta\right)$$

贝叶斯估计

- **定理**：若采用平方误差损失函数，则θ的贝叶斯估计量是在给定 x 时θ 的条件期望，即

$$\hat{\theta} = E(\theta \mid x) = \int_{\Theta} \theta p(\theta \mid x) d\theta$$

同理可得到，在给定样本集X 下，θ 的贝叶斯估计是:

$$\hat{\theta} = E(\theta \mid X) = \int_{\Theta} \theta p(\theta \mid X) d\theta$$

自学证明过程，可参考《模式识别》-边肇祺张学工P52

贝叶斯估计

- **算法步骤**（平方误差损失下）
  - 确定θ的先验分布：　$p(\theta)$
  - 求样本集的联合分布：$p(X\mid\theta)=\prod_{i=1}^{N}p(x_i\mid\theta)$
  - 求θ的后验概率分布：$p(\theta\mid X)=\dfrac{p(X\mid\theta)p(\theta)}{\int_{\Theta}p(X\mid\theta)p(\theta)d\theta}$
  - 求θ的贝叶斯估计量：$\hat{\theta}=\int_{\Theta}\theta p(\theta\mid X)d\theta$

贝叶斯估计

- **示例-单变量正态分布**
  - **已知(一维)** : $p(x|\mu) \sim N(\mu, \sigma^2)$　σ^2 已知，估计μ
  - **求解(**自学证明过程，可参考《模式识别》-边肇祺张学工P56-57)
    - 假设先验分布：　$p(\mu) \sim N(\mu_0, \sigma_0^2)$
    - 可得：　$\hat{\mu} = \dfrac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} m_N + \dfrac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0,$　　$m_N = \dfrac{1}{N}\sum_{i=1}^{N} x_i$
    - 即：　N=0时，　$\hat{\mu} = \mu_0$　　　N→∞时，$\hat{\mu} = m_N$
    - 特例：若$\sigma_0^2 = 0$，则$\hat{\mu} = \mu_0$　　　若$\sigma_0 >> \sigma$，则$\hat{\mu} = m_N$

估计量的性质与评价标准

- **无偏性**：$E[\hat{\theta}(x_1, x_1, \ldots, x_N)] = \theta$
- **渐近无偏性**：$E[\hat{\theta}(x_N)] \overset{N\to\infty}{=} \theta$
- **有效性**：对估计 $\hat{\theta}_1$ 和 $\hat{\theta}_2$，若方差 $\sigma^2(\hat{\theta}_1) < \sigma^2(\hat{\theta}_2)$，则 $\hat{\theta}_1$ 更有效
- **无偏性和有效性**：对于多次估计，估计量能以较小的方差平均地表示真实值

- **一致性**：$\forall \varepsilon > 0, \lim\limits_{N\to\infty} p\left(\left|\hat{\theta}_N - \theta\right| > \varepsilon\right) = 0$

  当样本数无穷多时，每一次估计都在概率意义上任意接近真实值

非参数估计

- **问题：**
  - 前提：概率密度函数的形式非已知（不能写成某些参数的函数，非典型分布）
  - 目标：直接依据样本估计总体分布
- **典型方法**
  - Parzen窗估计
  - k_n近邻估计

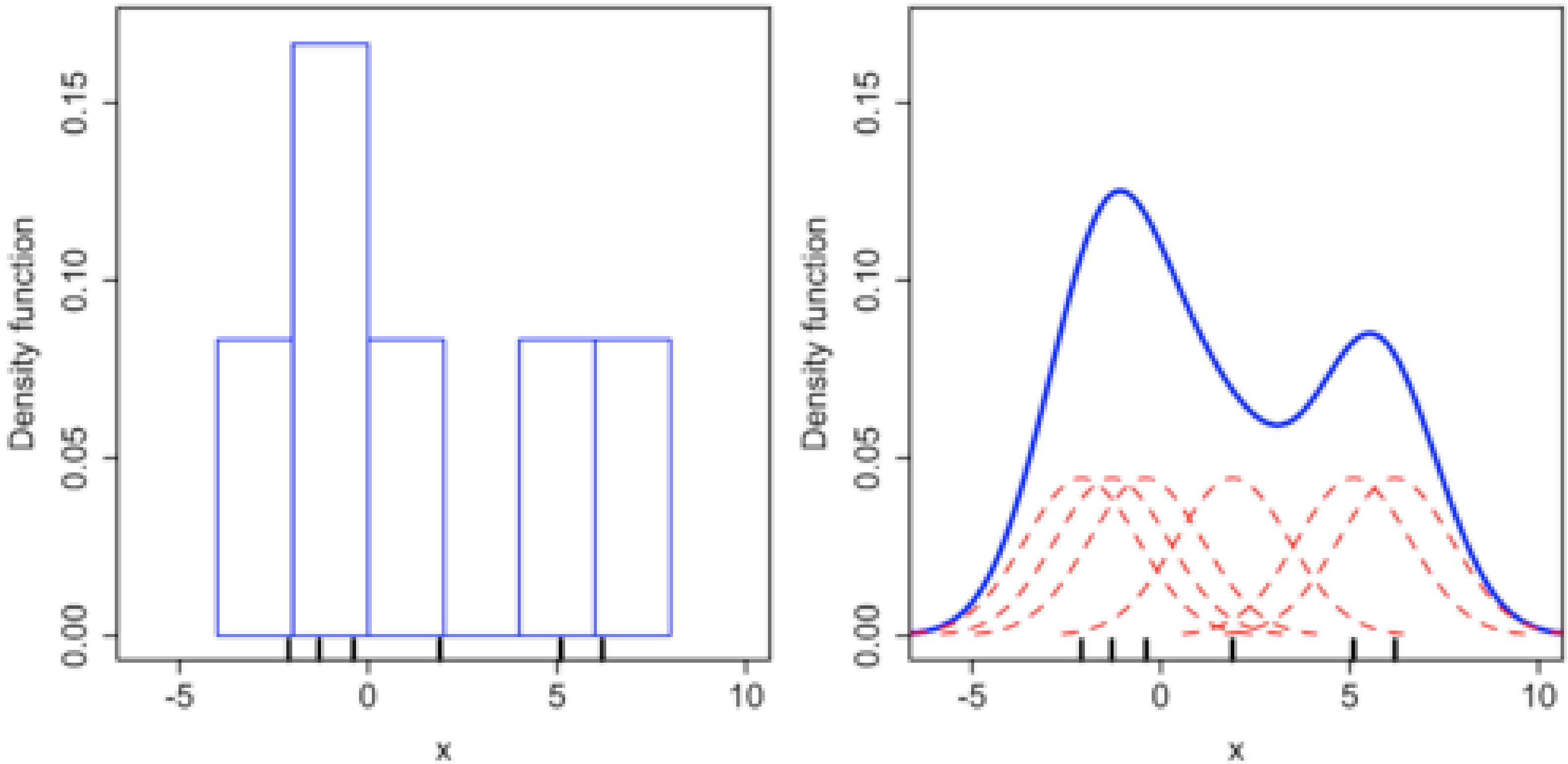
参考：《模式识别》-边肇祺张学工P65-72

非参数估计

- 基本方法-直方图方法
  - 基本思路：要估计x_i点的密度$p(x_i)$，可把所有样本在该点的"贡献"相加近似作为其概率密度，进而得到 $\hat{p}(x)$
  - 算法步骤：
    - 把x 的每个分量分成k 个等间隔bin小窗（$x \in E^d$，则形成k^d 个小舱）
    - 统计落入各个小舱内的样本数 q_i
    - 各小舱概率密度为q_i/(NV) (N:样本总数，V:小舱体积）

非参数估计

- 基本方法-直方图方法
  - 示意图



非参数估计

- **一般方法**
  - 设$p(x)$为x的总体概率密度函数，N个样本 $x=\{x_1, x_2, \ldots, x_N\}$从密度为$p(x)$的总体中独立抽取，估计 $\hat{p}(x)$ 近似 $p(x)$
  - 考虑随机变量x落入区域R的概率

$$P_R = \int_R p(x)dx$$

  - N个样本中k 个样本落入区域R的概率符合二项分布

$$P_k = C_N^k P_R^k (1-P_R)^{N-k}$$

其中 P_R 为样本 x 落入区域 R 的概率

非参数估计

●一般方法

– *k*的期望值 $$E[k]=NP_R$$

– P_R的估计 $$\hat{P}_R=\frac{k}{N}$$

– 设p(x)连续，且区域R足够小(体积V)也足够小，则有

$$\hat{P}_R=\int_R p(x)dx=\hat{p}(x)V$$

– 即： $$\hat{p}(x)=\frac{k}{NV}$$

与总样本数*N*、区域的体积 *V*及落入的样本数*k*有关

– *V*的选择：过大, 估计粗糙；过小, 可能某些区域无样本

非参数估计

一般方法

$$\hat{p}(x) = \frac{k}{NV}$$

$$\lim_{N\to\infty} V_N = 0$$

$$\lim_{N\to\infty} k_N = \infty$$ → $\hat{p}_N(x)$收敛于$p(x)$

$$\lim_{N\to\infty} \frac{k_N}{N} = 0$$

– **V的选择：过大，估计粗糙；过小，可能某些区域无样本**

非参数估计

- **Parzen窗法**
  - —使区域体积序列V_N以N的某个函数的关系不断缩小
  - —同时限制k_N和k_N/N。

有限的N, V选择很敏感

↓

- **k_n近邻法**
  - —使落入区域样本数k_N为N的某个函数
  - —选择V_N使区域包含x的k_N个近邻

动态变化V的取值

Parzen窗法

- **Parzen窗法**

$$\hat{p}(x)=\frac{1}{N}\sum_{i=1}^{N}k(x,x_i)$$

- **窗函数/核函数**
  - $k(x,x_i)$ **:反映 x_i 对 $p(x)$ 的贡献，实现小区域选择**
  - **条件：** $k(x,x_i)\geq 0,\quad \int k(x,x_i)\mathrm{d}x=1$
- **窗宽选择**
  - **原则：样本数多则选小些；样本数少则选大些**

Parzen窗法

- 常用窗函数

(1) 方窗函数(见图 3.5(a))

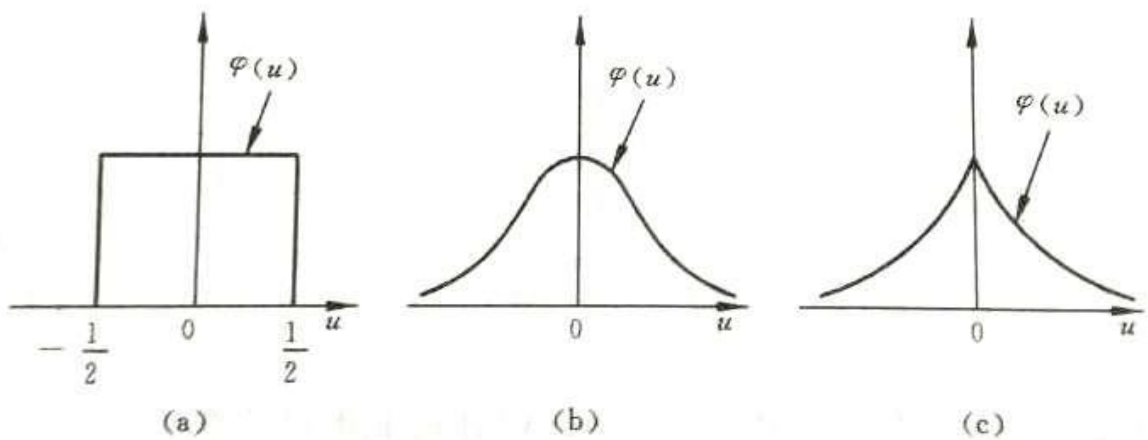
$$\varphi(u)=\begin{cases}1, & |u|\leqslant \frac{1}{2}\\ 0, & \text{其他}\end{cases}$$

(2) 正态窗函数(见图 3.5(b))

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}u^2\right\}$$

(3) 指数窗函数(见图 3.5(c))

$$\varphi(u)=\exp\{-|u|\}$$



(a)方窗　(b)正态窗　(c)指数窗

k_n近邻法

- **Parzen窗法问题**
  - 核和体积固定，若样本分布不均匀，则不能得到满意估计
- **解决办法**
  - 不使用固定区域，而是固定落在区域内的样本数；即通过控制小区域内的样本数k_N来确定小区域大小
  - 如共划分k个窗，每个窗内含k_N个样本

课外复习

矩阵理论基础

张永飞

2023年9月19日

矩阵的定义

- **定义:** 由$m \times n$个数a_{ij} $(i=1,2,\cdots,m\,;\,j=1,2,\cdots,n)$ 排成的m行n列的数表

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

称为m行n列矩阵，简称$m \times n$矩阵

- 简记为: $\boldsymbol{A}=\boldsymbol{A}_{m\times n}=(\boldsymbol{a}_{ij})_{m\times n}$
- 这$m\times n$个数称为矩阵A的元素，a_{ij}称为矩阵A的第i行第 j列元素

矩阵的定义

- **几种特殊矩阵**

(1) 行数与列数都等于n的矩阵A, 称为n阶方阵. 也可记作A_n,

例如: $\begin{pmatrix} 13 & 6 & 5 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{pmatrix}$ 是一个3 阶方阵.

(2) 只有一行的矩阵 $A=(a_1,a_2,\cdots,a_n)$,称为行矩阵(或行向量)

(3) 只有一列的矩阵 $B=\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$, 称为列矩阵(或列向量).

矩阵的定义

- **几种特殊矩阵**

(4) 元素全为零的矩阵称为零矩阵, 记作O.

注意： 不同阶数的零矩阵是不相等的.

例如 $\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \neq (0 \quad 0 \quad 0 \quad 0).$

(5) 形如 $\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$ 的方阵, 称为单位矩阵,

其中主对角线上元素都是1，其他元素都是0。记作:E_n 或 E

矩阵的定义

- **几种特殊矩阵**

(6) 形如 $\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$ 的方阵, 称为对角矩阵(或对角阵),

其中$\lambda_1, \lambda_2, \cdots, \lambda_n$不全为零.记作

$$A=\mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$$

(7) 设$A=(a_{ij})$为 n 阶方阵, 对任意 i, j, 如果$a_{ij}=a_{ji}$都成立, 则称A为对称矩阵.

例如: $A=\begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 4 \\ 3 & 4 & 6 \end{pmatrix}$ 为对称矩阵.

矩阵的定义

- **注：行列式与矩阵的区别:**
  - 一个是算式（数值），一个是数表
  - 一个行列数相同，一个行列数可不同
  - 对 n 阶方阵可求它的行列式. 记为: $|A|$

矩阵的运算-加法

•**定义:**设有两个$m\times n$ 矩阵$A=(a_{ij})$与 $B=(b_{ij})$，那么矩阵**A与B**的和记作**A+B**，规定为

$$A+B=\begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \cdots & a_{2n}+b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \cdots & a_{mn}+b_{mn} \end{pmatrix}$$

注意:仅当两个矩阵是同型矩阵时，才能进行加法运算

•**性质:**

(1) $A+B=B+A$ (2) $(A+B)+C=A+(B+C)$

(3) $A+(-A)=0$ $A-B=A+(-B)$

矩阵的运算-数乘

•**定义:**数λ与矩阵*A*的乘积记作**λA**或**Aλ**，规定为

$$\lambda A = A\lambda = \begin{pmatrix} \lambda a_{11} & \lambda a_{12} & \cdots & \lambda a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \cdots & \lambda a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \lambda a_{m1} & \lambda a_{m1} & \cdots & \lambda a_{mn} \end{pmatrix}$$

•**性质：** (设A、B都是$m\times n$ 矩阵, λ, μ 为数)：

$$(1)\ (\lambda\mu)A = \lambda(\mu A)$$

$$(2)\ (\lambda+\mu)A = \lambda A + \mu A$$

$$(3)\ \lambda(A+B) = \lambda A + \lambda B$$

矩阵相加与矩阵数乘合起来统称为**矩阵的线性运算**

矩阵的运算-乘法

•**定义:** 设$A=(a_{ij})$是一个 $m\times s$ 矩阵，$B=(b_{ij})$是一个$s\times n$ 矩阵，定义矩阵A与矩阵B的乘积 $C=(c_{ij})$是一个$m\times n$ 矩阵（ i=1,2,···, m; j=1,2,···, n ），其中

$$c_{ij}=a_{i1}b_{1j}+a_{i2}b_{2j}+\cdots+a_{is}b_{sj}=\sum_{k=1}^{s}a_{ik}b_{kj}$$

并把此乘积记作$C=AB$。记号AB常读作A左乘B或B右乘A

注意: 只有当第一个矩阵的列数等于第二个矩阵的行数时，两个矩阵才能相乘

矩阵的运算-乘法

- **性质:**
  - 结合律：$A(BC)=(AB)C$
  - 分配率：$A(B+C)=AB+AC$, $(B+C)A=BA+CA$
  - $\lambda(AB)=(\lambda A)B=A(\lambda B)$ $AE=EA=A$
- **注意：** (1)矩阵乘法不满足交换律，即: $AB \neq BA$

$$A=\begin{pmatrix}1 & 1\\ 0 & 2\end{pmatrix}, B=\begin{pmatrix}0 & 0\\ 2 & 1\end{pmatrix}, \text{但}AB=\begin{pmatrix}2 & 1\\ 4 & 2\end{pmatrix} \neq BA=\begin{pmatrix}0 & 0\\ 2 & 4\end{pmatrix}$$

(2)若AB=0；不能推出A=0或B=0

$$A=\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}, B=\begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix}, AB=\begin{pmatrix}0 & 0\\ 0 & 0\end{pmatrix}, \quad \text{但}A \neq 0,\ B \neq 0.$$

(3) 不满足消去率，即若AB=AC且A!=0，不能推出B=C

$$A=\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}, B=\begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix}, C=\begin{pmatrix}0 & 0\\ 0 & 0\end{pmatrix}. \quad AB=AC\text{，但}B \neq C$$

矩阵的运算-幂

•**定义:**若A是n 阶方阵, 则A^k为A的k次幂，即

$$A^{k+1}=A^kA=\underbrace{AA\cdots A}_{k}A$$

且满足幂运算律: $A^kA^m=A^{k+m}$, $(A^m)^k=A^{mk}$, 其中k, m为正整数

•**注意:** 由于矩阵乘法不满足交换律，则:

$(1)(AB)^k\neq A^kB^k$

$(2)A^2-B^2\neq(A+B)(A-B)$

$(3)(A+B)^2\neq A^2+2AB+B^2$

$(4)(A-B)^2\neq A^2-2AB+B^2$

矩阵的运算-转置

- **定义:** 把矩阵A的行换成同序数的列得到一个新矩阵，叫做A的**转置矩阵**，记作A^{T}

例: $A=\begin{pmatrix}1&2&3\\4&5&6\end{pmatrix}$, $A^T=\begin{pmatrix}1&4\\2&5\\3&6\end{pmatrix}$

- **性质** (假设运算都是可行的):

(1) $(A^T)^T=A$; (2) $(A+B)^T=A^T+B^T$;

(3) $(\lambda A)^T=\lambda A^T$; (4) $(AB)^T=B^TA^T$;

矩阵的运算-共轭矩阵

- **定义**: 当 $A=(a_{ij})$ 为复矩阵时, 用 $\overline{a_{ij}}$表示a_{ij} 的共轭复数, 记 $\overline{A}=(\overline{a_{ij}})$ ，称 $\overline{A}$ 为A 的**共轭矩阵**

- **性质**(设A, B为复矩阵, λ为复数, 且运算都是可行的):

(1) $\overline{A+B}=\overline{A}+\overline{B}$

(2) $\overline{\lambda A}=\overline{\lambda}\,\overline{A}$

(3) $\overline{AB}=\overline{A}\,\overline{B}$

(4) $\overline{(A^T)}=(\overline{A})^T$

矩阵的运算-逆矩阵

- **定义:** 对于n阶矩阵A，如果有一个n阶矩阵B，使

$$AB = BA = E$$

则说矩阵A是可逆的，并把矩阵B称为A的逆矩阵，简称逆阵。记作：$A^{-1} = B$

唯一性： 若A是可逆矩阵，则A的逆矩阵是唯一的

矩阵的运算-逆矩阵

- **性质**

(1) 若矩阵A可逆，则A^{-1}亦可逆，且 $(A^{-1})^{-1} = A$

(2) 若矩阵A可逆，且$\lambda \neq 0$，则 λA 亦可逆，且

$$(\lambda A)^{-1} = \frac{1}{\lambda} A^{-1}$$

(3) 若A, B为同阶可逆方阵，则AB亦可逆，

$$(AB)^{-1} = B^{-1}A^{-1}$$

(4) 若矩阵A可逆，则A^T亦可逆，且$(A^T)^{-1}=(A^{-1})^T$.

(5) 若矩阵A可逆， 则有$|A^{-1}|=|A|^{-1}$

矩阵的运算-秩

- **定义0:** 在$m \times n$矩阵A中任取 k 行 k 列($k \le m$, $k \le n$)，位于这 k 行 k 列交叉处的 k^2个元素，不改变它们在A中所处的位置次序而得到的 k 阶行列式，被称为矩阵A的k阶子式
- **定义:** 设在矩阵A中有一个不等于0的r阶子式D，且所有$r+1$阶子式(如果存在的话)全等于0，那么D称为矩阵A的一个最高阶非零子式，数r称为**矩阵A的秩，** 记作 $R(A)$

矩阵的运算-秩

- 规定：零矩阵的秩等于0
- 说明：$m \times n$矩阵A的秩$R(A)$是A中不等于零的子式的最高阶数
- 可逆矩阵的秩等于阶数。故又称可逆(非奇异)矩阵为满秩矩阵，奇异矩阵又称为降秩矩阵
- **性质**

1: $0 \le R(A_{m \times n}) \le \min\{m, n\}$

2: $R(A^T) = R(A)$

3: 若$A \sim B$, 则$R(A) = R(B)$

4: 若P, Q可逆, 则$R(PAQ) = R(A)$

5: $\max\{R(A), R(B)\} \le R(A \mid B) \le R(A) + R(B)$

6: $R(A+B) \le R(A) + R(B)$

7: $R(AB) \le \min\{R(A), R(B)\}$

8: 若$A_{m \times n} B_{n \times l} = O$, 则$R(A)+R(B) \le n$

矩阵的运算-迹

- **定义:** 如果一个矩阵A是 $n \times n$ 的方阵，则该矩阵的迹(trace)为

$$\mathbf{tr}(\mathrm{A}) = \sum_{i=1}^{n} a_{ii}$$

即：等于所有主对角线元素之和，一个实数的迹是它本身

- **性质:**
  - $\mathrm{trA}^{\mathrm{T}} = \mathrm{trA}$
  - $\mathrm{trAB} = \mathrm{trBA}$
  - $\mathrm{trABC} = \mathrm{trCAB} = \mathrm{trBCA}$

矩阵的运算-求导

定义:针对函数

$$\mathbf{y}=\Psi(\mathbf{x})$$

其中 $y\in R^{m}\times 1$，$x\in R^{n}\times 1$，则向量y关于x的导数可以表示为：

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}=\begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}$$

- 该矩阵也称为Jacobian矩阵(m×n)；
- 如果x 是一个标量，则Jacobian矩阵是一个m×1的矩阵
- 如果y是一个标量，则Jacobian矩阵是一个1×n的矩阵

矩阵的运算-求导

- 如果$\boldsymbol{y}\in R^{m}\times 1$，$\boldsymbol{x}\in R^{n}\times 1$，$A\in R^{m\times n}$，　y=Ax ，则　$\dfrac{\partial \mathbf{y}}{\partial \mathbf{x}}=\mathbf{A}$

- 如果 x 是关于 z 的函数，y=Ax，则　$\dfrac{\partial \mathbf{y}}{\partial \mathbf{z}}=\mathbf{A}\dfrac{\partial \mathbf{x}}{\partial \mathbf{z}}$

- 如果：$\alpha=y^{T}Ax$ 则：　$\dfrac{\partial \alpha}{\partial \mathbf{x}}=\mathbf{y}^{T}\mathbf{A}$　$\dfrac{\partial \alpha}{\partial \mathbf{y}}=\mathbf{x}^{T}\mathbf{A}^{T}$

- 如果：$\alpha=x^{T}Ax$, $x\in R^{n}\times 1$, $A\in R^{n\times n}$ 则：　$\dfrac{\partial \alpha}{\partial \mathbf{x}}=\mathbf{x}^{T}(\mathbf{A}+\mathbf{A}^{T})$

- 设 $\alpha=y^{T}x$, 其中 x 和 y 是关于 z 的函数，则

$$\frac{\partial \mathbf{y}}{\partial \mathbf{z}}=\mathbf{x}^{T}\frac{\partial \mathbf{y}}{\partial \mathbf{z}}+\mathbf{y}^{T}\frac{\partial \mathbf{x}}{\partial \mathbf{z}}$$