计算机模式识别

----- 理论、技术与编程

主　讲：图像处理与模式识别研究所
赵群飞

邮　箱：zhaoqf@sjtu.edu.cn
办公室：电院 2-441
电　话：13918191860

- 本节学习目标

  - ✓掌握贝叶斯公式在机器学习中的应用思路
  - ✓能够熟练运用贝叶斯决策方法
  - ✓明确分类器相关的基本概念
  - ✓掌握基于高斯分布的贝叶斯分类器
  - ✓理解朴素贝叶斯分类器
  - ✓了解各种参数估计方法（参考内容）

目录

- 贝叶斯决策是统计决策理论的基本方法。
- 理论上，在给定类条件概率密度函数和类先验概率条件下，贝叶斯决策是

  最小分类错误率和最小风险一致最优的决策。
- 对于模式分类任务而言，贝叶斯决策与估计的核心任务是：

  利用统计学中的贝叶斯定理来估计类后验概率密度函数，采用期望效用最大化和类别误判损失最小化等准则构建分类判别函数，确定样本的最优类别标记。

- 作为规范性理论，在类条件概率密度函数和类先验概率等经验知识条件下，最小错误率贝叶斯决策和最小风险贝叶斯决策的理论与方法已较完善。在这一前提下，贝叶斯决策所构建的模式分类器在统计上是最优的。
- 在最小错误率贝叶斯决策和最小风险贝叶斯决策准则的基础上，模式分类方法得到充分的发展，建立起了基于训练样本直接构建分类器的方法体系。

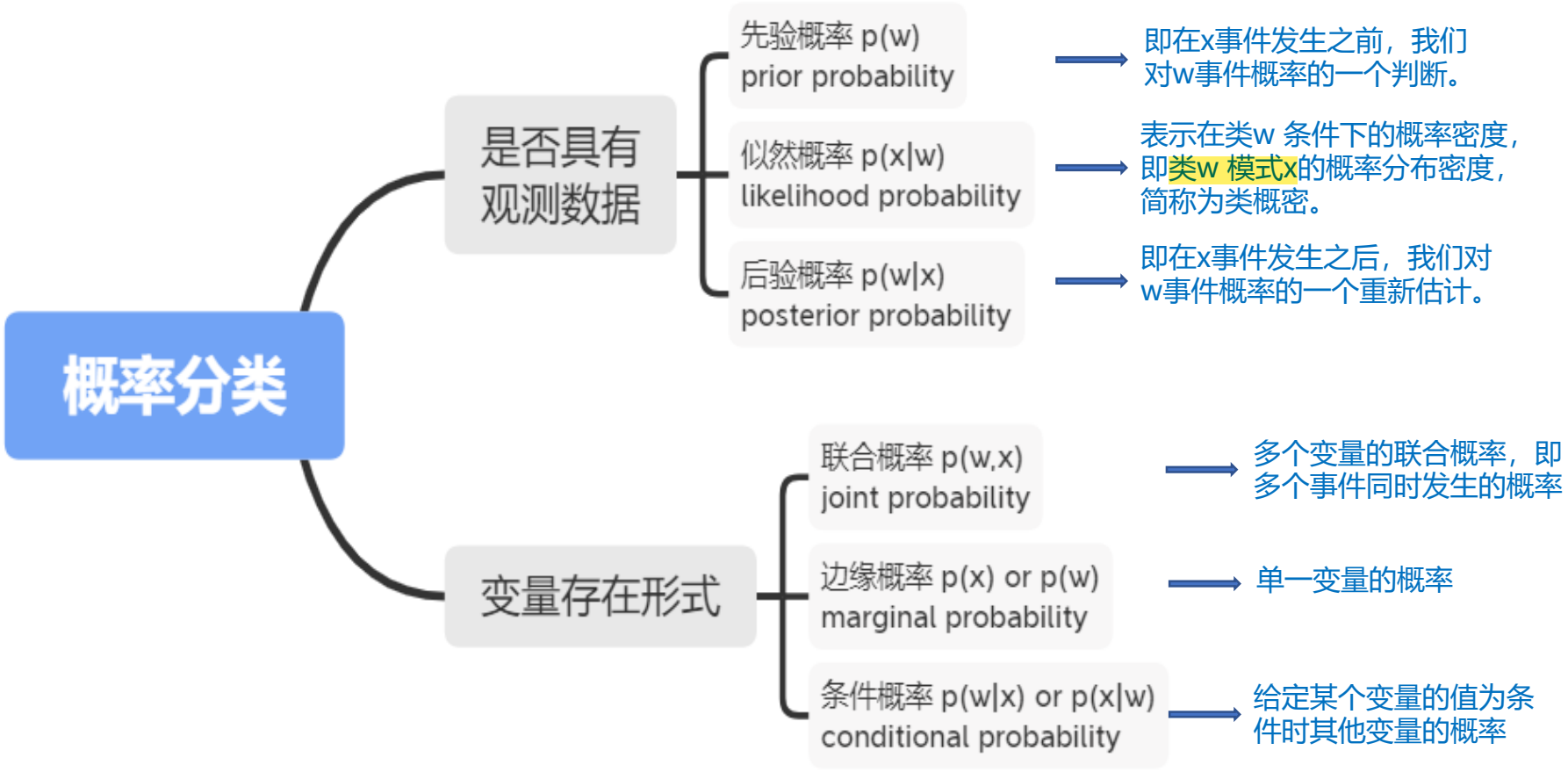
- 在技术上，针对不同的类条件概率密度函数，可构造不同的分类器。比如，常见的最近邻分类器、线性分类器、二次判别函数等均可在类条件概率密度函数为正态分布的情形下通过最小错误率贝叶斯决策来获得。
- 在此基础上，产生了带拒识决策、Neyman-Pearson决策方法、ROC曲线性能评估、连续类条件概率密度下的分类决策、离散概率模型下的统计决策、两类分类错误率估计、正态分布类条件概率密度的分类错误率估计、高维独立随机变量分类错误率估计、贝叶斯估计、贝叶斯学习、K近邻分类器的错误率界、决策树模型、朴素贝叶斯模型等基本理论与方法。

- 基于贝叶斯学习和核函数方法发展了关联向量机方法，一定程度上克服了经典支持向量机中支持向量过多且其分类性能易受正则化参数影响的缺点。
- 贝叶斯深度学习在无监督表示学习、数据生成、半监督学习、深度神经网络训练、网络结构搜索等中得到广泛应用。
- 最近几年，以贝叶斯决策与估计为基础，贝叶斯隐变量学习模型、代价敏感学习、代价缺失学习、信息论模式识别、鲁棒分类器设计、正则化方法、贝叶斯统计推断、变分贝叶斯学习等得到了充分的发展，拓展了贝叶斯决策与估计的应用范围，也进一步拓展了贝叶斯决策的方法体系。

目录

- 贝叶斯学习的思想和方法
- 贝叶斯公式
- 贝叶斯决策
- 分类器的相关概念
- 基于高斯分布的贝叶斯分类器
- 朴素贝叶斯分类器
- 参数估计（参考内容）

假设x表示观测变量，w表示模型参数：



先验概率：根据以往经验和分析得到的概率，用$P(\mathrm{w})$来代表在没有训练数据前假设 w 拥有的初始概率。

后验概率：根据已经发生的事件来分析得到的概率。

$P(\mathrm{w}|\mathrm{x})$ 反映了在看到训练数据x后w成立的置信度。

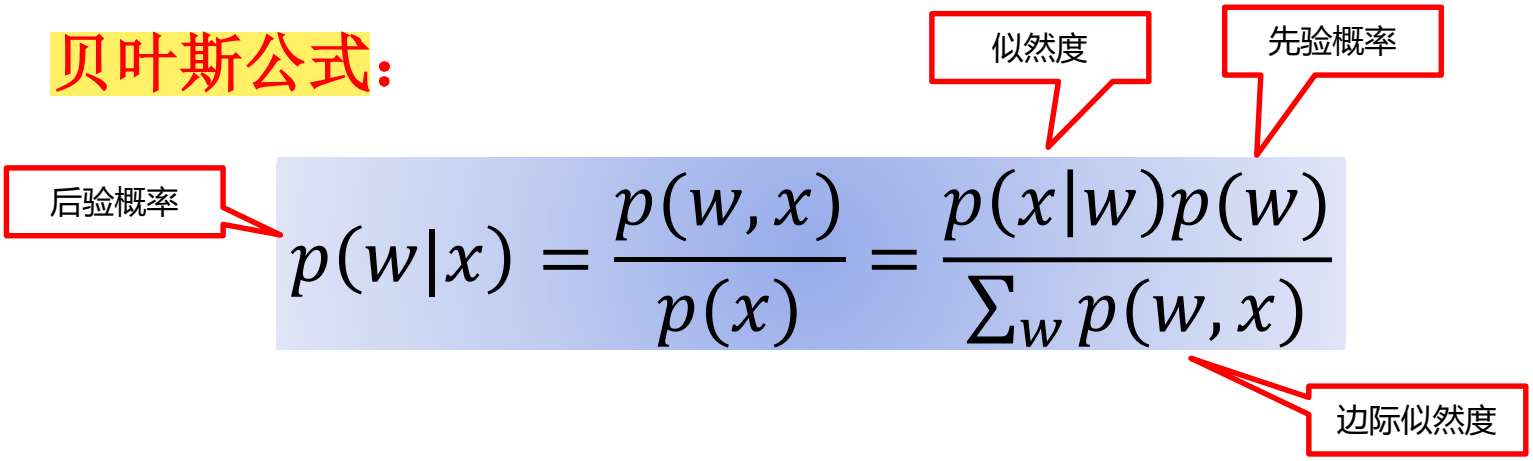
联合概率 = 条件概率×边缘概率

$$p(w,x) = p(x|w)p(w) = p(w|x)p(x)$$

某变量的边缘概率等于其他变量的概率之和,即有

$$p(x) = \sum_{w} p(w,x) \quad p(w) = \sum_{x} p(w,x)$$

贝叶斯公式:



例：假设某个动物园里的雌性和雄性熊猫的比例是4：6，雌性熊猫中90%的熊猫是干净整洁的，雄性熊猫中20%是干净整洁的。

1. 求解“**正向概率**”：

在动物园中看到一只干净整洁的雄性熊猫的概率是多少？

2. 求解“**逆向概率**”：

如果看到一只熊猫是干净整洁的，它是雄性的概率是多少？

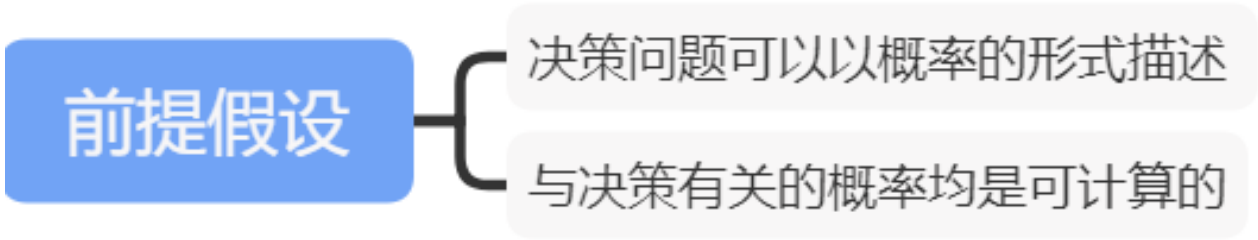
这个逆向概率的问题就可以用贝叶斯公式求解

目录

- 贝叶斯决策

贝叶斯决策（Bayesian decision）是概率框架下实施决策的基本方法，它通过综合考虑决策的后验分布和错误决策的损失来做出决策。其中，贝叶斯公式被用于计算后验分布。

贝叶斯决策的前提是假设：



例：根据熊猫的形态特征来判断熊猫的性别。

- 设w表示性别，$w=1$表示雌性，$w=2$表示雄性。

  $p(w=1)$ 熊猫为雌性的先验概率

  $p(w=2)$ 熊猫为雄性的先验概率

  那么

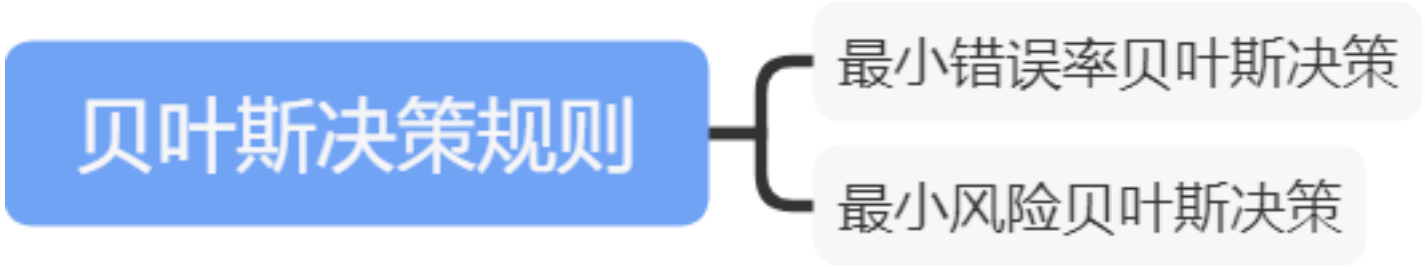
  $$p(w=1)+p(w=2)=1$$

- 假设x表示观测变量，刻画熊猫的形态特征（本例只选两种），

  $x=1$ 表示熊猫是干净整洁的

  $x=0$ 表示熊猫是不干净整洁的

在给定决策问题的概率描述（先验概率和似然概率）之后，贝叶斯决策使用贝叶斯公式推导出性别变量w的后验分布$p(w|x)$，然后通过决策规则做出决策。



- **最小错误率贝叶斯决策**

原则：决策的平均错误率尽可能地小。

熊猫分类问题**分类错误**:

- 某样本类别是雄性$w=2$，但被分为雌性$w=1$;
- 某样本类别是雌性$w=1$，但被分为雄性$w=2$;

$$p(error|x)=\begin{cases}p(w=1|x) & \text{如果}x\text{被判定为雄性}w=2\\ p(w=2|x) & \text{如果}x\text{被判定为雌性}w=1\end{cases}$$

- 最小错误率贝叶斯决策

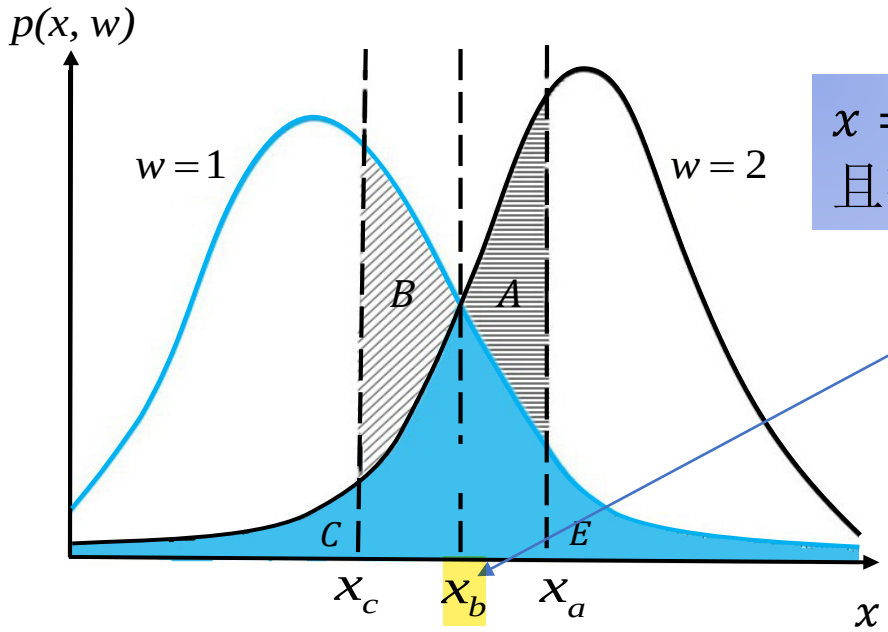
那么，**决策的平均错误率：**

$$p(error) = \int_{-\infty}^{\infty} p(error|x)p(x)dx$$

$$= \int_{R_1} p(x, w = 2)dx + \int_{R_2} p(x, w = 1)dx$$

对于二分类问题，最小错误率贝叶斯决策：

$$\begin{cases} x\text{被判定为第一类} & \text{如果} p(w = 1|x) > p(w = 2|x) \\ x\text{被判定为第二类} & \text{如果} p(w = 1|x) < p(w = 2|x) \end{cases}$$



$x = x_b$ 为错误率最小的决策边界
且有 $p(x, w = 1) = p(x, w = 2)$

图2-1 二分类的最小错误率贝叶斯决策示意图

（图中近似三角形阴影区域A和B的面积分别表示相比于$x = x_b$为决策边界，$x = x_a$和$x = x_c$作为决策边界所增加的平均错误率）

考虑更一般的多分类问题，假设存在C个类别，将特征空间分为区域$R_1, R_2, \cdots, R_C$。每一类都会错分成其他类，那么决策的平均错误率可表示为

$$p(error) =$$

$$\left[\int_{R_2} p(x, w=1)dx + \int_{R_3} p(x, w=1)dx + \cdots + \int_{R_C} p(x, w=1)dx\right]$$

$$+\left[\int_{R_1} p(x, w=2)dx + \int_{R_3} p(x, w=2)dx + \cdots + \int_{R_C} p(x, w=2)dx\right] + \cdots$$

$$+\left[\int_{R_1} p(x, w=C)dx + \int_{R_2} p(x, w=C)dx + \cdots + \int_{R_{C-1}} p(x, w=C)dx\right]$$

$$= \sum_{j=1}^{C} \sum_{j=1, j\neq i}^{C} \int_{R_j} p(x, w=i)dx$$

可能错分的情况存在$C \times (C-1)$种，涉及到的计算很多，所以通常采样计算平均正确率$p(correct)$来计算$p(error)$

$$\boldsymbol{p(error)}$$

$$= 1 - p(correct)$$

$$= 1 - \left[\int_{R_1} p(x, w=1)dx + \int_{R_2} p(x, w=2)dx + \cdots + \int_{R_C} p(x, w=C)dx\right]$$

$$= 1 - \sum_{1}^{C} \int_{R_c} p(x, w=c)dx$$

- 对于更一般化的多类分类问题，最小错误率决策表示为最大化平均正确率，平均正确率$p(correct)$的计算如下：

$$p(correct) = \sum_{1}^{C} \int_{R_c} p(x, w = c)dx$$

- 由上式可见，$p(correct)$最大化等价于**将x判别为联合概率$p(x, w)$最大类别**，即决策输出$h(x)$表示为

$$h(x) = \mathrm{argmax}_c \, p(w = c|x)$$

- 在实际分类应用中，往往不必计算后验概率。

  根据贝叶斯公式，后验概率可以表示为联合概率除以边缘概率$p(x)$，对于所有类别，分母都是相同的，所以决策时实际上只需比较分子即可，也就是说只需要计算 $p(x|w)p(w)$ ，将样本判别为其值最大类别。

- 最小风险贝叶斯决策

最小化决策带来的平均损失，也叫做*最小化风险（risk）*。

➢ 考虑一个多类分类问题，样本的真实类别为第j类，但是被误判为第i类的损失为

$$\lambda_{ij} = \lambda(h(x) = i|w = j)$$

对于C类分类问题，损失矩阵是一个$C \times C$的矩阵

$$\mathrm{L} = (\lambda_{ij})_{C \times C}$$

● 根据损失的定义可知，损失矩阵的对角元素通常为0。

平均损失的两重含义：

1. 获得观测值后，决策造成的损失对实际所属类别的各类可能的平均，称为条件风险（conditional risk）：

$$R(h(x)|x) = \sum_i \lambda(h(x)|w = i)p(w = i|x)$$

2. 条件风险对x的数学期望，称为总体风险：

$$R\big(h(x)\big) = \mathbb{E}\big(R(h(x)|x)\big) = \int R(h(x)|x)p(x)dx$$

决策函数：

$$h(x) = \underset{j}{\operatorname{argmin}} \sum_i \lambda(h(x) = j|w = i)p(w = i|x)$$

以二分类问题为例：

标记 α_1 表示把样本判别为第一类，α_2 表示把样本判别为第二类。

二分类问题中的损失矩阵 λ_{ij} 是一个 2×2 的矩阵，条件风险为：

$$R(\alpha_1|x) = \lambda_{11}p(w=1|x) + \lambda_{12}p(w=2|x)$$

$$R(\alpha_2|x) = \lambda_{21}p(w=1|x) + \lambda_{22}p(w=2|x)$$

根据最小风险贝叶斯决策规则，如果满足

$$(\lambda_{21} - \lambda_{11})p(w=1|x) > (\lambda_{12} - \lambda_{22})p(w=2|x)$$

或者满足

$$(\lambda_{21} - \lambda_{11})p(x|w=1)p(w=1) > (\lambda_{12} - \lambda_{22})p(x|w=2)p(w=2)$$

则x将被判别为第一类，否则被判别为第二类。

- 最小风险贝叶斯决策与最小错误率贝叶斯决策

假设决策损失定义为$0-1$损失，即

$$\lambda(\alpha_i|w=j)=\begin{cases}0 & i=j\\ 1 & i\neq j\end{cases}$$

此时，条件风险=条件错误率

$$\begin{aligned}R(\alpha_i|x)&=\sum_{j=1}^{C}\lambda(\alpha_i|w=j)p(w=j|x)\\&=\sum_{j\neq i}p(w=j|x)\\&=1-p(w=i|x)\end{aligned}$$

目录

- 贝叶斯公式
- 贝叶斯决策
- 分类器的相关概念
- 基于高斯分布的贝叶斯分类器
- 朴素贝叶斯分类器
- 参数估计（参考内容）

二类分类问题： 要机器来判断一张图像是大熊猫还是小熊猫

多类分类问题： 区分一张图片是大熊猫、小熊猫还是棕熊

（a）大熊猫

（b）小熊猫

（c）棕熊

分类器是一个计算系统，它通过计算出一系列判别函数的值做出分类决策，实现对输入数据进行分类的目的。

判别函数是一个从输入特征映射到决策的函数，其结果可以直接用于做出分类决策。

分类问题中，分类器会把输入空间划分成多个决策区域，这些决策区域之间的边界称作**决策面**或**决策边界**。

例：对于一个C类图像识别任务，分类器将提取的特征$\mathbf{x}$作为输入向量（例如使用图像的SIFT特征表示一张图像），然后输出一个对应的类标签。

首先，分类器计算出C个判别函数

$$\mathrm{g}_i(\mathbf{x}), i = 1,2,\cdots,C.$$

其次，分类器会把一个特征向量$\mathbf{x}$划分为类别c，如果满足：

$$\mathrm{g}_c(\mathbf{x}) > \mathrm{g}_{c'}(\mathbf{x}), \forall c' \neq c.$$

在输入空间中，使得$\mathrm{g}_c(\mathbf{x}) = \mathrm{g}_{c'}(\mathbf{x}), \forall c' \neq c$成立的超平面就是决策面。

分类器的构建方法有很多种，常用的方法大致可以分为三大类，这里按照复杂度依次降低的顺序罗列。

其中生成式模型和判别式模型都是基于概率框架，生成式模型构建所有观测的联合分布，而判别式模型只关心给定输入数据时输出数据的条件分布。

判别模型（Discriminative Model）	生成模型（Generative Model）
由数据直接学习决策函数$y = h(x)$或者条件概率分布$p(y\|x)$作为预测的模型，即判别模型。基本思想是有限样本条件下建立判别函数，不考虑样本的产生模型，直接研究预测模型。 即：直接估计$p(y\|x)$	由训练数据学习联合概率分布$p(x, y)$，然后求得后验概率分布$p(y\|x)$。具体来说，利用训练数据学习$p(x\|y)$和$p(y)$的估计，得到联合概率分布：$p(x, y) = p(y)p(x\|y)$，再利用它进行分类。 即：估计$p(x\|y)$ 然后推导 $p(y\|x)$
线性回归、逻辑回归、感知机、决策树、支持向量机……	朴素贝叶斯、HMM、深度置信网络(DBN)……

在得到一个训练好的分类器后，我们需要去评估这个分类器的性能好坏。当分类器确定后，其错误率亦随之确定了。分类器的错误率可以用于比较对于同一问题设计的多种分类器的优劣。分类器的错误率计算通常有三种方法：

1. 根据错误率的定义按照公式进行计算。
2. 计算错误率的上界。
3. 通过在测试数据上进行分类实验来估计错误率。

$$\text{Error} = \frac{1}{M}\sum_{i=1}^{M} I[h(x_i) \neq y_i]$$

其中M为测试样本的总数，$I[\cdot]$表示单位函数，当且仅当括号中的条件满足时取值为1，否则为0。把$1 - \text{Error}$称作精度Accuracy

$$\text{Accuracy} = \frac{1}{M}\sum_{i=1}^{M} I[h(x_i) = y_i]$$

目录

- 现实问题中，特征是连续的，非离散的。
- 对于连续变量，需要对类条件概率密度分布建模。
- **常用高斯分布，即，正态分布。因为计算和分析简单。**

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2)$$

- 高斯密度函数/正态密度函数
  - 一元高斯密度函数：

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

其中μ是均值，σ^2是方差，分别表示为

$$\mu=\mathbb{E}[x]=\int xp(x)dx$$

$$\sigma^2=\mathbb{E}[(x-\mu)^2]=\int(x-\mu)^2p(x)dx$$

随机变量称x为正态随机变量,服从正态分布，记为

$$\boldsymbol{x\sim\mathcal{N}(\mu,\sigma^2)}$$

方差$\sigma>0$描述不确定性： σ 越大，不确定性越大；

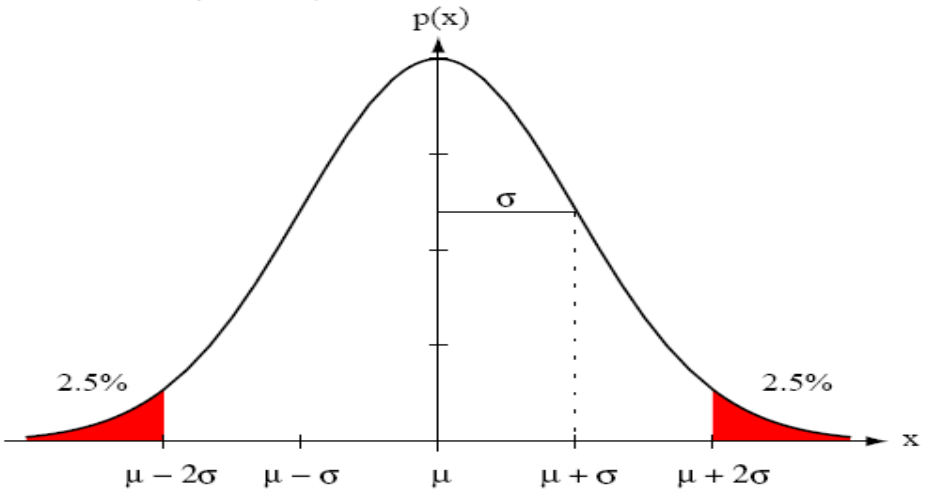
σ越小，不确定越小； σ =0，则没有不确定性。

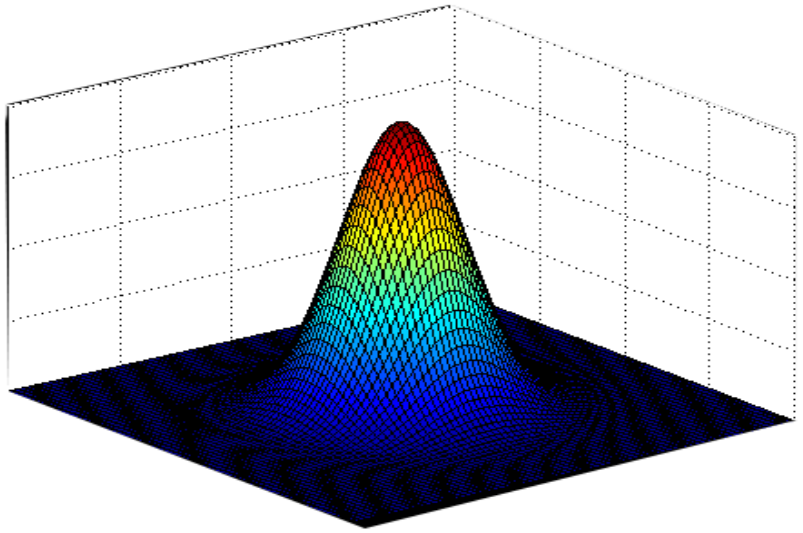
- 多元高斯密度函数：

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

其中$\boldsymbol{\mu}$是均值，Σ是协方差，d为数据维度。

- 按照定义，协方差Σ只满足非负定，不一定是正定矩阵，即有可能为 0。





- 基于高斯分布的贝叶斯决策

➢ 假设类条件概率分布为高斯分布：

$$p(\mathbf{x}|w=i)=\mathcal{N}(\boldsymbol{\mu}_i,\Sigma_i),\qquad i=1,2,\cdots,C$$

则贝叶斯决策得到的**判别函数**为：

$$\begin{aligned}\mathbf{g_i(x)}&=\ln p(\mathbf{x}|w=i)+\ln p(w=i)\\&=-\frac{d}{2}\ln 2\pi-\frac{1}{2}\ln|\Sigma_i|+\ln p(w=i)\\&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\end{aligned}$$

➢ 通过判别函数可以得到**决策面**为 $g_i(x)=g_j(x)$，即

$$-\frac{1}{2}\left[(x-\boldsymbol{\mu}_i)^{\mathrm{T}}\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)-(\mathbf{x}-\boldsymbol{\mu}_j)^{\mathrm{T}}\Sigma_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right]+\ln\frac{p(w=i)}{p(w=j)}-\frac{1}{2}\ln\frac{|\Sigma_i|}{|\Sigma_j|}=0$$

考虑当所有类别的协方差矩阵都相等的情况下，即

$$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_C = \Sigma$$

则判别函数可化简为

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{i})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{i}) + \ln p(w = i)$$

忽略与i无关的项$\mathbf{x}^\mathbf{T}\Sigma^{-1}\mathbf{x}$，判别函数进一步简化为

$$g_i(\mathbf{x}) = (\Sigma^{-1}\boldsymbol{\mu}_i)^T \mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^T \Sigma^{-1}\boldsymbol{\mu}_\mathbf{i} + \ln p(w = i)$$

同样，我们有

$$g_j(\mathbf{x}) = \left(\Sigma^{-1}\boldsymbol{\mu}_j\right)^T \mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_j^T \Sigma^{-1}\boldsymbol{\mu}_\mathbf{j} + \ln p(w = j)$$

此时判别函数是$\mathbf{x}$线性函数，决策面是一个超平面。

当决策区域R_i与R_j相邻时，决策面满足

$$\mathbf{g_i(x) - g_j(x) = 0}$$

即 $$\mathbf{\left[\Sigma^{-1}(\mu_i - \mu_j)\right]^T(x - x_0) = 0}$$

当各类别的先验概率相等时，可以得到

$$\mathbf{x_0 = \frac{1}{2}(\mu_i + \mu_j)}$$

即为$\boldsymbol{\mu_i}$与$\boldsymbol{\mu_j}$连线的中点。

图2-3展示了此时基于非对角协方差矩阵的二维高斯分布的贝叶斯决策面。

当各类别的先验概率不相等时，则$\mathbf{x_0}$不在$\boldsymbol{\mu_i}$与$\boldsymbol{\mu_j}$连线的中点上，并且向先验概率小的方向偏移。图中椭圆形的环表示类条件概率密度等高线。

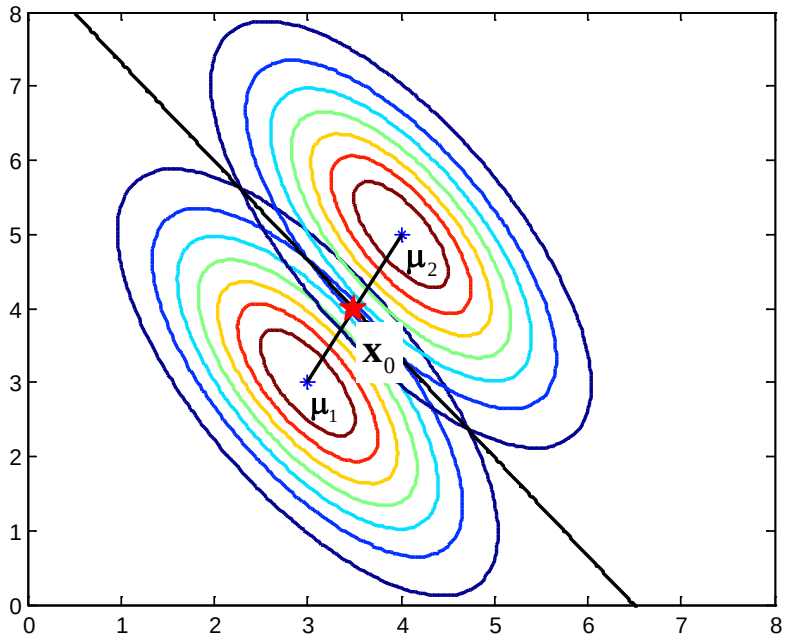


图2-3

- 基于高斯分布的贝叶斯决策的错误率

以二类问题为例，讨论两类协方差相同的情况。

为了计算错误率，这里引入最小错误率决策的负对数似然比：

$$r(\mathbf{x}) = -\ln p(\mathbf{x}|w=1) + \ln p(\mathbf{x}|w=2)$$

最小错误率贝叶斯决策可以表示为

$$\begin{cases} x\text{被判定为第一类} \quad \text{若} r(x) = -\ln p(\mathbf{x}|w=1) + \ln p(\mathbf{x}|w=2) < \dfrac{\ln p(w=1)}{\ln p(w=2)} \\ x\text{被判定为第二类} \quad \text{若} r(x) = -\ln p(\mathbf{x}|w=1) + \ln p(\mathbf{x}|w=2) > \dfrac{\ln p(w=1)}{\ln p(w=2)} \end{cases}$$

由于r(x)也是随机变量，其条件概率密度函数为$p(\mathbf{r}|w)$，贝叶斯平均错误率可转换为关于r(x)的积分。前面的式中，令

$p_1(\text{error})$ 表示将第一类样本判定为第二类的错误率，

$p_2(\text{error})$ 表示将第二类样本判定为第一类的错误率，

则通过先验概率加权可得**平均错误率**

$$p(\text{error}) = p(w=1)p_1(\text{error}) + p(w=2)p_2(\text{errror})$$

其中每一类错误率为

$$p_1(error) = \int_{R_2} p(\mathbf{x}|w=1)dx = \int_{r_B}^{\infty} p(r|w=1)dr$$

$$p_2(error) = \int_{R_1} p(\mathbf{x}|w=1)dx = \int_{-\infty}^{r_B} p(r|w=2)dr$$

这里关于$r(\mathbf{x})$的决策边界

$$r_B = \ln\frac{p(w=1)}{p(w=2)}$$

根据$p(x|w=i)=\mathcal{N}(\boldsymbol{\mu_i},\Sigma_i), i=1,2$，以及$r(x)$的定义，可得

$$r(\mathbf{x})=-\ln p(\mathbf{x}|w=1)+\ln p(\mathbf{x}|w=2)$$

$$=\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_1})^{\mathrm{T}}\Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu_1})-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_2})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu_2})+\frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|}$$

当两类协方差矩阵相等时，即$\Sigma_1=\Sigma_2=\Sigma$，上式可以化简为

$$r(\mathbf{x})=(\boldsymbol{\mu_2}-\boldsymbol{\mu_1})^{\mathrm{T}}\Sigma^{-1}\mathbf{x}+\frac{1}{2}(\mathbf{u}_1^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu_1}-\mathbf{u}_2^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu_2})$$

由上式可见，$r(\mathbf{x})$是一维随机变量，服从一维高斯分布。

设一维高斯分布$p(r(\mathbf{x})|w=1)$的期望m_1和方差σ_1^2：

$$m_1 = \mathbb{E}[r(\mathbf{x})|w=1]$$

$$= (\boldsymbol{\mu_2} - \boldsymbol{\mu_1})^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu_1} + \frac{1}{2}\left(\boldsymbol{\mu_1}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu_1} - \boldsymbol{\mu_2}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu_2}\right)$$

$$= -\frac{1}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^{\mathrm{T}}\Sigma_1^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$$

令 $$m = \frac{1}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^{\mathrm{T}}\Sigma_1^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}),$$

则期望 $$m_1 = -m,$$

方差为

$$\sigma_1^2 = \mathbb{E}[(r(\mathbf{x}) - m_1)^2 \big| w=1] = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^{\mathrm{T}}\Sigma_1^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) = 2m$$

同理可得$p(r(\mathbf{x})|w=2)$期望m_2和方差σ_2^2

$$m_2 = -\frac{1}{2}(\boldsymbol{\mu_1}-\boldsymbol{\mu_2})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{\mu_1}-\boldsymbol{\mu_2}) = m$$

$$\sigma_2^2 = (\boldsymbol{\mu_1}-\boldsymbol{\mu_2})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{\mu_1}-\boldsymbol{\mu_2}) = 2m$$

由于$\sigma_1^2 = \sigma_2^2$，记为σ^2.

$$p_1(\text{error}) = \int_{r_B}^{\infty} p(r|w=1)dr = \int_{r_B}^{\infty} \frac{1}{(2\pi)^{1/2}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{r+m}{\sigma}\right)^2\right\} dr$$

$$= \int_{r_B}^{\infty} (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{r+m}{\sigma}\right)^2\right\} d\left(\frac{r+m}{\sigma}\right)$$

$$= \int_{\frac{r_B+m}{\sigma}}^{\infty} (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\varphi^2\right) d\varphi$$

$$p_2(\text{error}) = \int_{-\infty}^{\mathrm{r_B}} p(r|w=2)dr$$

$$= \int_{-\infty}^{\frac{r_B+m}{\sigma}} (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{r-m}{\sigma}\right)^2\right\} d\left(\frac{r-m}{\sigma}\right)$$

$$= \int_{-\infty}^{\frac{r_B+m}{\sigma}} (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\varphi^2\right) d\varphi$$

式中$r_B = \ln\frac{p(w=1)}{p(w=2)}$，$\sigma = \sqrt{2m}$，

$p_1(\text{error})$和$p_2(\text{error})$表示为标准$\boldsymbol{\mathcal{N}(0,1)}$的*概率值*。

目录

- 贝叶斯学习的思想和方法
- 贝叶斯公式
- 贝叶斯决策
- 分类器的相关概念
- 基于高斯分布的贝叶斯分类器
- 朴素贝叶斯分类器
- 参数估计（参考内容）

- 贝叶斯决策的关键是计算类条件概率分布和类先验概率，往往需要在不同情况下，确定所有特征的联合概率分布。
- 数据的特征多，空间大，如有D个2值特征，2的D次方。
- 类条件概率的分布比较稀疏，只有少数事件的概率大，大部分较小，尤其是离散特征，很多情况下为0.
- 为了解决大空间和概率分布稀疏问题，朴素贝叶斯（naïve Bayes）分类器对条件概率分布提出了**特征条件独立**的假设。

- 朴素贝叶斯分类器

朴素贝叶斯法是典型的生成学习方法。由训练数据学习联合概率分布$p(x,y)$，然后求得后验概率分布$p(y|x)$。具体来说，利用训练数据学习$p(x|y)$和$p(y)$的估计，得到联合概率分布：

$$p(x,y)=p(y)p(x|y)$$

再利用它进行分类。

一般可以用极大似然估计或贝叶斯估计，估计$p(x|y)$ 然后推导 $p(y|x)$

例：对于一张大熊猫图像，它的词袋特征可以表示为一个 D维的向量，朴素贝叶斯假设向量的D个元素之间相互独立，其联合分布可以写成 个独立的概率分布相乘。

$$p(x = x_d|w = c_k) = P(x_1, \dots, x_d\,|w^k) = \prod_{j=1}^{d} P\left(x_j|w = c_k\right)$$

c_k代表类别，k代表类别个数。

这是一个较强的假设，可使模型包含的条件概率的数量大为减少，朴素贝叶斯法的学习与预测大为简化。因而朴素贝叶斯法高效，且易于实现。其缺点是分类的性能不一定很高。

基于此假设，类别w的后验概率为：

$$p(w|\mathbf{x}) = \frac{p(w)p(\mathbf{x}|w)}{p(\mathbf{x})} \propto p(w)\prod_{d=1}^{D} p(\mathbf{x}_d|w)$$

其中D为特征的个数，x_d为第d个特征上的值。

因此，基于朴素贝叶斯分类器的分类结果为

$$\text{argmax}_{\text{w}}\, p(w)\prod_{d=1}^{D} p(x_d|w)$$

目录

- 贝叶斯学习的思想和方法
- 贝叶斯公式
- 贝叶斯决策
- 分类器的相关概念
- 基于高斯分布的贝叶斯分类器
- 朴素贝叶斯分类器
- 参数估计（参考内容）

- 在贝叶斯决策中，类条件概率密度函数被假定是已知的。
- 然而由于模式分类任务通常是面向给定样本集的，其类条件概率密度函数往往是未知的。
- 因此，对类条件概率密度函数进行估计是贝叶斯决策的基础。决策过程中的一个核心环节。
- 给定一个观测样本集，概率密度估计的基本任务是采用某种规则估计出生成这些样本的概率密度函数。观测样本的分布能代表样本的真实分布，且观测样本足够充分。
- 概率密度估计的基本思路是若一个样本在观测中出现则认为在该样本所处的区域其概率密度较大而离观测样本较远的区域其概率密度较小。

- 概率密度估计方法主要包含参数估计和非参数估计。
- 参数估计方法假定概率密度函数的形式已知，所含参数未知。
- 参数法进一步分为频率派和贝叶斯两大类学派。频率派认为待估计的概率密度函数的参数是客观存在的，样本是随机的；而贝叶斯派假定待估参数是随机的，但样本是固定的。
- 频率派的代表方法为最大似然估计，贝叶斯派的代表性方法则包含贝叶斯估计和贝叶斯学习。
- 最大似然估计被广泛地应用于确定型参数的类条件概率密度函数估计，而贝叶斯估计则应用于随机型参数的类条件概率密度函数估计。

◆ 根据大数定律，先验概率$p(w)$可通过样本数据在各类别出现的频率直接估计。

例如：多类分类模型中，N为样本总数，N_c为第c类样本数，则先验概率估计为 $p(w=c)=\frac{N_c}{N}$

◆ 通常，类条件概率分布估计有两种策略

- 最大似然估计或最大后验估计，确定唯一的参数值
- 使用贝叶斯参数估计，认为要估计的参数也是随机变量，假定它服从一个先验分布，并计算这个参数在给定观测数据集$\mathcal{D}$的后验分布概率 $p(\theta|\mathcal{D})$。预测时就是用参数的后验分布。

- 最大似然估计（maximum likelihood estimation）

最大似然估计是一种给定观测时估计模型参数的方法，它试图在给定观测的条件下，找到最大化似然函数的参数值。

例：假设数据的分布是联合高斯分布的，那么似然函数就是所有观测数据以均值与协方差为参数的联合高斯密度函数，此时$\mathrm{p}(\mathcal{D}|\theta) = \mathcal{N}(\mathcal{D}|\boldsymbol{\mu}, \Sigma)$。最大似然方法找到使得似然函数$p(\mathcal{D}|\theta)$最大的模型参数的值$\hat{\theta}_{ml}$，即

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{ml}} = \mathbf{argmax}_{\boldsymbol{\theta}}\, \boldsymbol{p}(\boldsymbol{\mathcal{D}}|\boldsymbol{\theta})$$

为了计算方便，通常使用似然函数的自然对数作为优化目标，称作对数似然（log-likelihood），那么

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{ml}} = \boldsymbol{argmax}_{\boldsymbol{\theta}} \ln \boldsymbol{p}(\boldsymbol{\mathcal{D}}|\boldsymbol{\theta})$$

- 如果数据是独立同分布的且样本个数为N，那么所有训练数据的对数似然函数表示为

$$\ln p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \ln p(\mathbf{x_i}|\theta)$$

- 考虑基于高斯分布的贝叶斯分类器，给出高斯分布的最大似然估计。假设某类别具有N个样本，则类条件密度/似然密度函数的对数为

$$\sum_{i=1}^{N} \ln p(\mathbf{x_i}|\theta) = \sum_{i=1}^{N} \ln \mathcal{N}(\mathbf{x_i}|\boldsymbol{\mu}, \Sigma)$$

- 关于均值和协方差进行求导，对上式求极值，以得到均值与协方差的估计值：

$$\boldsymbol{\mu_{ml}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x_i}, \qquad \Sigma_{ml} = \frac{1}{N}(\mathbf{x_i} - \boldsymbol{\mu_{ml}})(\mathbf{x_i} - \boldsymbol{\mu_{ml}})^{\mathrm{T}}$$

- 最大后验估计
  - 最大后验估计是在最大似然估计的基础上考虑参数的先验分布，通过贝叶斯公式获得参数的后验分布$p(\theta|\mathcal{D})$，并以后验分布作为估计的优化目标。
  - 参数θ的最大后验估计$\hat{\theta}_{\text{map}}$表示为

$$\hat{\theta}_{\text{map}} = \text{argmax}_\theta \ln p(\theta|\mathcal{D}) = \text{argmax}_\theta \ln \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$
$$= \text{argmax}_\theta \ln p(\mathcal{D}|\theta) + \ln p(\theta)$$

  - 基于高斯分布的贝叶斯分类器，假设协方差已知情况下给出对均值的最大后验估计。首先假设均值是服从高斯分布的，如$\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \Sigma_\mu)$，则其对数后验概率为

$$\ln p(\boldsymbol{\mu}|\mathcal{D}) = \sum_{i=1}^{N} \ln \mathcal{N}\left(\mathbf{x_i}|\boldsymbol{\mu}, \Sigma_\mu\right) + \ln \mathcal{N}\left(\boldsymbol{\mu}|\mathbf{0}, \Sigma_\mu\right) - \ln p(\mathcal{D})$$

- 期望最大化算法（expectation maximization，EM）
  - 对不完整数据建模时，使用隐变量定义缺失数据；对复杂的观测数据建模时，使用隐变量定义潜在因素。
  - 考虑一个概率模型，X表示观测变量集，Z表示隐变量集，θ表示模型参数，目标是最大化观测变量X对参数θ的对数似然函数：

$$L(\theta) = \ln p(X|\theta) = \ln \int (X, Z|\theta) dZ$$

  - **EM算法**是一种迭代算法，常用于求解带有隐变量的概率模型的最大似然或者最大后验估计。
    - **E步**：根据给定观测变量X和当前参数θ推理隐变量Z的后验概率分布，并计算观测数据X和隐变量Z的对数联合概率关于Z的后验概率分布的期望；
    - **M步**：最大化E步求得的期望，获得新的参数θ。

算法 2-1 期望最大化算法（expectation maximization，EM）

输入：观测数据 X

1: 初始化参数 $\theta^{(1)}$，

2: REPEAT

3: E 步：记 $\theta^{(t)}$ 为第 t 次迭代参数的估计值，计算对数联合概率分布 $\ln p(X,Z\mid\theta)$ 关于隐变量 Z 的后验概率分布 $p(Z\mid X,\theta^{(t)})$ 的期望：

$$Q(\theta\mid\theta^{(t)})=\mathbb{E}_{Z\mid X,\theta^{(t)}}\ln p(X,Z\mid\theta)=\int p(Z\mid X,\theta^{(t)})\ln p(X,Z\mid\theta)dZ.$$

4: M 步：求解使 $Q(\theta\mid\theta^{(t)})$ 最大化的 θ，得到第 $t+1$ 次迭代的参数估计：

$$\theta^{(t+1)}=\arg\max_{\theta}Q(\theta\mid\theta^{(t)}).$$

5: UNTIL 满足收敛条件：

$$\|\theta^{(t+1)}-\theta^{(t)}\|<\varepsilon_1 \text{ 或 } \|Q(\theta^{(t+1)}\mid\theta^{(t)})-Q(\theta^{(t)}\mid\theta^{(t-1)})\|<\varepsilon_2,$$

其中 $\varepsilon_1,\varepsilon_2$ 是非常小的正数

输出：参数 $\theta^{(t+1)}$

- 贝叶斯参数估计

贝叶斯参数估计不直接估计参数的值，而是通过贝叶斯公式推理出参数的后验分布。因此贝叶斯参数估计得到的是参数θ在给定观测数据集$\mathcal{D}$的后验分布

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$

参数估计从训练数据X学习出参数的后验分布$p(\theta_c|X, w = c)$。

在训练完成后，利用该后验分布可以得到测试样本 的类条件概率分布为

$$p(\mathbf{x}_*|w = c) = \int p(\mathbf{x}_*|w = c, \theta_c)p(\theta_c|X, w = c)d\theta_c$$

贝叶斯参数估计量的步骤如下：

1.确定 θ 的先验分布 $p(\theta)$

2.由样本集 $X=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$ 求出样本联合分布 $p(X\mid\theta)$，它是 θ 的函数

3.利用贝叶斯公式，求出 θ 的后验分布

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

4.利用定理求出贝叶斯估计量

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

- 贝叶斯参数估计

考虑基于高斯分布的贝叶斯分类器，假设协方差已知，且

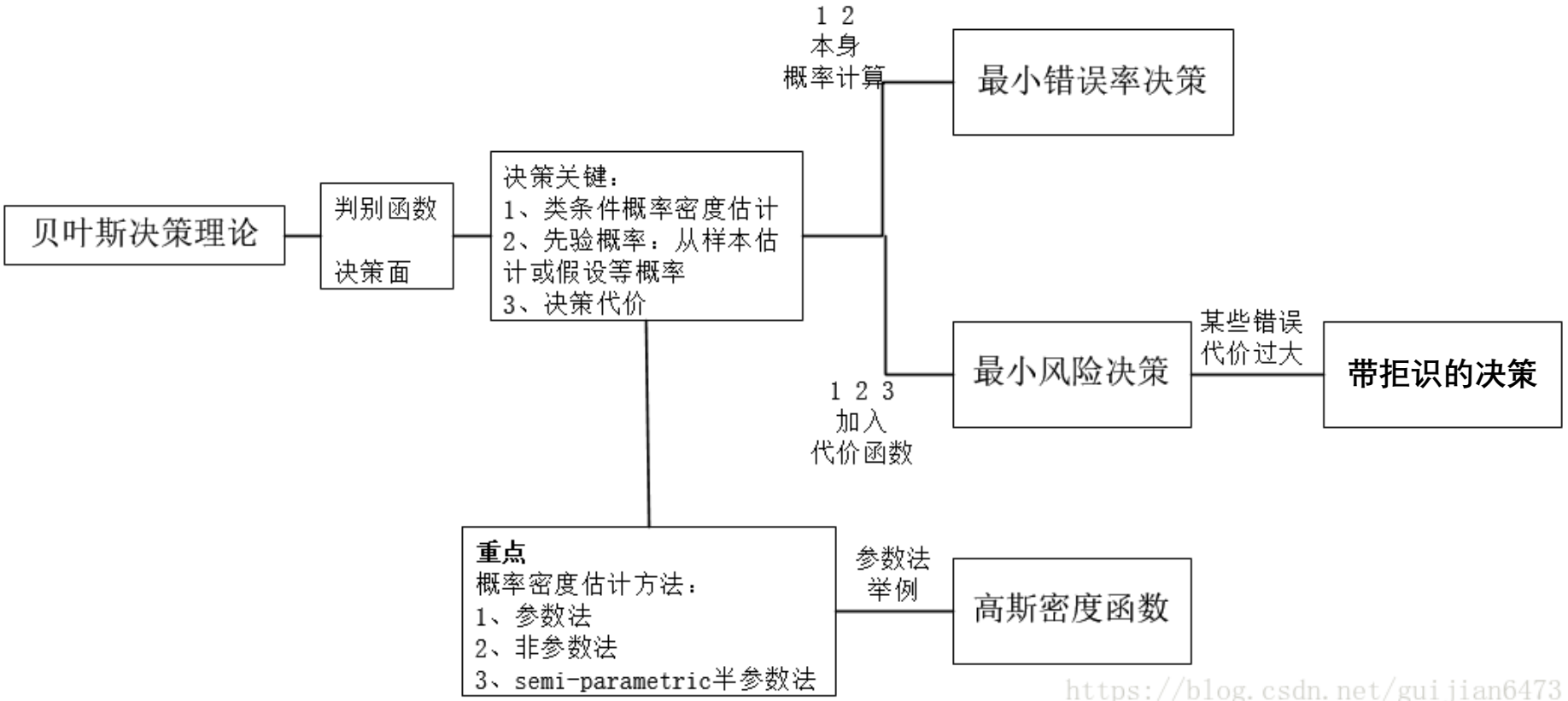
$\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \Sigma_\mu)$，

则均值参数的后验分布为

$$p(\boldsymbol{\mu}|x) = \mathcal{N}(N\Sigma_\mu\left(\mathrm{N}\Sigma_\mu + \Sigma\right)^{-1}\boldsymbol{\mu}_{ml}, \left(\Sigma_{\boldsymbol{\mu}}^{-1} + N\Sigma^{-1}\right)^{-1})$$

贝叶斯决策	贝叶斯估计
决策问题	估计问题
样本$\mathbf{x}$	样本集$\mathcal{D}$
决策α_k	估计量$\hat{\theta}$
真实状态w_i	真实参数θ
状态空间$\mathcal{A}$是离散空间	参数空间Θ是连续空间
先验概率$P(w_i)$	参数的先验分布$p(\theta)$

- 针对样本的类别是否已知，参数法估计又可分为有监督和无监督的估计方法。在每类样本独立同分布的假定下，两类方法主要依靠最大似然估计的技术路线来实现。
- 有监督的估计假定每类样本的类别标签已知，无监督的估计假定每类样本的类别标签未知。
- 无监督的估计通常需要同时对观测变量和隐变量进行估计，因此在最大似然估计的框架下，该类方法大多采用期望最大化方法来具体实现。
  - 在此基础上，衍生了概率图模型参数估计、混合高斯模型概率函数估计、Poly-tree模型参数估计、Copula 密度函数估计、隐狄利克莱分配（Latent Dirichlet Allocation）模型估计、受限玻尔滋曼机参数估计等方法。



例题1： 鱼类加工厂对鱼进行自动分类，ω_1:鲈鱼； ω_2:鲑鱼。模式特征x=x(长度)。

已知： （统计结果）

先验概率： $P(\omega_1)=1/3$（鲈鱼出现的概率）

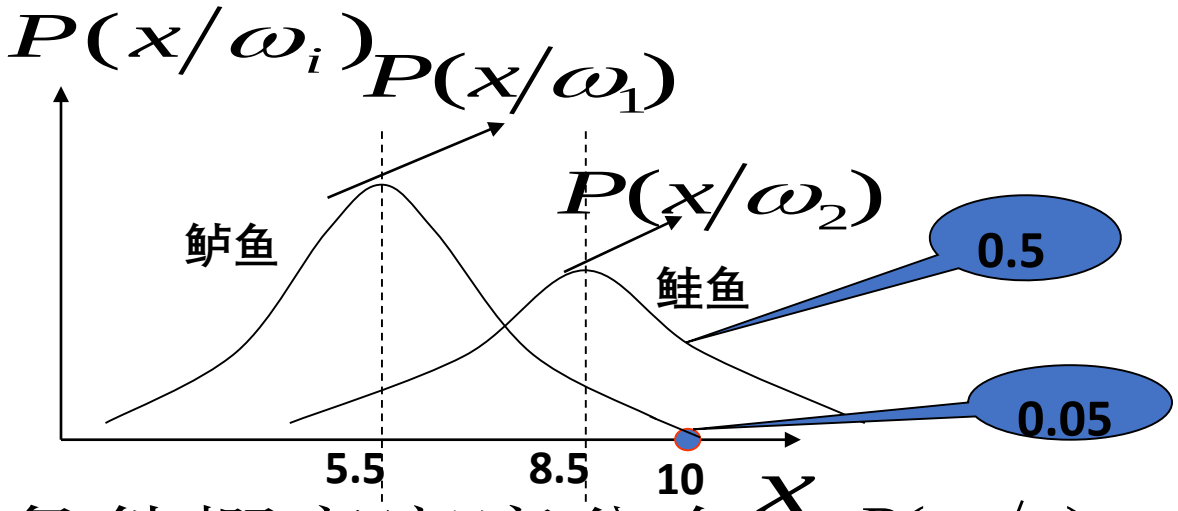
$P(\omega_2)=1-P(\omega_1)=2/3$ （鲑鱼出现的概率）

条件概率： $p(x|\omega_1)$ 见图示（鲈鱼的长度特征分布概率）

$p(x|\omega_2)$见图示（鲑鱼的长度特征分布概率）

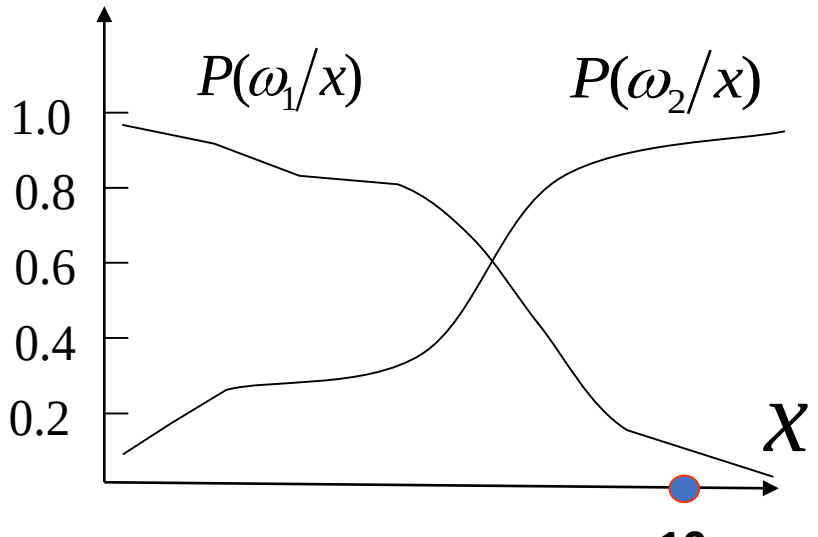
求：后验概率：$P(\omega|x=10)=?$

（如果一条鱼x＝10，是什么类别？）



条件概率密度分布

$P(\omega_i/x)$



后验概率分布

利用贝叶斯公式

$$P(\omega_1 \mid x=10)=\frac{p(x=10\mid\omega_1)P(\omega_1)}{p(x=10)}$$

$$=\frac{p(x=10\mid\omega_1)P(\omega_1)}{p(x=10\mid\omega_1)P(\omega_1)+p(x=10\mid\omega_2)P(\omega_2)}$$

$$=\frac{0.05\times1/3}{0.05\times1/3+0.50\times2/3}=0.048$$

因为，$P(\omega_2|x=10)=1-P(\omega_1|x=10)=1-0.048=0.952$

$$P(\omega_1|x=10)<P(\omega_2|x=10)$$

故判决：$(x=10)\in\omega_2$，即是鲑鱼。

例题2：对一批人进行癌症普查，患癌症者定为属ω_1类，正常者定为属ω_2类。统计资料表明人们患癌的概率 $P(\omega_1)=0.005$，从而 $P(\omega_2)=0.995$。设有一种诊断此病的试验，其结果有阳性反应和阴性反应之分，依其作诊断。化验结果是一维离散模式特征。统计资料表明：癌症者有阳性反映的概率为 0.95

即 $P(x=阳|\omega_1)=0.95$，从而可知 $P(x=阴|\omega_1)=0.05$，正常人阳性反映的概率为0.01即 $P(x=阳|\omega_2)=0.01$，可知 $P(x=阴|\omega_2)=0.99$。

问有阳性反映的人患癌症的概率有多大？

$$P(\omega_1|x=\text{阳}) = \frac{P(x=\text{阳}|\omega_1)P(\omega_1)}{P(x=\text{阳})}$$

$$= \frac{P(x=\text{阳}|\omega_1)P(\omega_1)}{P(x=\text{阳}|\omega_1)P(\omega_1)+P(x=\text{阳}|\omega_2)P(\omega_2)}$$

$$= \frac{0.95\times 0.005}{0.95\times 0.005+0.01\times 0.995}$$

$$= 0.323$$

说明有阳性反应的人其患癌的概率有32. 3%

写成似然比形式：

$$l_{12}(x)=\frac{P(x=\text{阳}|\omega_1)}{P(x=\text{阳}|\omega_2)}=\frac{0.95}{0.01}=95$$

$$\theta_{12}=\frac{P(\omega_2)}{P(\omega_1)}=\frac{0.995}{0.005}=197$$

$$\because l_{12}(x)<\theta_{12} \qquad \therefore x\in\omega_2$$

朴素贝叶斯 实战编程案例
https://cloud.tencent.com/developer/article/1014913

人人都懂EM算法
https://zhuanlan.zhihu.com/p/36331115

谢谢大家！