

2024-2025学年秋季学期

模 式 识 别

第7章：统计分类器及其学习

主讲人：张治国

zhiguo.zhang@hit.edu.cn

本章内容



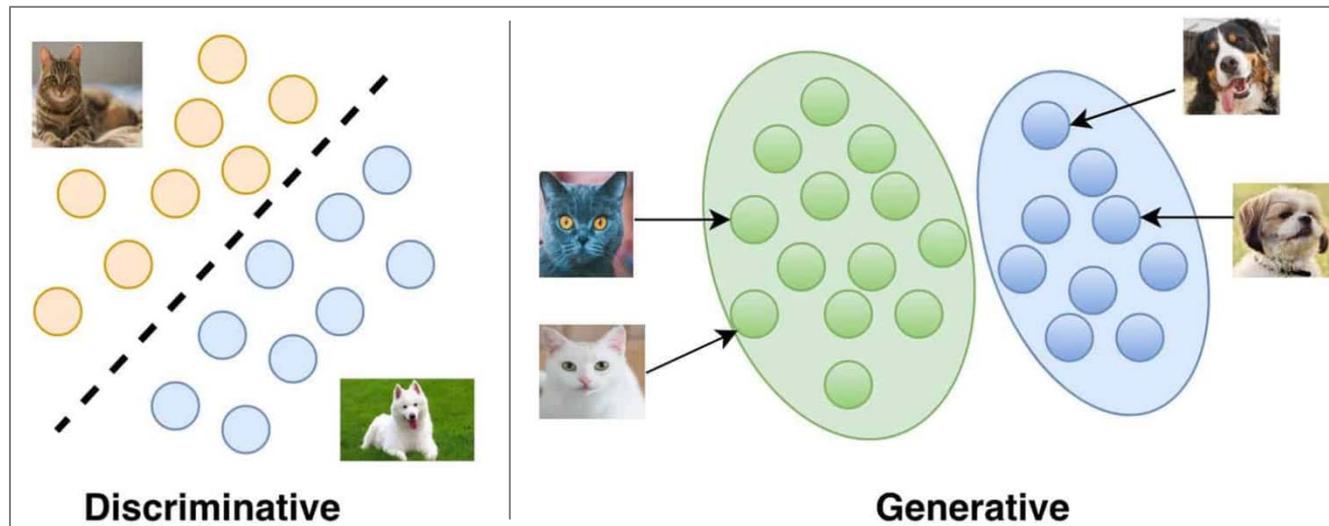
- 贝叶斯决策理论
 - 常用的概率表示形式
 - 最小错误率准则贝叶斯分类器
 - 最小平均风险准则贝叶斯分类器
 - 高斯分布贝叶斯分类器
 - 高斯分布的判别函数
 - 朴素贝叶斯分类器
-
-
- 概率密度函数的非参数估计
 - Parzen窗法
 - 近邻法

本章内容

- 概率密度函数的参数估计
 - ✓ – 最大似然估计
 - 高斯混合模型
 - 期望最大化算法
 - 贝叶斯估计
- 隐含马尔可夫模型
 - 马尔可夫模型
 - ✓ – 隐含马尔可夫模型
 - 估值问题、解码问题、学习问题

模式识别方法分类

- 分类器有两大类：判别式模型（Discriminative Model）和产生式模型（Generative Model）。
- 判别式模型：**样本模式确定地处在特征空间不同区域，通过训练得到类别边界。
- 产生式模型：**样本模式是特征空间的随机变量，估计概率密度以确定类别属性。



模式识别方法分类

- **判别式模型**
 - 将待识别模式 x 看做特征空间中的点
 - 构建判别函数 $g(x)$ 来决定 x 属于哪个类别
 - 关键在于计算 x 与训练样本间的距离关系
- **产生式模型**
 - 将待识别模式 x 看做随机变量
 - 根据 x 属于各类别的概率大小，来决定其类别
 - 关键在于计算不同类别产生待识别模式的概率
 - 产生式模型的基础是贝叶斯决策理论

贝叶斯理论

- 频率学派和贝叶斯学派是数理统计的两大流派。
- 频率学派（Frequentist）：事物是确定的，有一个本体，这个本体的真值是不变的，我们的目标就是要找到这个真值或真值所在的范围；
- 贝叶斯学派（Bayesian）：事物是不确定的，人们对事物有一个先验预判，而后通过数据调整预判，目标是找到最优的描述该事物的概率分布。

频率学派	贝叶斯学派
没有先验	有先验，随数据更新
数据是可重复的随机采样	数据是固定的
模型参数是固定的	参数以概率形式描述
相对计算简单	计算较复杂

推理

- **推理**: 从已知的条件B，依据因果关系和逻辑规则，得出某个结果A的过程。

$\Box B \rightarrow A$	确定性推理
$\Box B \rightarrow P(A B)$	概率推理
$\Box P(B A) \leftarrow A$	逆概率推理

- **确定性推理**: 如果条件B存在，一定有结果A。
 - 例：如果考试作弊，该科成绩就一定是 0 分。
- **概率推理**: 如条件B存在，则结果A发生的概率为 $P(A|B)$ ，称为条件概率。
 - 例：如果考前未复习，该科有50%的可能性不及格。

推理

- 我们更关注的是：如果发现了某个结果A，那么造成这种结果的原因条件B是什么？
- 逆概率推理：如果发现结果A出现了，求条件B存在的概率 $P(B|A)$ 是多少？
 - 例：已知某人如果是罪犯，他留下某些线索的概率；那么如果发现了一些线索，他是罪犯的概率是多少？
 - 例：已知患有某种疾病，会出现某种症状的概率；那么如果医生发现某人出现了该症状，他患有该种疾病的概率是多少？
- 解决逆概率推理问题的理论就是以贝叶斯公式为基础的贝叶斯决策理论。

贝叶斯分类器原理

- 统计分类是依据样本在各个维度上的特征值分布来进行分类决策的模式识别算法。
- 如果把样本真实所属的类别作为条件，样本的特征值作为结果，那么，
 - 样本真实类别：条件B
 - 样本特征值：结果A
 - 分类决策：逆推理 $P(B|A) \leftarrow A$

常用的概率表示形式

- 模式识别中的类别 ω 一般是离散随机变量，只有有限个取值可能；离散随机变量用它每一个可能取值的概率描述 $P(\omega_i) = P(\omega = \omega_i)$ 。
- 模式识别中的特征多是连续随机变量，需要用概率密度函数描述 $p(x)$ 。
- 多维特征矢量 $x = [x_1, \dots, x_d]^T$ 的概率密度函数是所有元素的联合概率密度函数 $p(x) = p(x_1, \dots, x_d)$ 。

常用的概率表示形式

- 假设模式的特征矢量是 x , 分类器的目的是将它分类到 $\omega_1, \dots, \omega_c$ 中的某个类别。
- 类别先验概率 (prior probability) $P(\omega_i)$: 观测数据 x 前根据经验分析得到的类别 ω_i 发生概率。
 - 所有类别的先验概率之和为1, $\sum_{i=1}^c P(\omega_i) = 1$ 。
 - 没有特征矢量 x 数据的情况下, 先验概率最大的类别即可以认为是 x 的类别。

常用的概率表示形式

- **类别后验概率** (posterior probability) $P(\omega_i | x)$: 在考虑和给出相关证据或数据后所得到的类别 ω_i 发生的条件概率。
 - 分类器的分类依据不是先验概率，而是后验概率：已知特征矢量 x 的条件下计算各类别发生概率的大小。
- **条件概率** (conditional probability) $p(x | \omega_i)$: 描述每一个类别 ω_i 样本特征的概率分布情况。
 - 条件概率可以由每个类别的训练样本估计。
 - 也称为似然 (likelihood)。

贝叶斯分类

- 已知：分类问题有 c 个类别 $\omega_1, \dots, \omega_c$ ，
各类别的先验概率 $P(\omega_i)$ ，
各类别特征的条件概率密度函数 $p(x | \omega_i)$ ，
- 决策：对于特征空间中观测到的向量 x ，应该
将 x 分到哪一类？
- 贝叶斯分类的基础是贝叶斯公式。

贝叶斯分类

- 贝叶斯公式 (Bayes Formula)

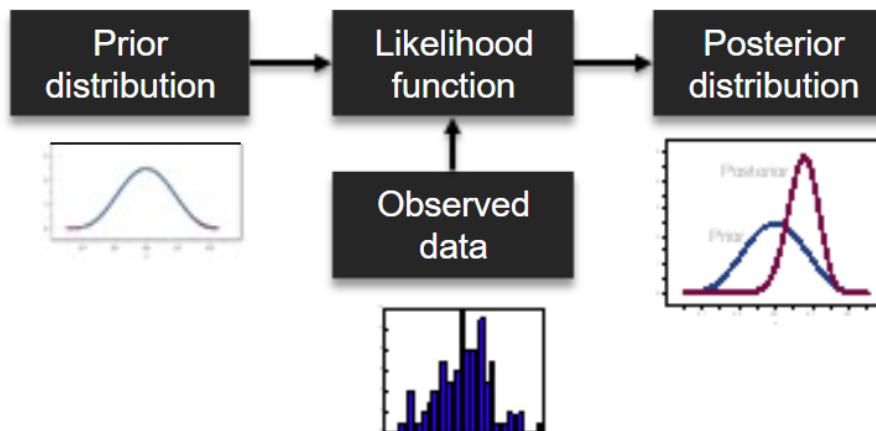
$$P(\omega_i | \mathbf{x}) = \frac{p(\mathbf{x} | \omega_i)P(\omega_i)}{p(\mathbf{x})}$$

↑ 条件概率
 (likelihood)

类別先驗概率
 (prior)

后验概率
 (posterior)

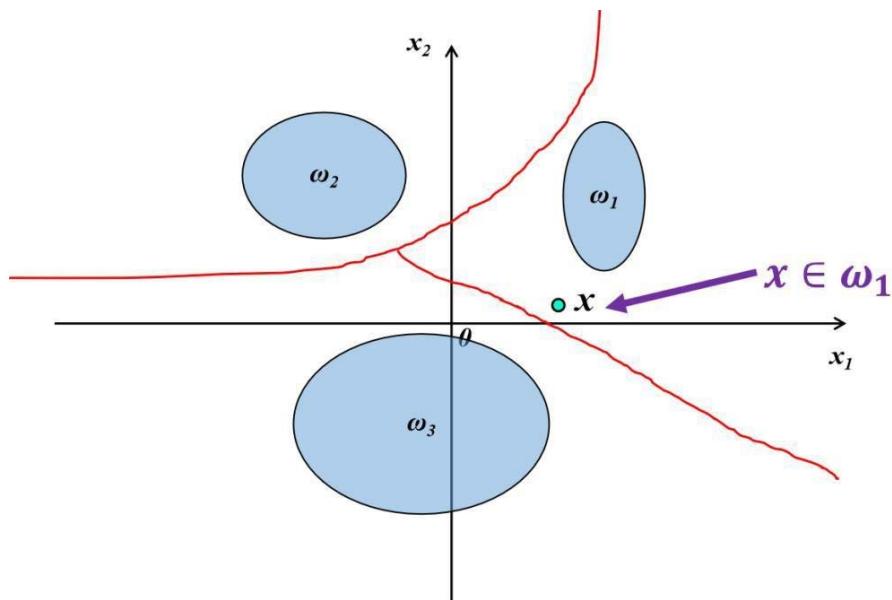
\mathbf{x} 的先验概率密度
 (evidence) , 与分类
无关, 可视为常数, 在
计算中可以忽略。



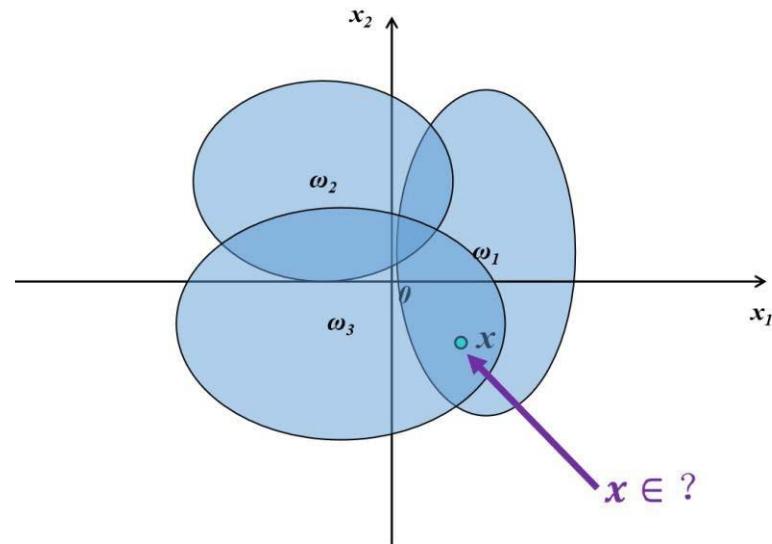
$$p(\mathbf{x}) = \sum_{i=1}^c p(\mathbf{x} | \omega_i)P(\omega_i)$$

贝叶斯分类

判别式模型
确定性的统计分类



产生式模型
不确定的统计分类



根据后验概率的大小来
做出最终的分类决策

贝叶斯分类

- 贝叶斯分类器通过每个类别的先验概率和每个类别中出现某种特征值情况的条件概率，来计算具有某有特征值的输入模式 x 属于每一类的后验概率，根据后验概率大小进行分类，将其判别为后验概率最大的类别。
- 实际中，后验概率经常计算为先验概率和条件概率密度的乘积。

贝叶斯分类实例

- 根据头发的长短（唯一特征 x ）
区分女性 (ω_1) 和男性 (ω_2)。



- 左图照片中长发的人是女性还是
男性？
- 计算后验概率 $P(\text{性别} | \text{长发})$ 。



贝叶斯分类实例

- 类别先验概率: $P(\text{女性}) = P(\text{男性}) = 50\%$
- 条件概率密度: $P(\text{长发} \mid \text{女性}) = 70\%$
 $P(\text{长发} \mid \text{男性}) = 5\%$
- 特征的先验概率密度:
$$p(x) = \sum_{i=1}^c p(x \mid \omega_i)P(\omega_i)$$
$$P(\text{长发}) = P(\text{长发} \mid \text{女性}) \times P(\text{女性}) +$$
$$P(\text{长发} \mid \text{男性}) \times P(\text{男性})$$
$$= 0.7 \times 0.5 + 0.05 \times 0.5 = 37.5\%$$

贝叶斯分类实例

- 后验概率：

$$P(\omega_i | x) = \frac{p(x | \omega_i)P(\omega_i)}{p(x)}$$

$$\begin{aligned} P(\text{女性} | \text{长发}) &= P(\text{长发} | \text{女性}) \times P(\text{女性}) / P(\text{长发}) \\ &= 0.7 \times 0.5 / 0.375 = 93.3\% \end{aligned}$$

$$\begin{aligned} P(\text{男性} | \text{长发}) &= P(\text{长发} | \text{男性}) \times P(\text{男性}) / P(\text{长发}) \\ &= 0.05 \times 0.5 / 0.375 = 6.7\% \end{aligned}$$

- 无论分类决策如何，都有可能犯错。



贝叶斯分类实例

- 如果该例中类别先验概率改变，对结果有何影响？
- 假设某一特定人群中，类别先验概率为
 $P(\text{女性}) = 25\%, P(\text{男性}) = 75\%$
- 条件概率密度不变：
 $P(\text{长发} \mid \text{女性}) = 70\%, P(\text{长发} \mid \text{男性}) = 5\%$
- 特征的先验概率密度： $P(\text{长发}) = 21.25\%$
- 后验概率：
 $P(\text{女性} \mid \text{长发}) = 82.4\%, P(\text{男性} \mid \text{长发}) = 17.6\%$
- 先验概率的改变会引起后验概率的改变。

贝叶斯分类的特点

- 先验概率必须是已知的
 - 先验概率是计算后验概率的基础
 - 先验概率可以由大量的重复实验所获得的各类样本出现的频率来近似获得
 - 先验概率可以通过新获得的信息进行更新
- 分类决策存在错误率

贝叶斯决策的方法

- 基于最小错误率的贝叶斯决策
- 基于最小平均风险的贝叶斯决策
- Neyman-Pearson决策：限定一类错误率，最小化另一类错误率
- 极小化极大准则：先验概率未知的情况下，使最大可能的风险最小化

最小错误率准则贝叶斯分类器

- 根据输入模式 x 的后验概率大小进行分类的贝叶斯分类器可以使得**错误率最小**。

- 首先检查两类问题。当观察到特征 x 时做出判别的错误率为：

$$P(e | x) = \begin{cases} P(\omega_1 | x), & \text{判定 } \omega_2 \\ P(\omega_2 | x), & \text{判定 } \omega_1 \end{cases}$$

- 两类问题最小错误率判别准则：

$$\begin{cases} \text{如果 } P(\omega_1 | x) > P(\omega_2 | x), & x \in \omega_1 \\ \text{如果 } P(\omega_2 | x) > P(\omega_1 | x), & x \in \omega_2 \end{cases}$$

Important

最小错误率准则贝叶斯分类器

- 多分类问题下，将特征 x 判别为 ω_i 类时的错误率计算为：

$$\begin{aligned} P_i(e | x) &= \sum_{j=1, j \neq i}^c P(\omega_j | x) = \sum_{j=1}^c P(\omega_j | x) - P(\omega_i | x) \\ &= 1 - P(\omega_i | x) \end{aligned}$$

- 如果希望错误率最小，则应该将 x 判别为后验概率最大的一个类别，即

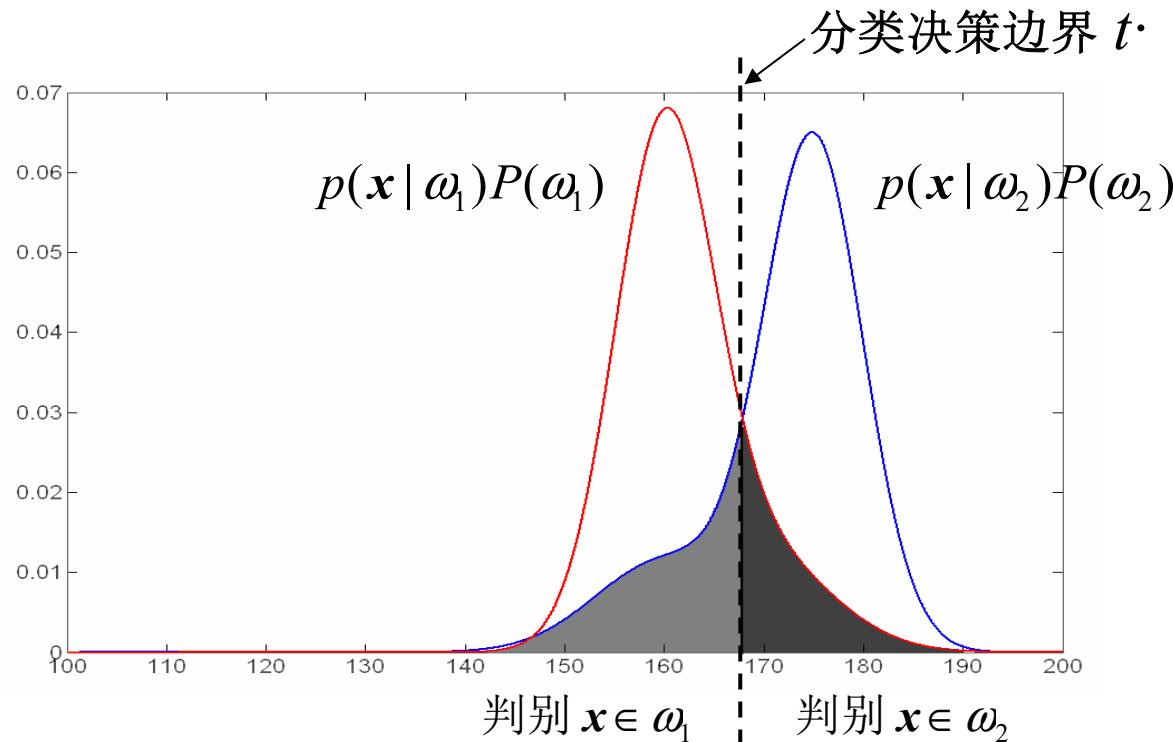
如果 $i = \arg \max_{j=1, \dots, c} P(\omega_j | x)$, 则判别 $x \in \omega_i$

Important

最小错误率准则贝叶斯分类器

- 后验概率需要使用贝叶斯公式由先验概率和条件概率密度间接计算：

如果 $i = \arg \max_{j=1, \dots, c} g_j(x) = p(x | \omega_j)P(\omega_j)$, 则判别 $x \in \omega_i$

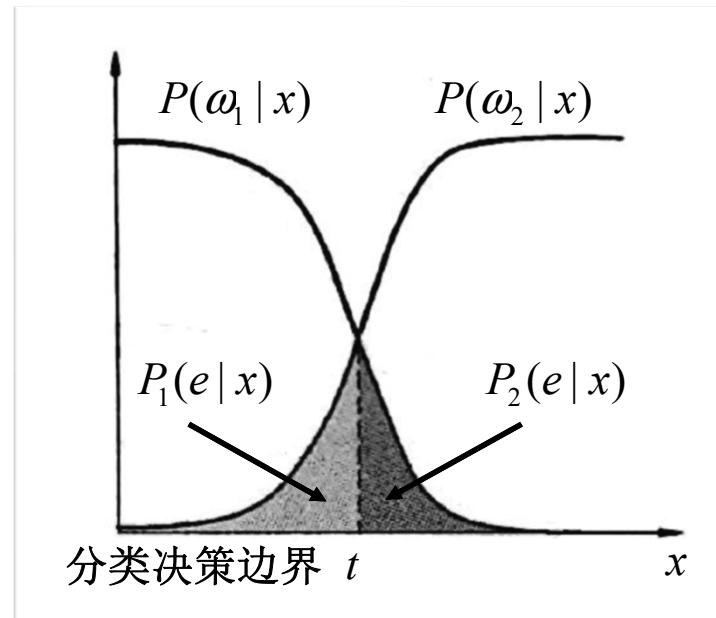


Important

最小错误率准则贝叶斯分类器

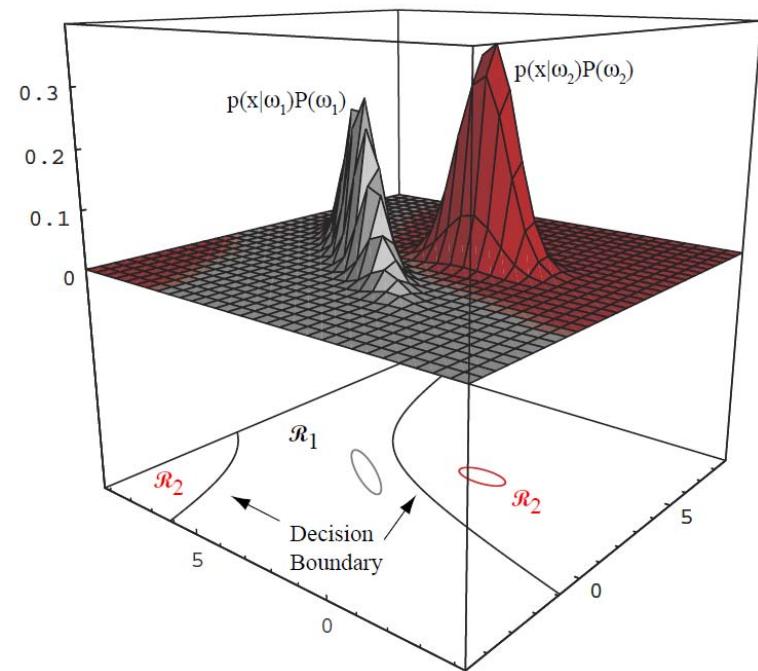
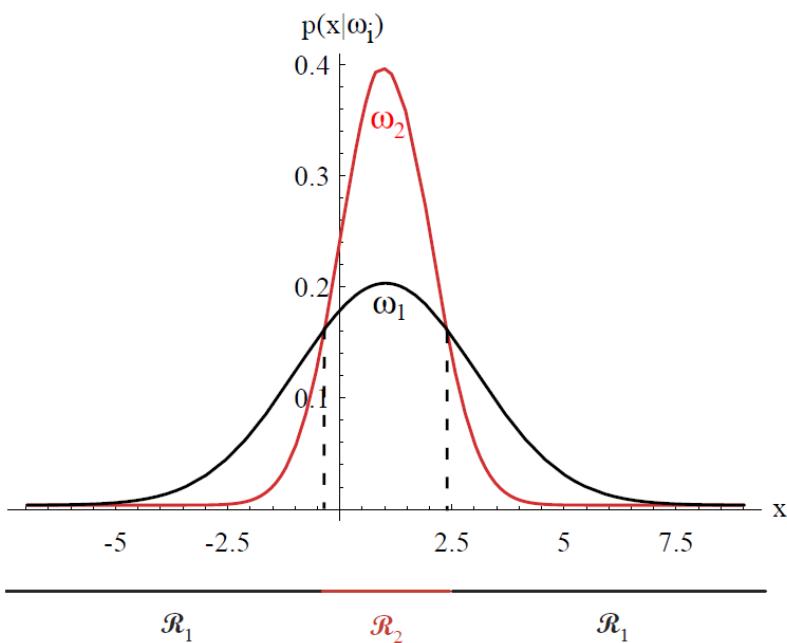
- 最小错误率准则贝叶斯分类器发生分类错误的概率可以计算如下：

$$\begin{aligned} P(e | x) &= P_1(e | x) + P_2(e | x) \\ &= \int_{-\infty}^t P(\omega_2 | x) dx + \int_t^{+\infty} P(\omega_1 | x) dx \end{aligned}$$



最小错误率准则贝叶斯分类器

- 注意：分类决策边界并非总是直线，也并非总是连通，与类别的分布有关（下节课介绍）



最近邻与贝叶斯分类的正确率

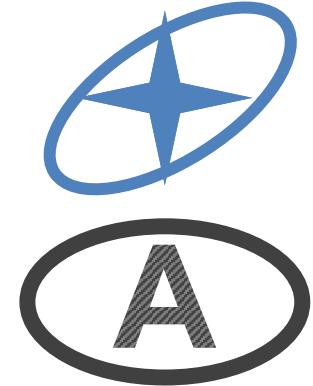
- 最近邻法/K近邻错误率和贝叶斯错误率的关系：
最近邻/K近邻法的错误率上下界是在一倍到两倍
贝叶斯决策方法的错误率范围内。
 - 最近邻规则导致的错误率大于最小错误率（即
贝叶斯错误率）。
 - 在有无限训练样本的情况下，最近邻分类的错
误率不会超过贝叶斯分类错误率的两倍。

延伸阅读：

<https://www.math.pku.edu.cn/teachers/ganr/course/pr2010/06.pdf>

贝叶斯分类实例

- 例：交通事故肇事逃逸事件
 - 目击者证词： 99%可能是下方车标
 - 特征： 目击者认成上方或下方车标
 - 类别： 真实车标， 上方或下方
 - 决策： 认成下方车标情况下， 真实车标是什么？
- 如果不考虑先验概率（或者认为两类车先验概率一样， 即市场占有率一样）， 则真实车标为下方的概率为99%。
- 如果考虑先验概率（上方车市场占有率99%， 下方1%）， 则真实车标为下方的概率为50%。



最小错误率准则贝叶斯分类器

- 例：对大批人进行某种癌症的普查，设 ω_1 类代表患癌症， ω_2 类代表正常人。已知先验概率：

$$P(\omega_1) = 0.005, \quad P(\omega_2) = 0.995$$

以一个化验结果作为特征 x : {阳性/阴性}，患癌症的人和正常人化验结果为阳性的概率分别为：

$$P(\text{阳性} | \omega_1) = 0.95, \quad P(\text{阳性} | \omega_2) = 0.01$$

现有一人化验结果为阳性，根据最小错误贝叶斯准则判断此人是否患癌症？

最小错误率准则贝叶斯分类器

- 解：利用最小错误贝叶斯准则计算

$$g_1(x) = P(x = \text{阳性} | \omega_1)P(\omega_1) = 0.95 \times 0.005 = 0.00475$$

$$g_2(x) = P(x = \text{阳性} | \omega_2)P(\omega_2) = 0.01 \times 0.995 = 0.00995$$

- 因为 $g_2(x) > g_1(x)$ ，所以判断该人正常。
- 思考：现实中医生不可能将化验指标为阳性的人诊断为正常。为什么？

最小平均风险准则贝叶斯分类器

- 以最小错误率为准则的贝叶斯分类器对于某些问题不合适，因为误判的后果可能不一样：有些类别的样本被误判的后果严重，有些类别的样本被误判的后果不严重。
- 对于某些模式识别问题，不仅需要考虑结果是否正确，也要考虑误判的后果和风险。
- 对于此类问题，需要引入**风险值**评估误判所付出的代价。

最小平均风险准则贝叶斯分类器

- 令 λ_{ij} 为将 ω_i 类的样本判别为 ω_j 类所承担的风险或付出的代价。如果识别正确， λ_{ij} 为很小的值或等于零。
- 如果分类器将待识别模式 x 识别为 ω_j 类，而 x 的真实类别属性可能是 $\omega_1, \dots, \omega_c$ 中的任何一个。做出这个判别的平均风险 $\gamma_j(x)$ 是每个类别的样本被分类为 ω_j 所带来的风险由后验概率加权求和：

$$\gamma_j(x) = \sum_{i=1}^c \lambda_{ij} P(\omega_i | x)$$

最小平均风险准则贝叶斯分类器

- 为了使由于误判而蒙受的损失最小，根据最小平均风险准则构建的贝叶斯分类器为：

如果 $k = \arg \min_{j=1, \dots, c} \gamma_j(\mathbf{x})$, 则判别 $\mathbf{x} \in \omega_k$

- 上式中的后验概率由先验概率和条件概率密度间接计算后，可得最小平均风险贝叶斯判别准则：

如果 $k = \arg \max_{j=1, \dots, c} g_j(\mathbf{x})$, 则判别 $\mathbf{x} \in \omega_k$

$$\text{其中 } g_j(\mathbf{x}) = -\sum_{i=1}^c \lambda_{ij} p(\mathbf{x} | \omega_i) P(\omega_i)$$

最小平均风险准则贝叶斯分类器

- 如果风险值是0-1损失函数（正确决策0损失，错误决策损失为1），最小平均风险贝叶斯决策就等价于最小错误率贝叶斯决策。

$$\lambda_{ij} = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases}$$

- 可以把最小错误率决策看成是最小平均风险决策的一个特例。由于损失函数的调整会造成不同的分类结果，当两类错误决策所造成的错误相差悬殊时，损失就会起到主导作用。

最小平均风险准则贝叶斯分类器

- 例：在前例癌症普查中引入判别风险

$$\lambda_{11} = \lambda_{22} = 0, \lambda_{12} = 100, \lambda_{21} = 25$$

根据最小平均风险准则判断此人是否患癌症？

- 解：

$$\begin{aligned}g_1(x) &= -\lambda_{11}p(x=\text{阳性} | \omega_1)P(\omega_1) - \lambda_{21}p(x=\text{阳性} | \omega_2)P(\omega_2) \\&= -0.24875\end{aligned}$$

$$\begin{aligned}g_2(x) &= -\lambda_{12}p(x=\text{阳性} | \omega_1)P(\omega_1) - \lambda_{22}p(x=\text{阳性} | \omega_2)P(\omega_2) \\&= -0.475\end{aligned}$$

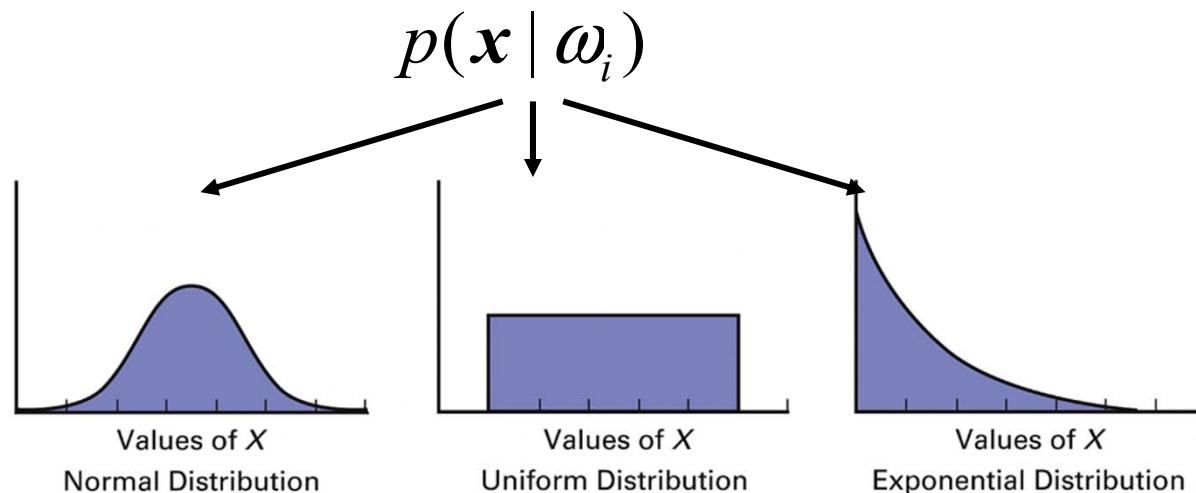
- 因为 $g_1(x) > g_2(x)$ ，所以判断该人患癌症。

贝叶斯分类器与高斯分布

- 贝叶斯分类器是根据类条件概率密度和先验概率来判别样本的类别属性。

$$P(\omega_i | x) = \frac{p(x | \omega_i)P(\omega_i)}{p(x)}$$

- 因此，构建贝叶斯分类器时必须确定每个类别的先验概率，并且估计类别的条件概率密度。



贝叶斯分类器与高斯分布

- 高斯分布（或正态分布）常被用于作为贝叶斯分类器的概率模型。
 - 高斯分布是自然界中最常见的概率分布形式，它呈现对称的单峰分布，大多数样本都会聚集在中央。
 - 高斯分布的参数简单，只需要用均值和标准差/协方差矩阵描述。
 - 中心极限定理：随机变量序列的分布会收敛于正态分布，即当样本量足够大时，样本均值的分布慢慢变成正态分布，这与样本总体的数据无关。

高斯分布

- 单变量 x 的高斯密度函数

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

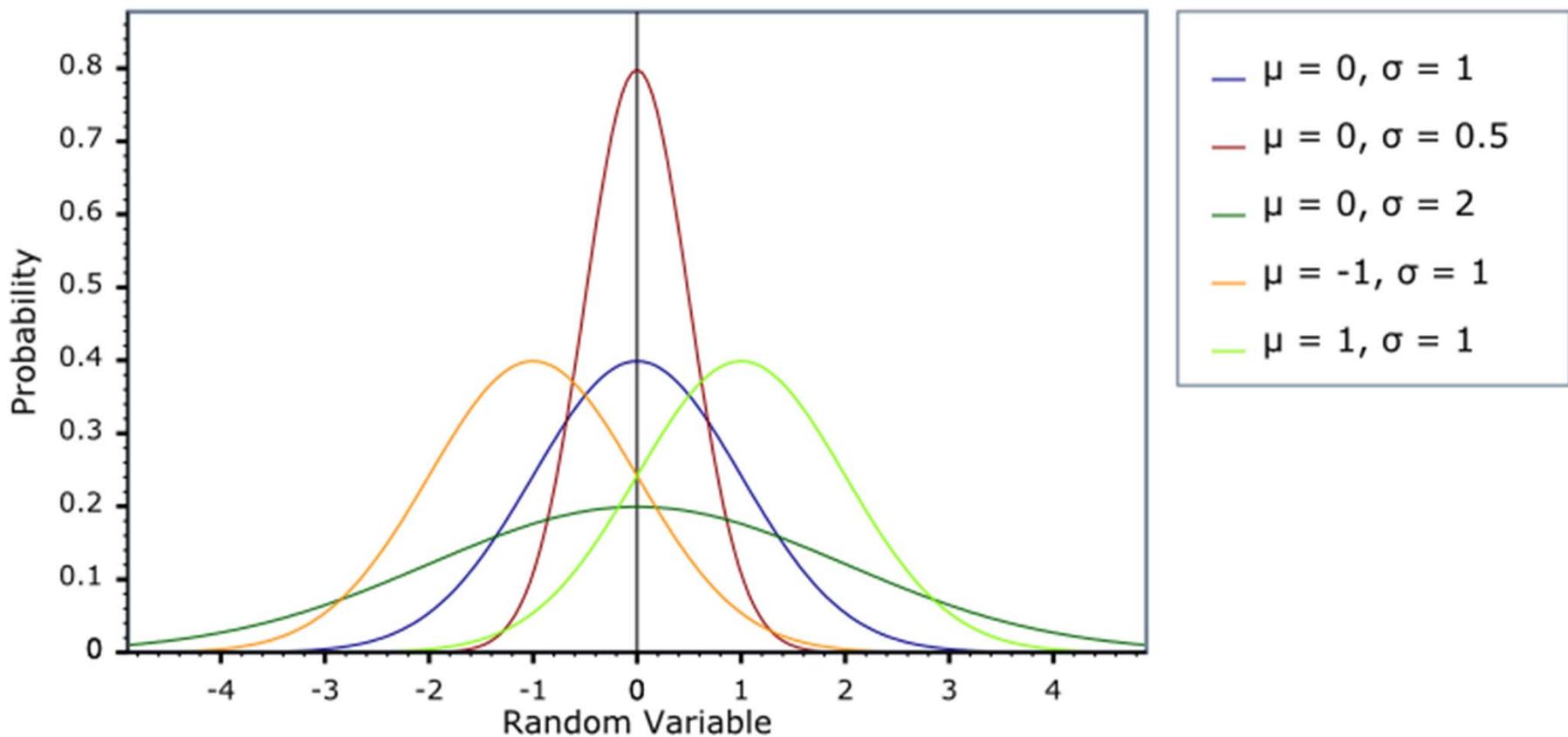
其中 μ 是分布的均值， σ^2 为分布的方差。

- 如果有 n 个服从此高斯分布的样本 x_1, \dots, x_n ，则可以对均值和方差做出估计：

$$\mu \approx \frac{1}{n} \sum_{i=1}^n x_i, \quad \sigma^2 \approx \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

高斯分布

Normal Distribution PDF



高斯分布

- 多变量 $x \in \mathbb{R}^d$ 的高斯密度函数

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

其中 μ 是均值矢量， Σ 是协方差矩阵。

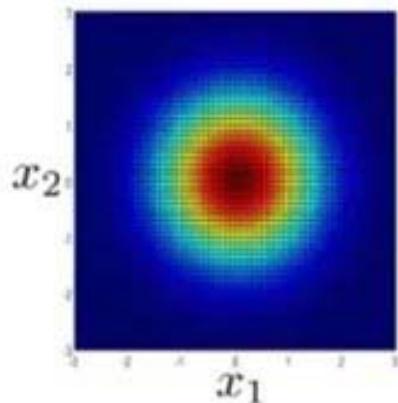
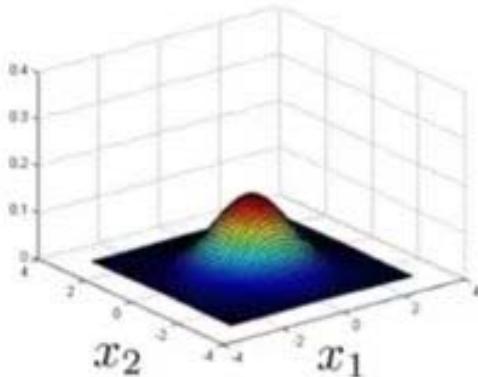
- 如果有 n 个服从此高斯分布的样本 x_1, \dots, x_n ，则可以对均值矢量和协方差矩阵做出估计：

$$\mu \approx \frac{1}{n} \sum_{i=1}^n x_i, \quad \Sigma \approx \frac{1}{n} \sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T$$

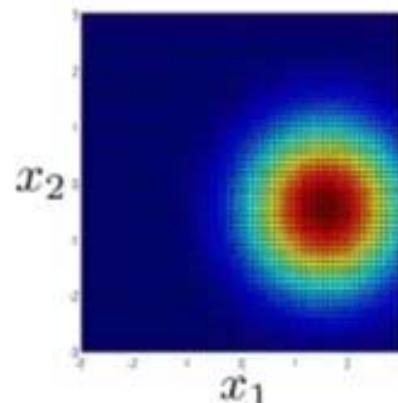
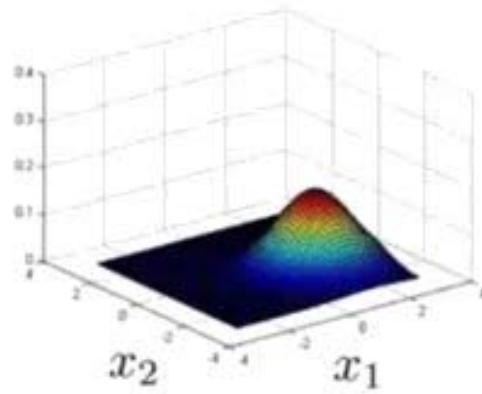
高斯分布

- 例：二维高斯分布（ bivariate Gaussian ）

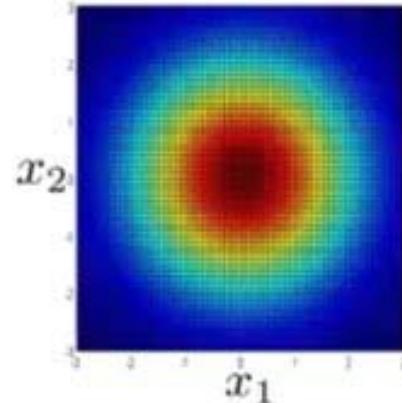
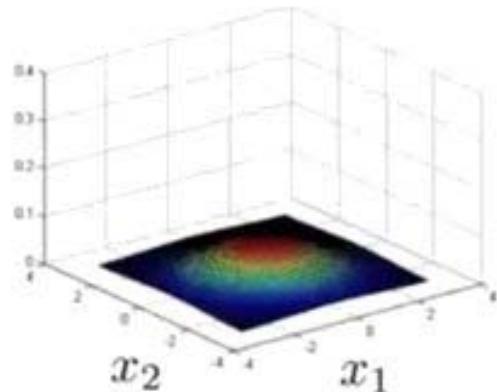
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 1.5 \\ -0.5 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



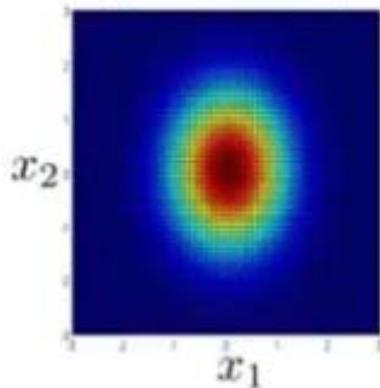
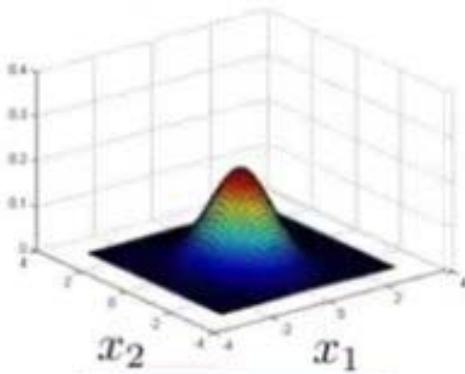
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$



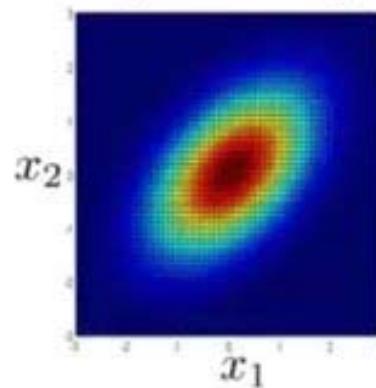
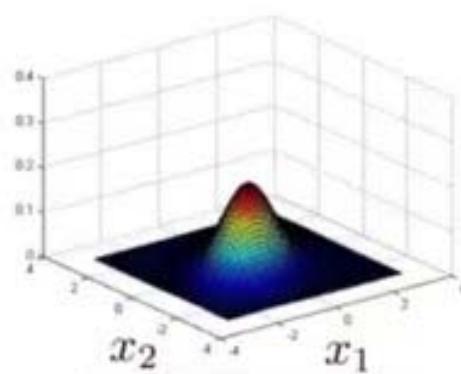
高斯分布

- 例：二维高斯分布（ bivariate Gaussian ）

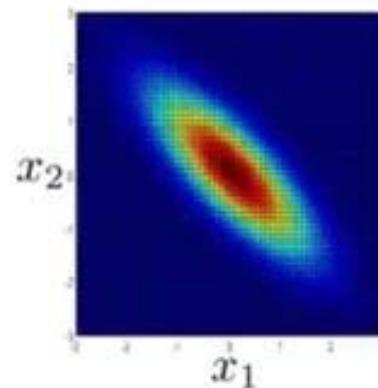
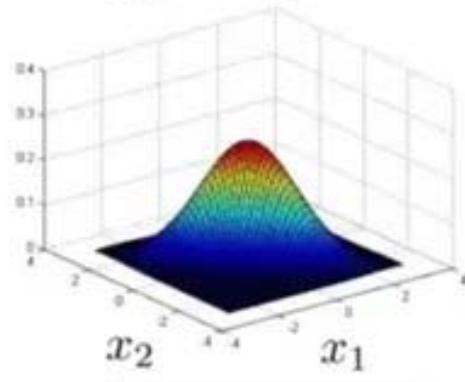
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$



高斯分布下的贝叶斯分类

- 最小错误率贝叶斯分类器
 - 分类决策规则简单
 - 类条件概率分布的形式不确定
 - 很难找到形式简单的分类决策边界
- 简化的最小错误率贝叶斯分类器
 - 各类样本特征向量取值的类条件概率满足高斯分布

高斯分布下的贝叶斯分类

- 最小错误率贝叶斯分类器判别函数：

$$g_i(x) = P(\omega_i | x) \propto p(x | \omega_i)P(\omega_i)$$

- 假设样本空间被划分为 c 个类别决策区域，则分类判决规则为

$$g_i(x) > g_j(x), j = 1, 2, \dots, c, j \neq i \iff x \in \omega_i$$

- 决策边界方程：

$$g_i(x) - g_j(x) = 0$$

高斯分布的判别函数

- 为计算方便，可将判别函数 $g_i(x) = p(x | \omega_i)P(\omega_i)$ 取对数（不会影响判别函数值的大小关系）：

$$\begin{aligned} g_i(x) &= \ln[p(x | \omega_i)P(\omega_i)] \\ &= \ln p(x | \omega_i) + \ln P(\omega_i) \end{aligned}$$

- 将高斯密度函数代入得： $p(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right]$

$$\begin{aligned} g_i(x) &= -\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i) \\ &\quad - \frac{d}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i) \end{aligned} \tag{公式 (*)}$$

高斯分布的判别函数

- 下面分几种情况讨论高斯分布判别函数的形式。
 - 情况1：类别先验概率相同
 - 1.1 特征不相关：协方差矩阵为相同的对角矩阵
 - 1.2 特征相关：协方差矩阵为相同的一般形式矩阵
 - 情况2：类别先验概率不同
 - 2.1 特征相关：协方差矩阵为相同的一般形式矩阵
 - 2.2 特征不相关：协方差矩阵为相同的对角矩阵
 - 情况3：一般情况

高斯分布的判别函数

- 情况1.1（类别先验概率相同，特征不相关）：

$$P(\omega_i) = 1/c, \Sigma_i = \sigma^2 I$$

- 类别先验概率相同，协方差矩阵为相同的对角矩阵，且对角线元素均为 σ^2 。
- 代入公式(*)中，忽略无关项，得

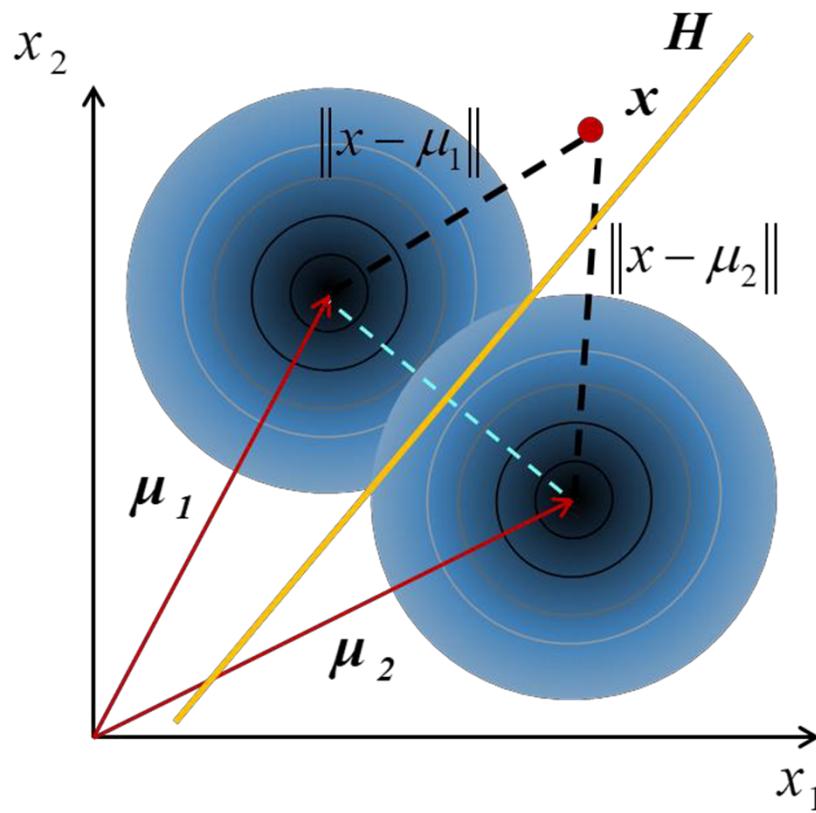
$$\begin{aligned} g_i(x) &= -\frac{1}{2\sigma^2} (x - \mu_i)^T (x - \mu_i) \\ &= -\frac{\|x - \mu_i\|^2}{2\sigma^2} \end{aligned}$$

高斯分布的判别函数

- 情况1.1的判别函数只同待识别样本 x 与每个类别均值矢量 μ_i 的欧氏距离有关：距离越近则判别函数值越大，距离越远则判别函数值越小。
- 因此，贝叶斯分类器会将 x 分类为与均值距离最近的类别。
- 实际上，情况1.1下的贝叶斯分类器就是单模板匹配距离分类器。

高斯分布的判别函数

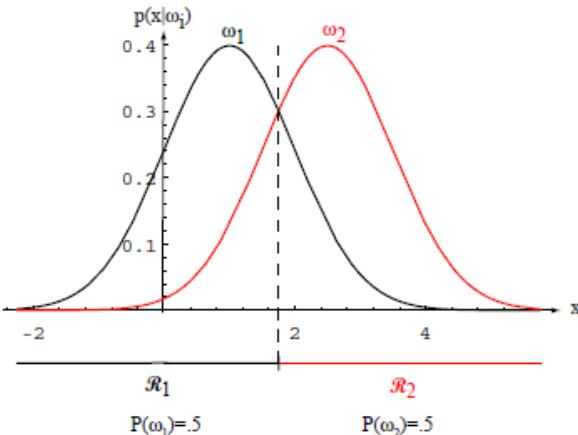
- 情况1.1（类别先验概率相同，特征不相关）：区分两类样本



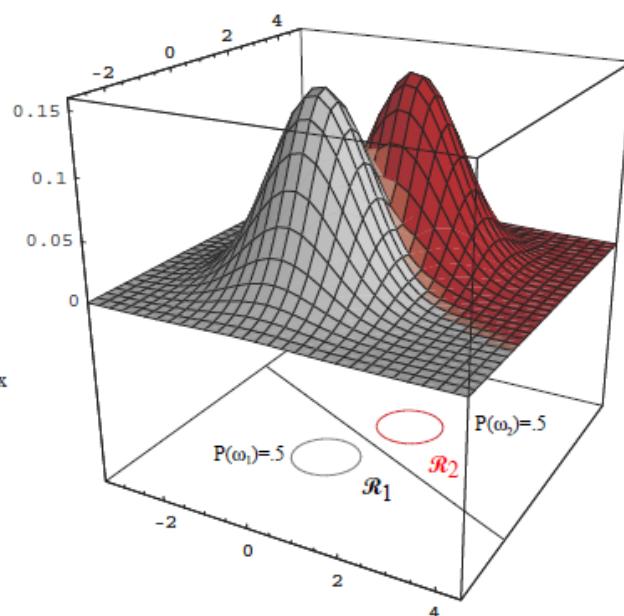
分类决策边界是两个类均值向量之间的垂直平分线

高斯分布的判别函数

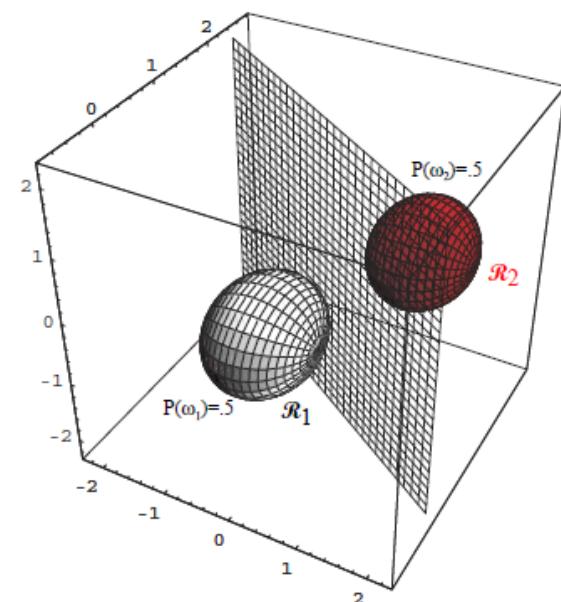
- 情况1.1（类别先验概率相同，特征不相关）：区分两类样本



1维空间



2维空间



3维空间

高斯分布的判别函数

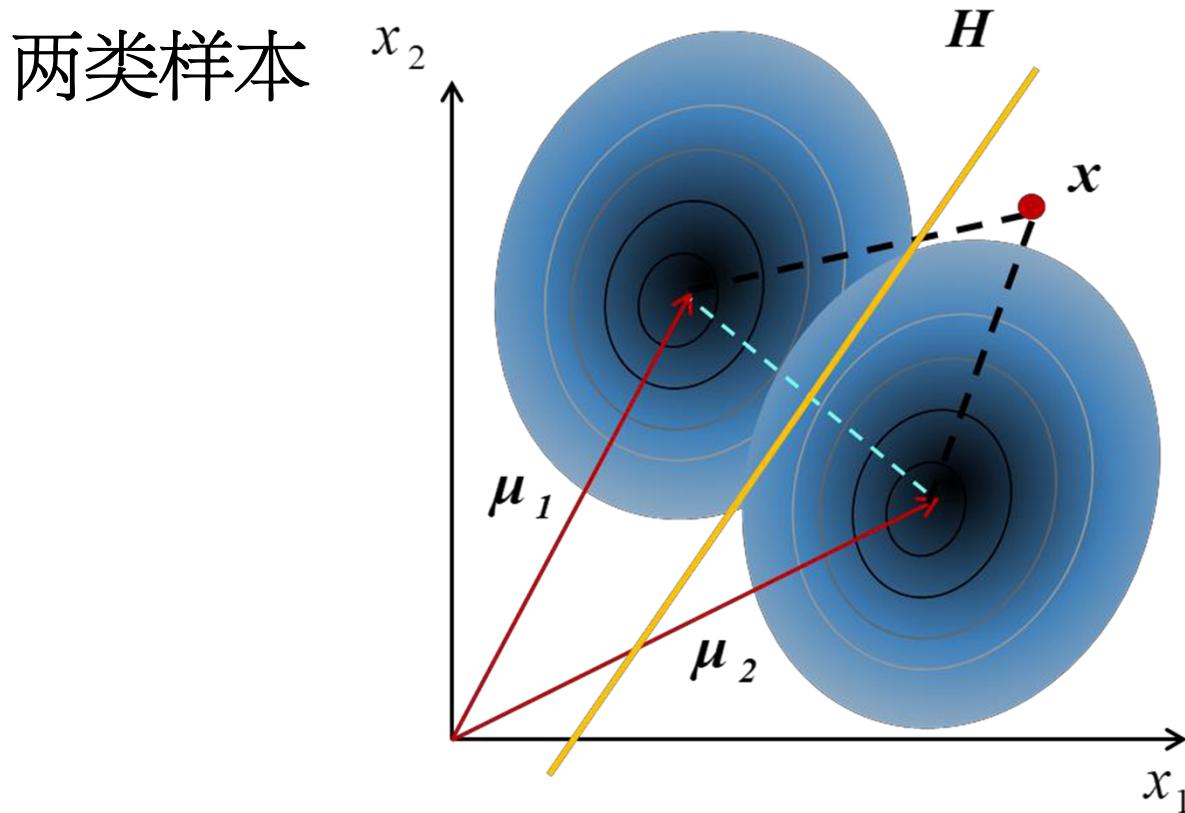
- 情况1.2（类别先验概率相同，特征相关）：

$$P(\omega_i) = 1/c, \Sigma_i = \Sigma$$

- 类别先验概率相同，协方差矩阵相同，但未必是对角矩阵。
- 代入公式(*)中，忽略无关项，得
$$g_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma^{-1} (x - \mu_i)$$
- 贝叶斯分类器仍然是一个距离分类器，但距离度量是马氏距离（而不是欧氏距离）。

高斯分布的判别函数

- 情况1.2（类别先验概率相同，特征相关）：区分两类样本



分类决策边界不再与两个类别均值向量之间的连线垂直，但是会通过两个均值向量连线的中点。

高斯分布的判别函数

- 情况2.1（类别先验概率不同，特征相关）：

$$\Sigma_i = \Sigma$$

- 类别先验概率不同，分布的均值不同，但是协方差矩阵相同（未必是对角矩阵）。
- 代入公式(*)中得

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma^{-1}(x - \mu_i) + \ln P(\omega_i)$$

$$= \boxed{-\frac{1}{2} x^T \Sigma^{-1} x} + \boxed{\mu_i^T \Sigma^{-1} x} \boxed{-\frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \ln P(\omega_i)}$$

与类别无关，可移除

w_i

w_{i0}

高斯分布的判别函数

- 上式中, $-\frac{1}{2} \mathbf{x}^T \Sigma^{-1} \mathbf{x}$ 与类别无关, 可忽略。
- 令 $\mathbf{w}_i = \Sigma^{-1} \boldsymbol{\mu}_i$, $w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^T \Sigma^{-1} \boldsymbol{\mu}_i + \ln P(\omega_i)$

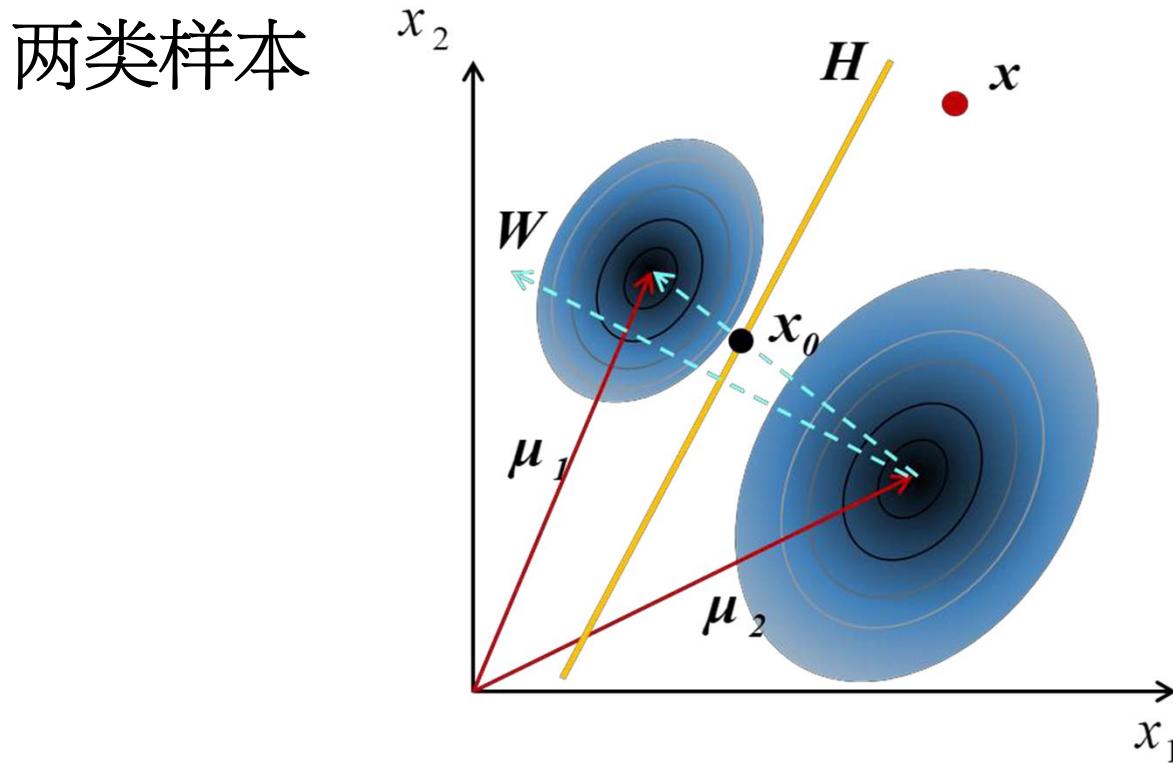
则上式变为

$$g_i(x) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

- 在此情况下, 贝叶斯分类器转变为线性分类器,
- \mathbf{w}_i 是线性判别函数的权值, w_{i0} 是偏置。

高斯分布的判别函数

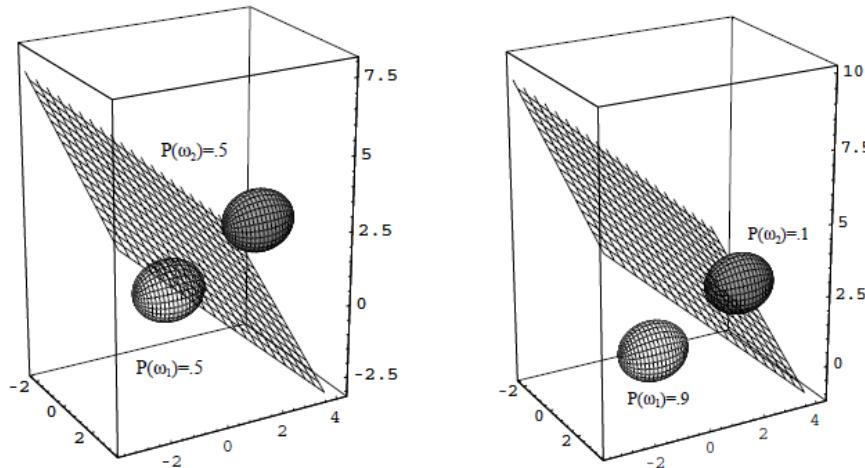
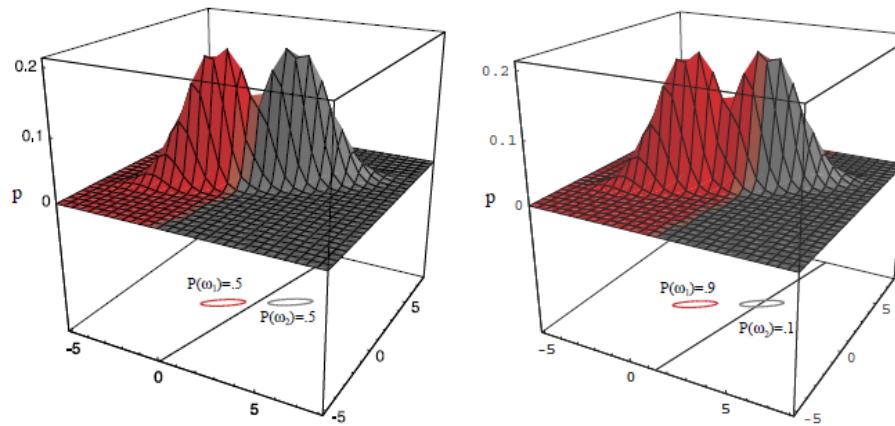
- 情况2.1（类别先验概率不同，特征相关）：区分两类样本



此时的分类器是在马氏距离基础上由先验概率进行修正的线性分类器。分类决策边界不与两个类别均值向量之间的连线垂直，也不会通过两个均值向量连线的中点，并且分类决策边界会偏向先验概率小的那一类。⁵⁶

高斯分布的判别函数

- 情况2.1（类别先验概率不同，特征相关）：区分两类样本



高斯分布的判别函数

- 情况2.2（类别先验概率不同，特征不相关）：

$$\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}$$

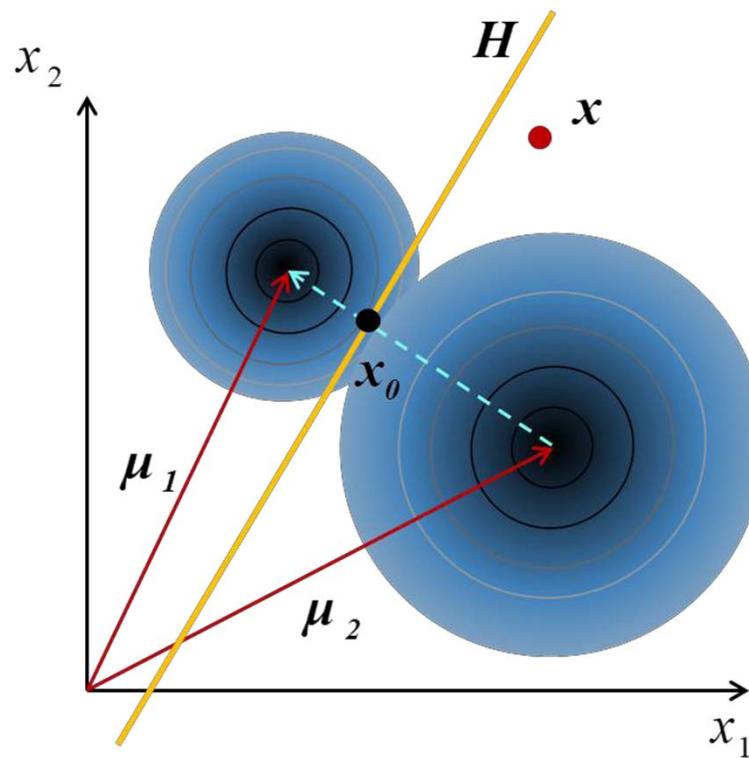
- 类别先验概率不同，协方差矩阵为相同的对角矩阵，且对角线元素均为 σ^2 。
- 类似地，可得线性判别公式

$$g_i(x) = \boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}$$

其中 $\boldsymbol{w}_i = \frac{1}{\sigma^2} \boldsymbol{\mu}_i, \quad w_{i0} = -\frac{1}{2\sigma^2} \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i + \ln P(\omega_i)$

高斯分布的判别函数

- 情况2.2（类别先验概率不同，特征不相关）：区分两类样本

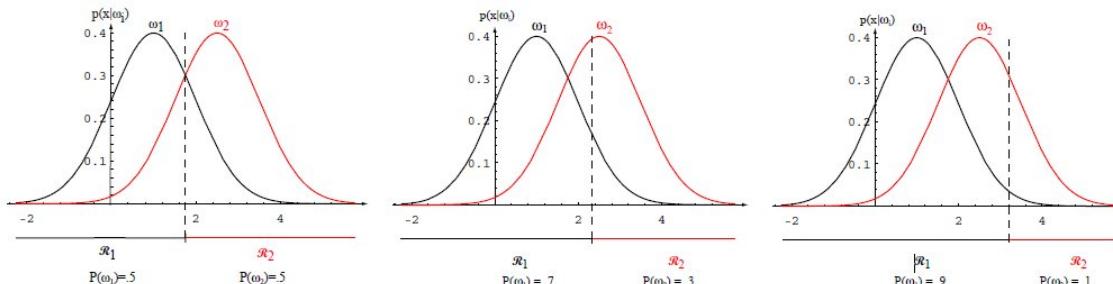


此时的分类器是在欧氏距离基础上由先验概率进行修正的线性分类器。分类决策边界垂直于两个类别均值向量之间的连线。分类决策边界会偏向先验概率小的那一类，使更多样本被识别为概率大的一类。

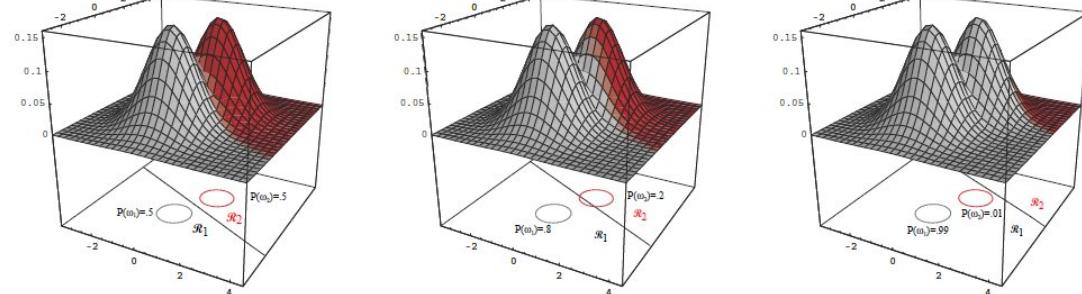
高斯分布的判别函数

- 情况2.2（类别先验概率不同，特征不相关）：区分两类样本

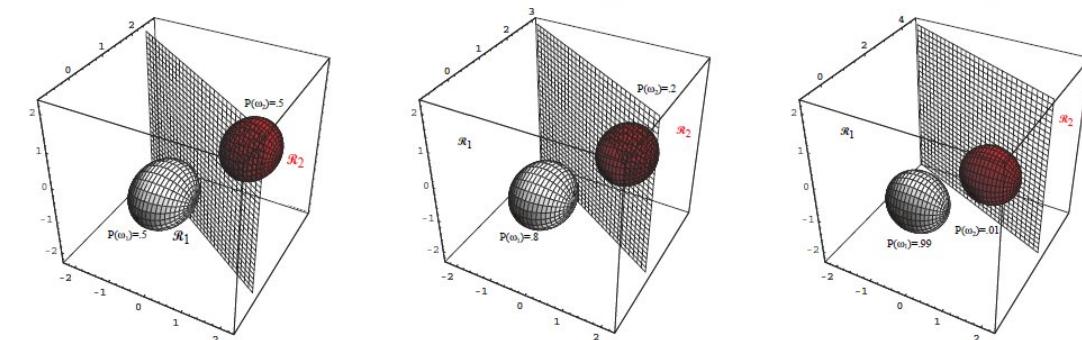
1维空间



2维空间



3维空间



高斯分布的判别函数

- 情况3：在一般情况下（参数可取任意合理值），将公式（*）展开并忽略无关项，得到

$$g_i(\mathbf{x}) = \underbrace{-\frac{1}{2} \mathbf{x}^T \Sigma_i^{-1} \mathbf{x}}_{\text{red line}} + \underbrace{\mu_i^T \Sigma_i^{-1} \mathbf{x}}_{\text{green line}} - \underbrace{\frac{1}{2} \mu_i^T \Sigma_i^{-1} \mu_i}_{\text{blue line}} - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i)$$

- 令 $W_i = -\frac{1}{2} \Sigma_i^{-1}$, $w_i = \Sigma_i^{-1} \mu_i$,

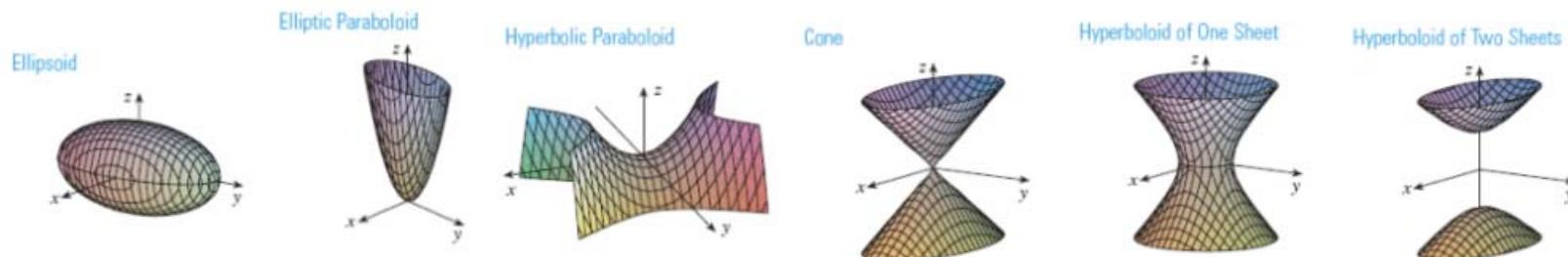
$$w_{i0} = -\frac{1}{2} \mu_i^T \Sigma_i^{-1} \mu_i - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i)$$

可得标准的二次判别函数

$$g_i(\mathbf{x}) = \mathbf{x}^T W_i \mathbf{x} + w_i^T \mathbf{x} + w_{i0}$$

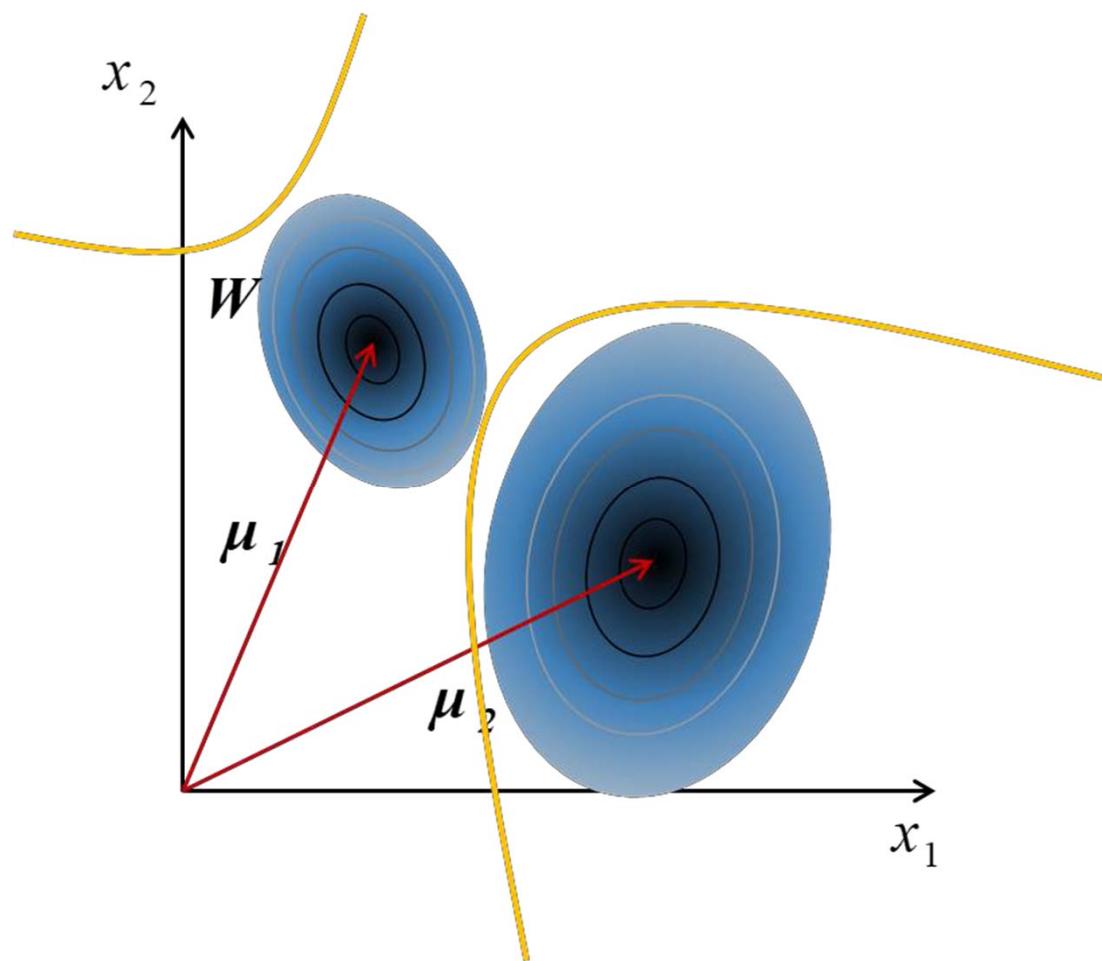
高斯分布的判别函数

- 此判别函数的二次项与类别有关，无法消去，因此，这是一个**非线性分类问题**。
- 这种情况下，分类决策边界是一个**超二次曲面**，随着参数的不同而呈现出为不同的形式，包括超球面、超抛物面、超双曲面或超平面等。



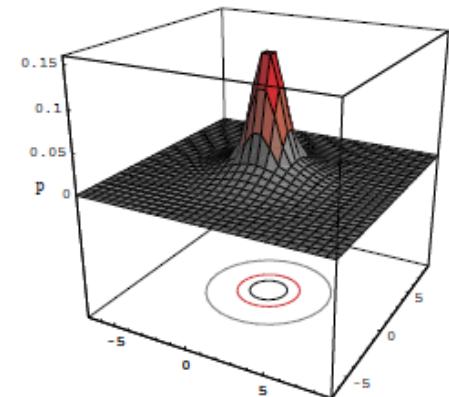
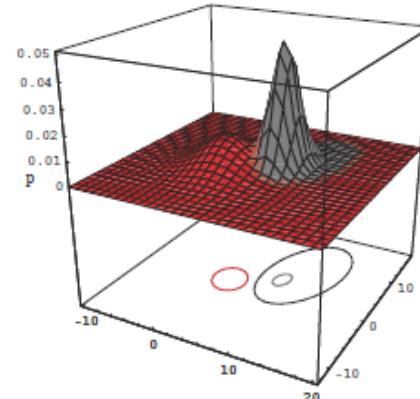
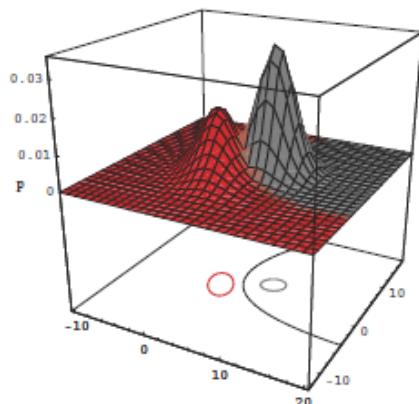
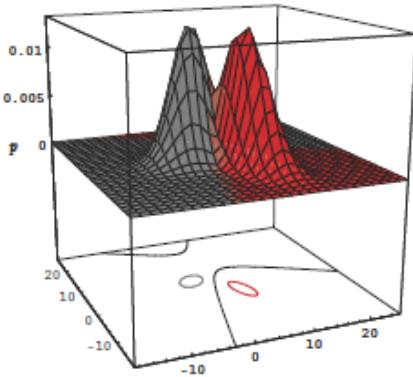
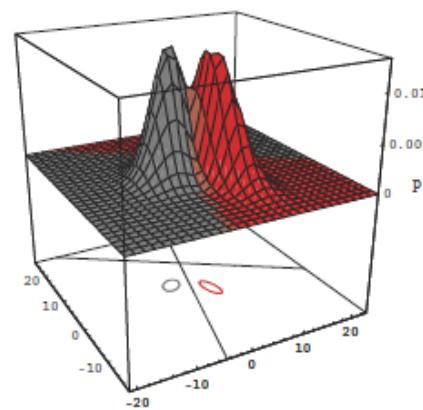
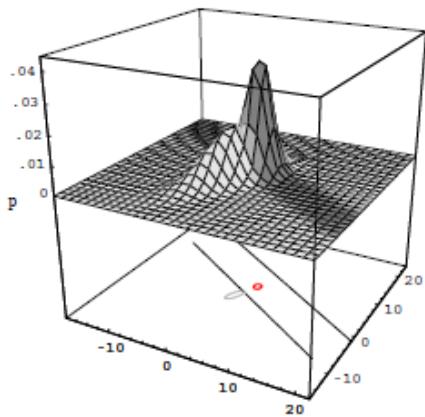
高斯分布的判别函数

- 情况3：区分两类样本



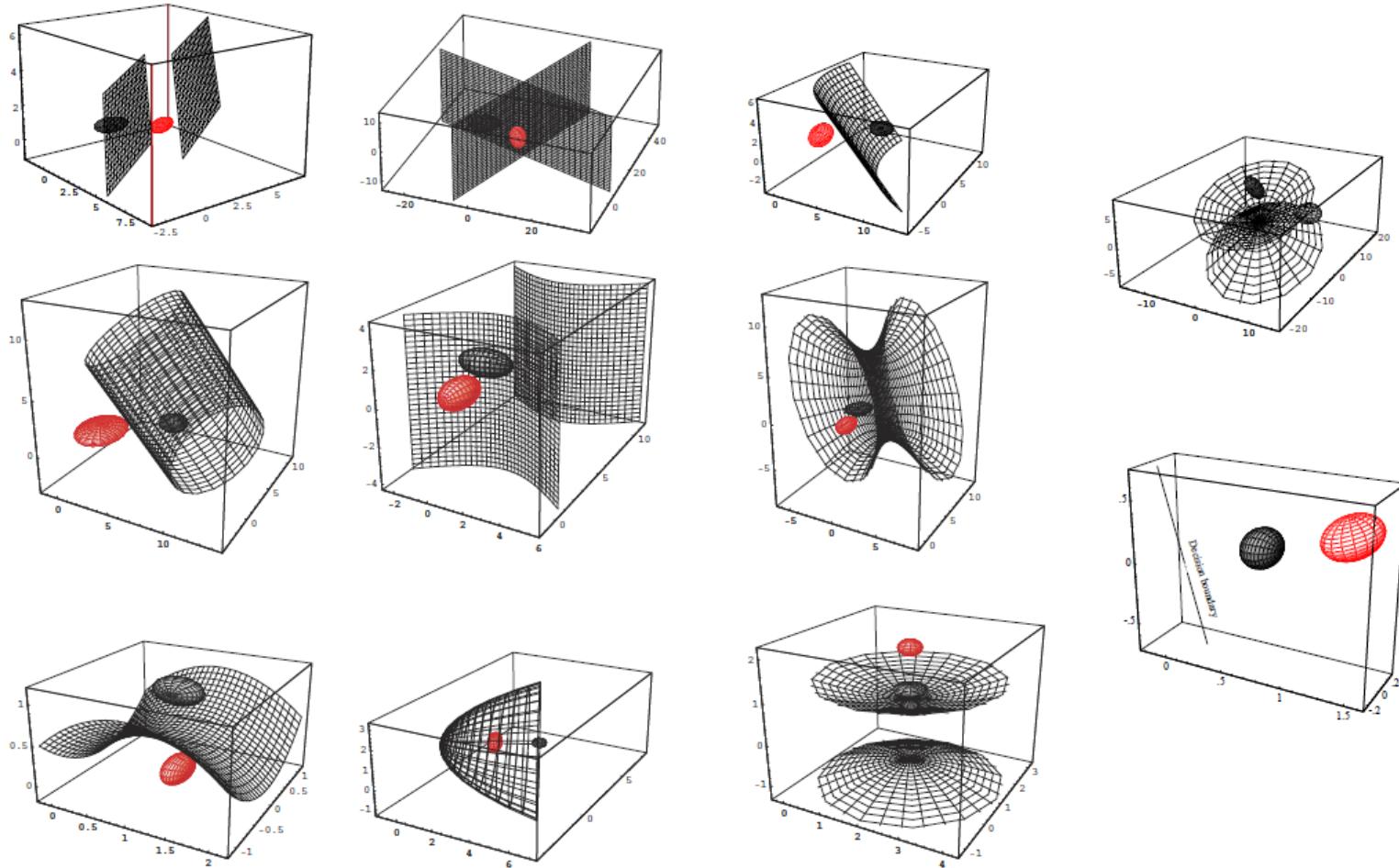
高斯分布的判别函数

- 情况3：任意高斯分布导致一般超二次曲面的贝叶斯判决边界



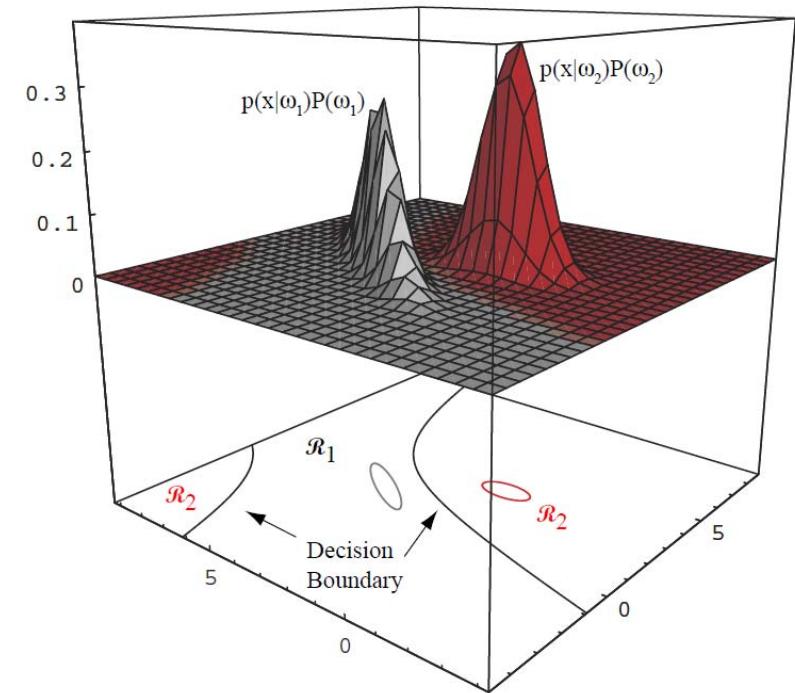
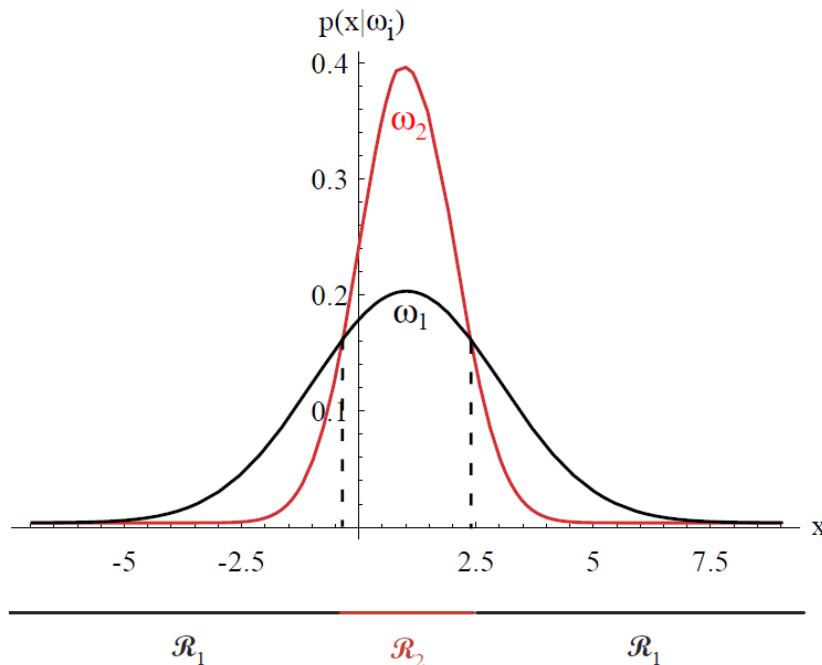
高斯分布的判别函数

- 情况3：任意三维高斯分布导致二维的超二次曲面的贝叶斯判决边界



高斯分布的判别函数

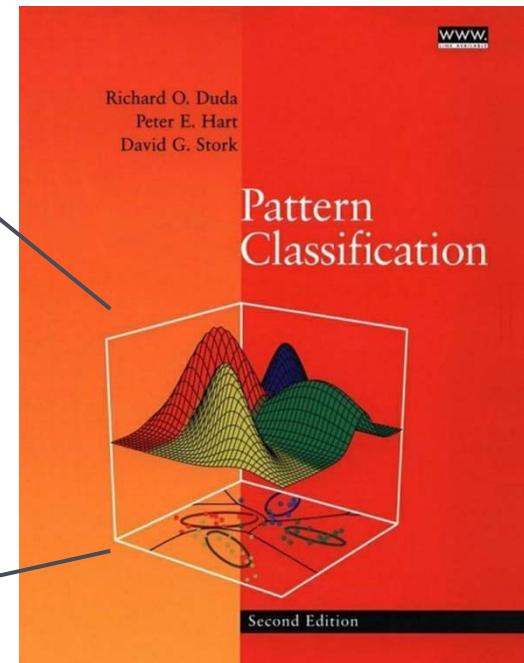
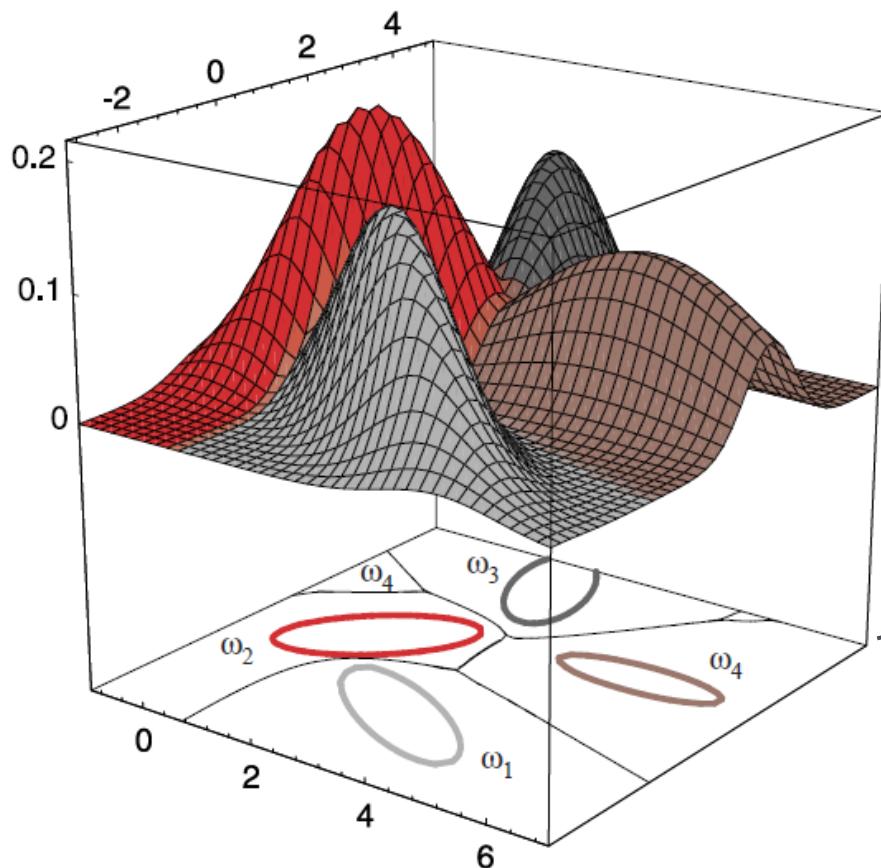
- 决策边界的复杂性：判决区域未必连通。



在方差不相等的一维高斯分布情况下，可能产生并非单连通的判决区域。

高斯分布的判别函数

- 决策边界的复杂性：简单的4个高斯分布的分类决策边界也非常复杂。



高斯分布的判别函数

- 例：高斯分布二次分类曲面

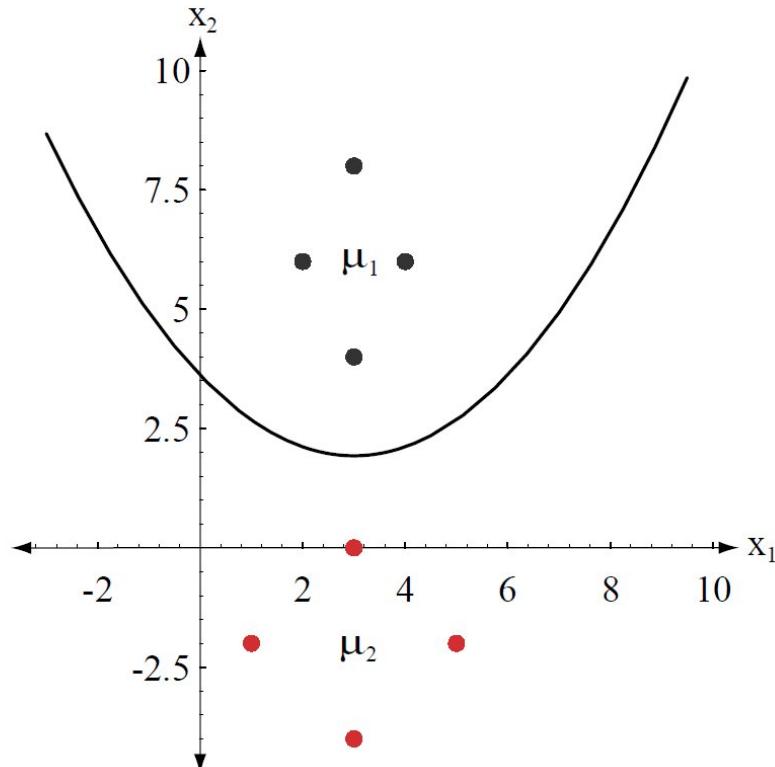
$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1/2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$\boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$P(\omega_1) = P(\omega_2) = 0.5$$

- 设 $g_1(x) = g_2(x)$ ，代入公式（28页）解得

$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2$$



思考：假如只有上述样本点，而无高斯分布参数，应该如何处理？

朴素贝叶斯分类器

- d 维特征的高斯分布 $p(x | \omega_i)$ 中，均值矢量和协方差矩阵共有 $(d^2 + 3d)/2$ 个参数。
- 如果特征维数较小而训练样本比较多，可以估计出每个类别的分布 $p(x | \omega_i)$ ，然后分类。
- 但是，当特征的维数较高，而每个类别的训练样本数量较少时，对高斯分布均值矢量和协方差矩阵的估计比较困难。从统计学的角度来说，使用少量样本来估计大量的参数，估计结果不可靠。

朴素贝叶斯分类器

- 解决多特征少样本贝叶斯分类的一种有效办法是采用朴素贝叶斯分类器（Naïve Bayes Classifier）。
- 朴素贝叶斯的一个基本假设是所有特征在类别已知的条件下是相互独立的，即

$$p(x | \omega_i) = p(x_1, \dots, x_d | \omega_i) = \prod_{j=1}^d p(x_j | \omega_i)$$

- 在构建分类器时，只需要逐个估计出每个类别的训练样本在每一维特征上的分布，就可以得到每个类别的条件概率密度，大大减少了需要估计参数的数量。

朴素贝叶斯分类器

- 朴素贝叶斯分类器可以根据具体问题来确定样本在每一维特征上的分布形式，最常用的假设是每一个类别的样本都服从各维特征之间相互独立的高斯分布：

$$p(\mathbf{x} | \omega_i) = \prod_{j=1}^d p(x_j | \omega_i) = \prod_{j=1}^d \left\{ \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp \left[-\frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2} \right] \right\}$$

其中 μ_{ij} 是第 i 类样本在第 j 维特征上的均值， σ_{ij}^2 是第 i 类样本在第 j 维特征上的方差。

朴素贝叶斯分类器

- 相应的对数判别函数为：

$$g_i(\mathbf{x}) = \ln p(\mathbf{x} | \omega_i) + \ln P(\omega_i)$$

$$= \sum_{j=1}^d \left[-\frac{1}{2} \ln(2\pi) - \ln \sigma_{ij} - \frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2} \right] + \ln P(\omega_i)$$

$$= -\frac{d}{2} \ln(2\pi) - \sum_{j=1}^d \ln \sigma_{ij} - \sum_{j=1}^d \frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2} + \ln P(\omega_i)$$

- 忽略无关项，得判别函数：

$$g_i(\mathbf{x}) = \ln P(\omega_i) - \sum_{j=1}^d \ln \sigma_{ij} - \sum_{j=1}^d \frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2}$$

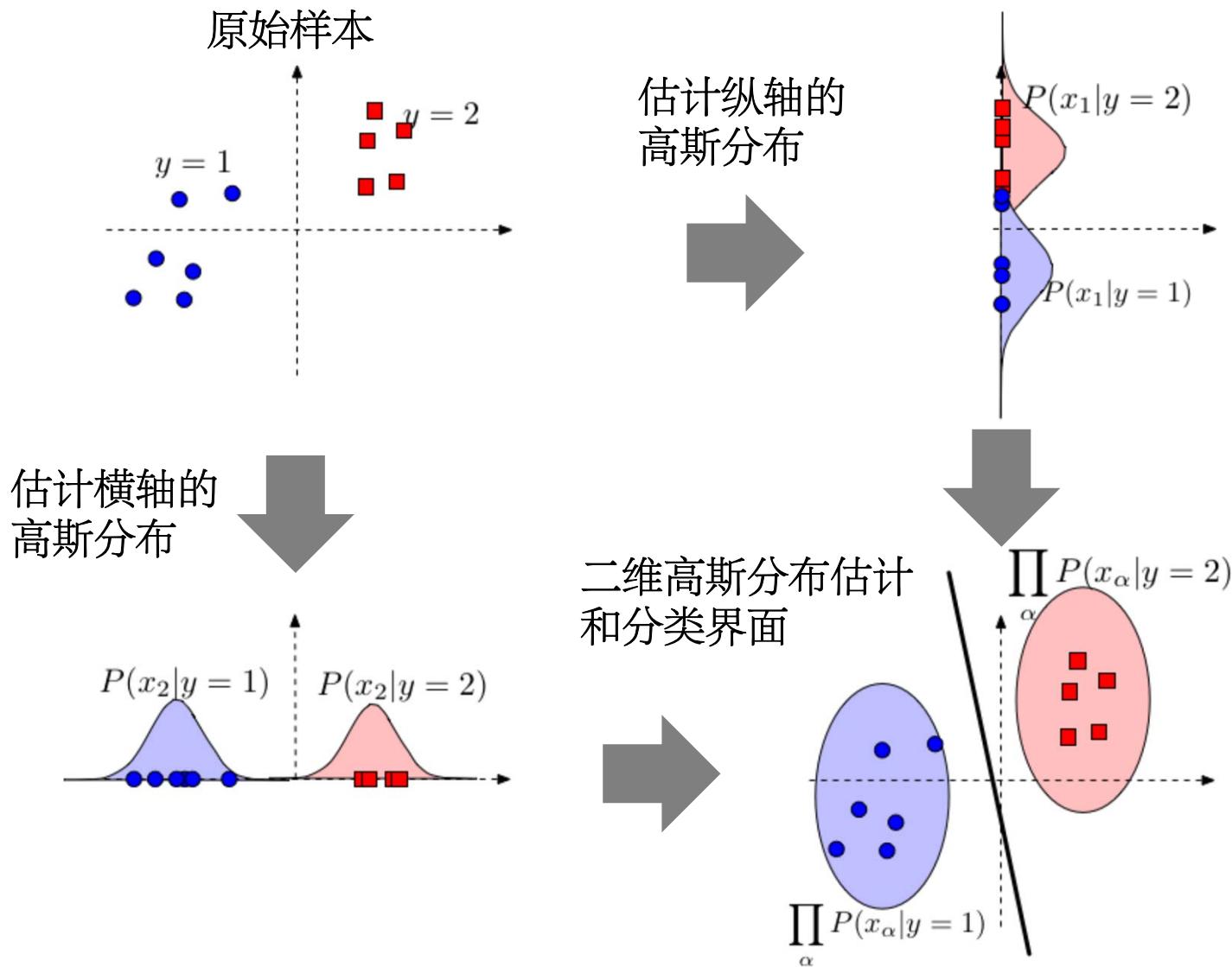
朴素贝叶斯分类器

- 朴素贝叶斯算法
- 分类器学习：由每个类别的训练样本估计出均值和方差 $\mu_{ij}, \sigma_{ij}^2, i = 1, \dots, c, j = 1, \dots, d$
- 分类判别：根据以下公式计算每个类别在待识样本 x 上的判别函数值：

$$g_i(x) = \ln P(\omega_i) - \sum_{j=1}^d \ln \sigma_{ij} - \sum_{j=1}^d \frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2}$$

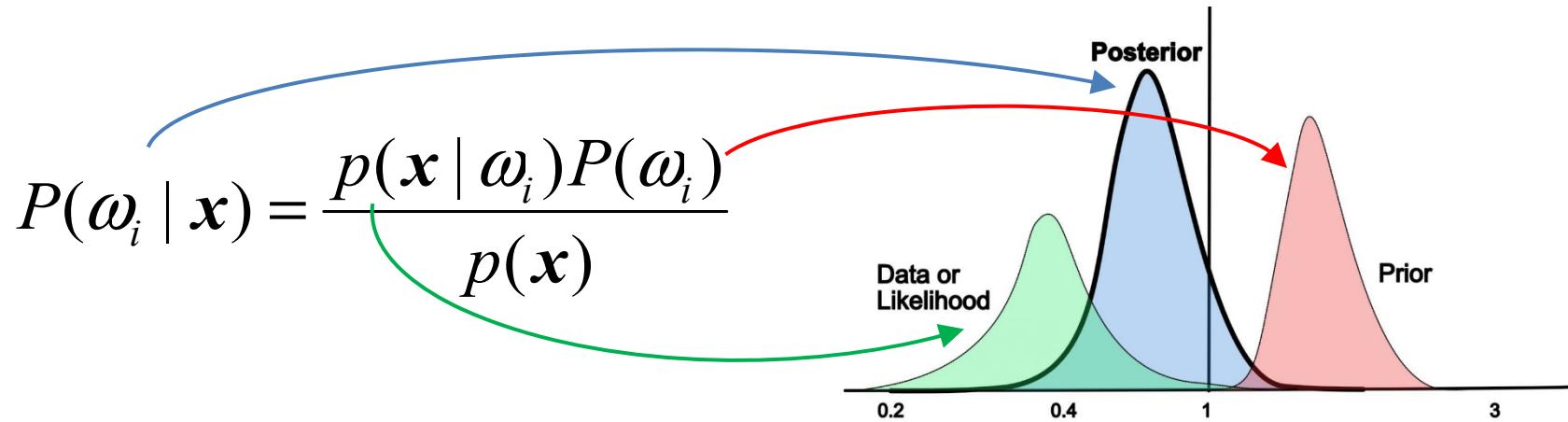
如果 $i = \arg \max_{j=1, \dots, c} g_j(x)$ ，则判别 $x \in \omega_i$ 。

朴素贝叶斯分类器



贝叶斯分类器与概率密度函数

- 贝叶斯分类器根据类条件概率密度和先验概率计算后验概率，进而判别样本的类别属性。



- 因此，贝叶斯分类器的学习主要是对类别的先验概率和类别的条件概率密度的估计。

贝叶斯分类器与概率密度函数

- 类别先验概率 $P(\omega_i)$ 的估计（相对较容易）：
 - 依靠经验
 - 用训练数据中各类出现的频率或比例估计
- 用频率或比例估计概率的优点：
 - 无偏性
 - 相合性
 - 收敛速度快

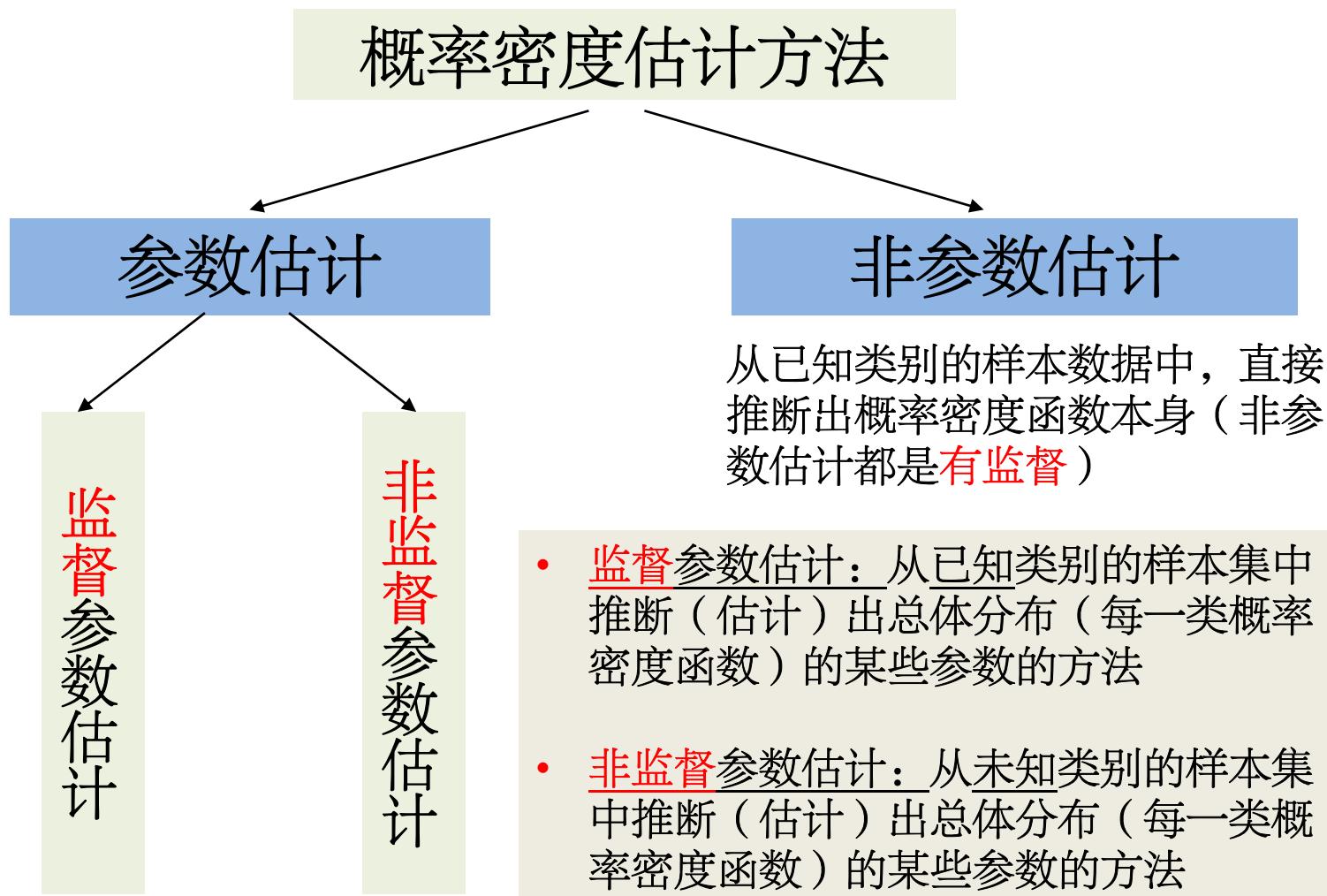
贝叶斯分类器与概率密度函数

- 类条件概率密度 $p(x | \omega_i)$ 的估计（相对较难）：
 - 概率密度函数包含一个随机变量的全部信息；
 - 概率密度函数可以是满足下面条件的任何一个函数： $p(x | \omega_i) \geq 0, \int p(x | \omega_i) dx = 1$
- 将某类别样本集合表示为 $D = \{x_1, \dots, x_n\}$ ，根据 D 估计概率密度函数 $p(x)$ 的基本假设是： D 中样本是独立抽样于 $p(x)$ 的同一分布（独立同分布，*independent & identically distributed, i.i.d.*）。

概率密度函数估计

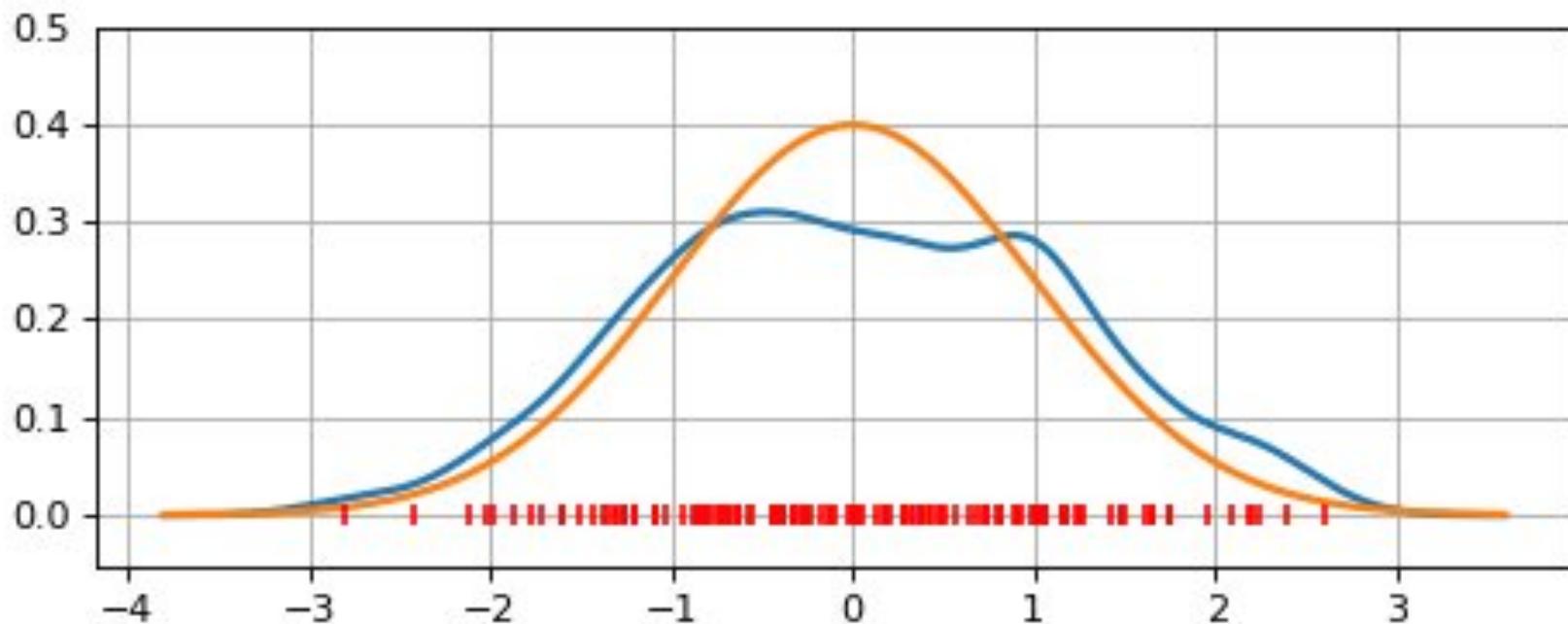
- 概率密度函数估计的两大类方法：非参数估计和参数估计。
- **参数估计**：根据对问题的一般性认识，假设随机变量服从某种分布形式，分布函数的参数通过训练数据来估计。
- **非参数估计**：不需要假设任何分布，只利用训练数据本身对概率密度函数做估计。

概率密度函数估计



概率密度函数估计

- 非参数估计和参数估计的实例比较：
 - 红点：用于估计概率密度函数的数据
 - 黄线：参数估计，假设数据服从高斯分布
 - 蓝线：非参数估计，对数据分布没有假设

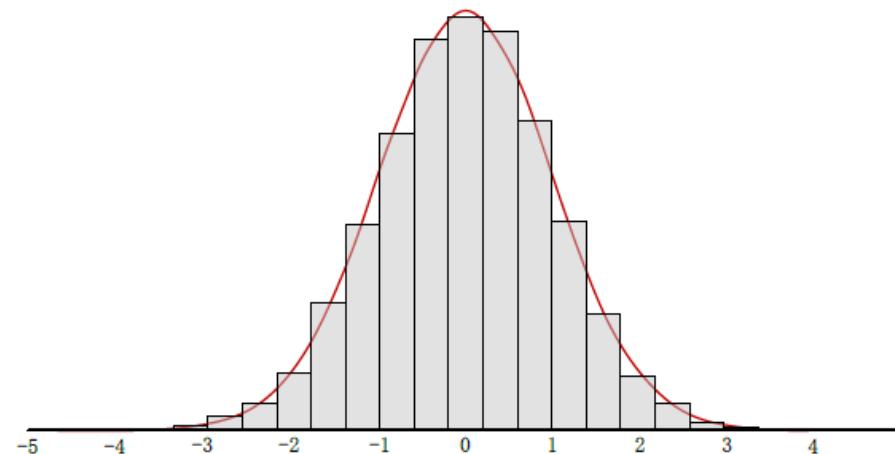


概率密度函数估计

参数估计	非参数估计
对分布函数有严格假设	对分布函数无任何假设
需要较少数据	需要较多数据
从数据估计概率密度函数的参数	从数据直接估计概率密度
结果依赖对分布的假设：如果对分布的假设准确，那么估计效果好；否则效果差	结果受估计方法参数选择的影响
计算相对较快	计算相对较慢
模型/假说驱动	数据驱动

概率密度函数的非参数估计

- 直方图 (histogram)：假设样本来自某分布，首先将样本分布空间划分为等间距的若干区间，统计每个区间包含的样本数，用柱状图画出，得到对真实概率密度分布（下图红线）的近似。



- 如何利用直方图近似计算概率密度函数？

概率密度函数的非参数估计

- 令 R 是 d 维空间中包含样本 x 的一个区域，
 n 个训练样本中有 k 个落入区域 R 的范围之内，
那么事件 “ x 出现在 R 中” 的概率估计如下：

$$P(x \in R) \approx k / n$$

- 假设在 R 中每一点的概率密度函数值相等，则 x 出现在 R 中的概率还可以估计为

$$P(x \in R) = \int_R p(x) dx = p(x) \int_R dx = p(x) \cdot V$$

其中 V 是区域 R 的体积。

概率密度函数的非参数估计

- 联合上两式 $P(x \in R) \approx k/n$ 和 $P(x \in R) = p(x) \cdot V$, 可以得到对概率密度函数的估计:

$$p(x) = \frac{k/n}{V}$$

- 非参数估计的过程:
 - 根据样本数 n 选择合适的区域 R ;
 - 计算区域的体积 V ;
 - 统计区域中包含的训练样本数量 k ;
 - 根据 $p(x) = \frac{k/n}{V}$ 计算 x 处的概率密度值。

概率密度函数的非参数估计

- 收敛性：对应样本数 $n = 1, 2, \dots$ ，构造一系列包含 x 的区域 $\mathbf{R}_1, \mathbf{R}_2, \dots$ ，则对 $p(x)$ 有一系列估计：

$$p_n(x) = \frac{k_n / n}{V_n}, \quad n = 1, 2, 3, \dots$$

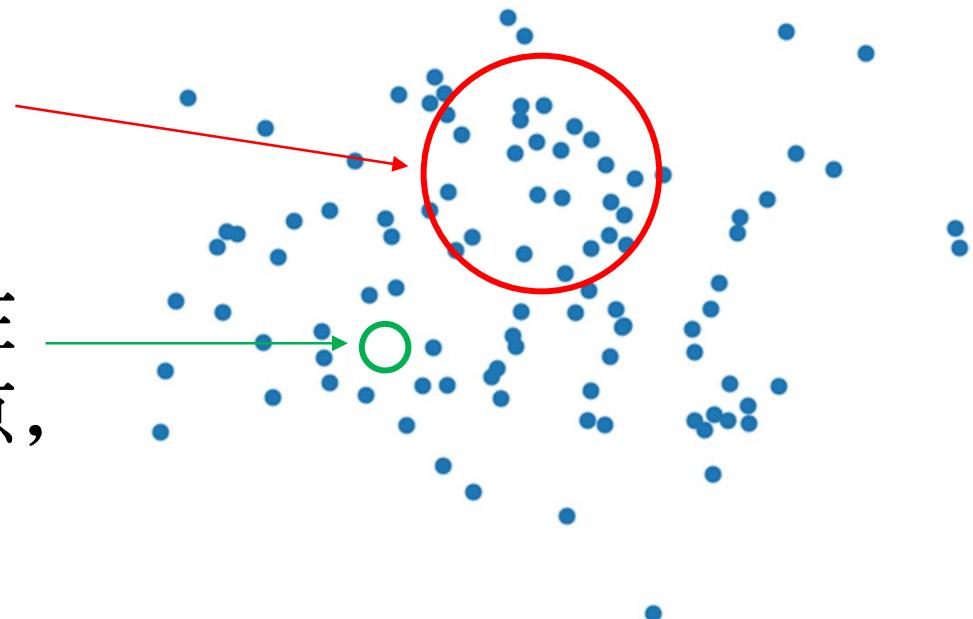
其中 V_n 是 \mathbf{R}_n 的体积， k_n 是落在 \mathbf{R}_n 内的样本数。

- 如果 \mathbf{R}_n 满足下列条件，那么 $p_n(x)$ 收敛于 $p(x)$ ：

- | | |
|---|---|
| 1) $\lim_{n \rightarrow \infty} V_n = 0$ | ✓ 随着样本数 n 的增多， \mathbf{R} 的体积 V 应该减小； |
| 2) $\lim_{n \rightarrow \infty} k_n \rightarrow \infty$ | ✓ 必须保证体积减小的同时，区域中包含的训练样本数 k 仍然增加； |
| 3) $\lim_{n \rightarrow \infty} k_n / n = 0$ | ✓ k 的增大速度应该小于 n 。 |

概率密度函数的非参数估计

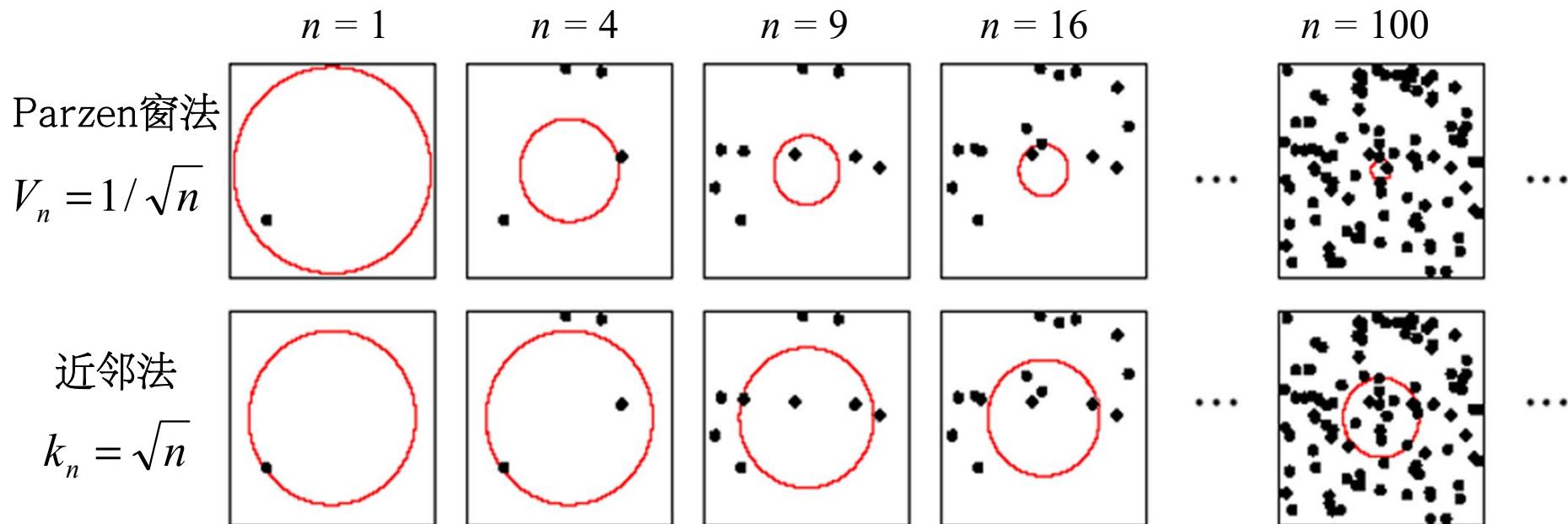
- 非参数估计 $p(x) = \frac{k/n}{V}$ 的有效性取决于样本数量 n 和区域体积 V 。
- 当 n 固定时， V 的大小对估计的效果影响很大：
 - 过大则平滑过多，估计不够精确；
 - 过小则可能导致在此区域内无样本点， $k = 0$ 。



Important

概率密度函数的非参数估计

- 区域选定的两个主要方法：
 - Parzen窗法：区域体积 V 是样本数 n 的函数
 - 近邻法：区域包含样本数 k 是样本数 n 的函数

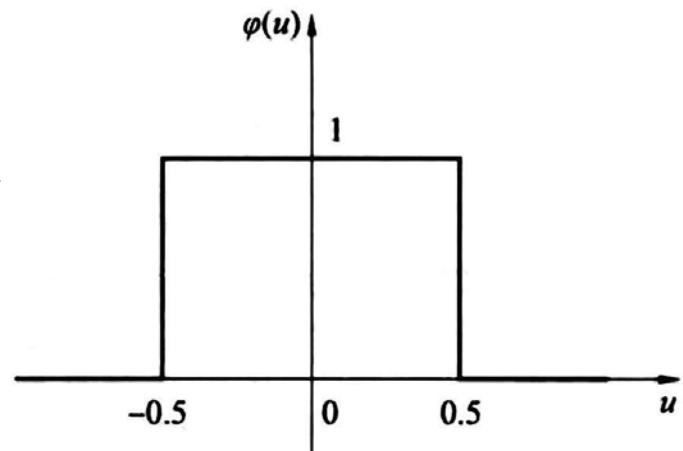


Parzen窗方法

- 为了用解析形式表示区域中包含样本数的技术过程，定义窗函数 $\varphi(\mathbf{u})$, $\mathbf{u} = (u_1, \dots, u_d)^T$

$$\varphi(\mathbf{u}) = \begin{cases} 1, & |u_j| \leq 1/2, \quad j = 1, \dots, d \\ 0, & \text{otherwise} \end{cases}$$

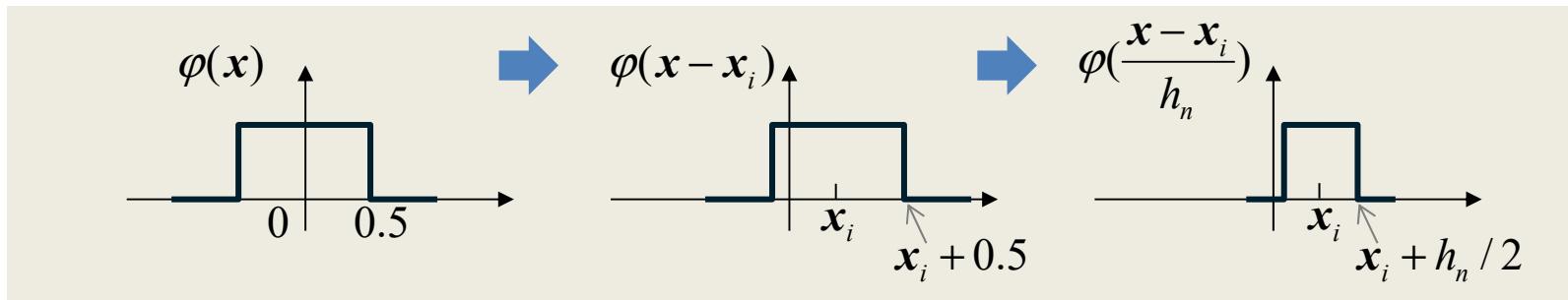
- 窗函数 $\varphi(\mathbf{u})$ 定义了一个中心位于坐标原点，边长为1的超立方体，立方体内部的函数值为1，外部的函数值为0。



Parzen窗方法

- 利用 $\varphi(u)$ 可以定义出一个中心位于训练样本 x_i , 边长为 h_n , 体积为 $V_n = h_n^d$ 的超立方体:

$$\varphi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right) = \begin{cases} 1, & |x_j - x_{ij}| \leq h_n/2, \quad j = 1, \dots, d \\ 0, & \text{otherwise} \end{cases}$$



- 如果某样本 x 落入该超立方体, 则 $\varphi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right) = 1$, 否则 $\varphi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right) = 0$ 。

Parzen窗方法

- 利用上式中 x 和 x_i 的对称性，在以 x 为中心的超立方体中包含训练样本数为

$$k_n = \sum_{i=1}^n \varphi\left(\frac{x - x_i}{h_n}\right)$$

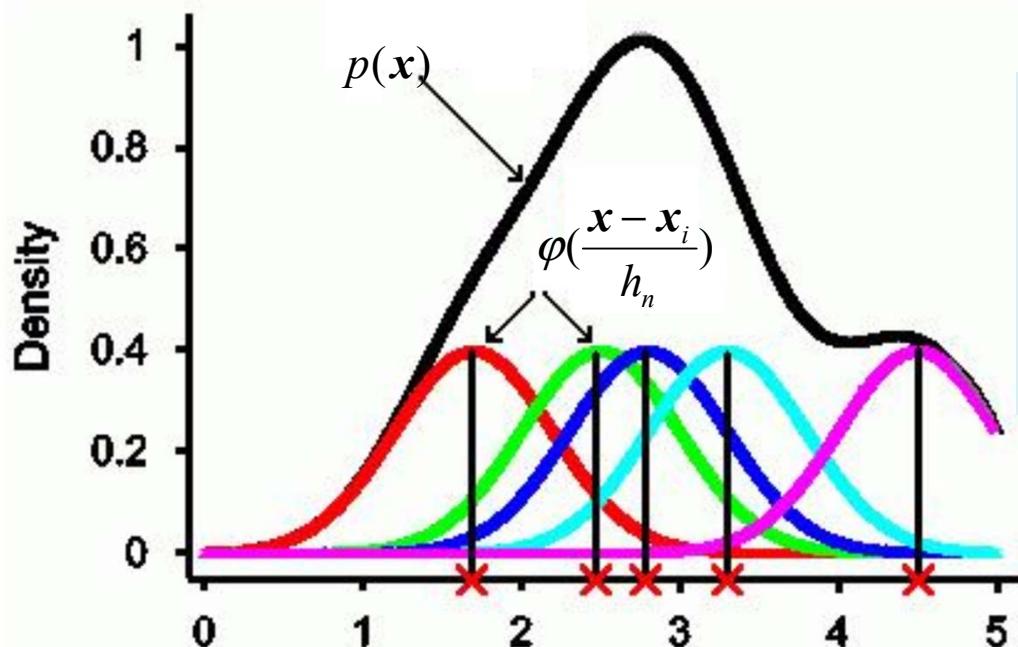
- 当样本 x_i 与 x 的距离较近时，它对概率密度函数估计的贡献为1，较远时贡献为0；关键看 x 是否落在以 x_i 为中心构造的窗函数内。

Important

Parzen窗方法

- 将 k_n 代入 $p_n(x) = \frac{k_n / n}{V_n}$ ，可得概率密度函数的估计式：

$$p(x) \approx \frac{1}{n} \sum_{i=1}^n \frac{1}{V_n} \phi\left(\frac{x - x_i}{h_n}\right)$$



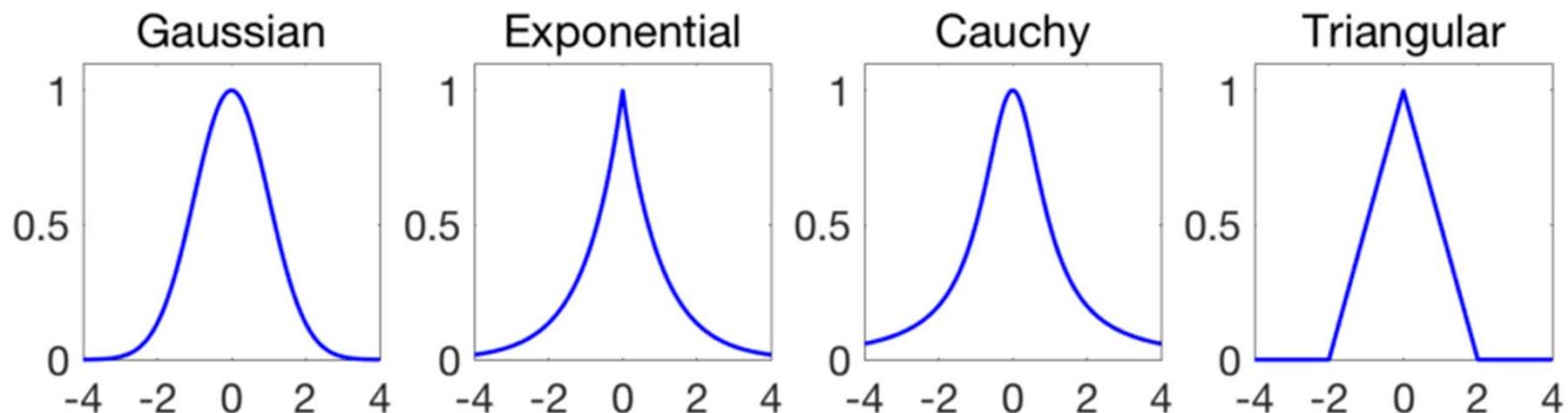
以高斯窗为例，概率密度函数是一系列以样本点为中心的窗口函数之和
(乘以系数)

x_i

Parzen窗方法

- 窗函数的形状是多种多样的。只要满足如下两个条件的函数都可作为窗函数（称为Parzen窗）：

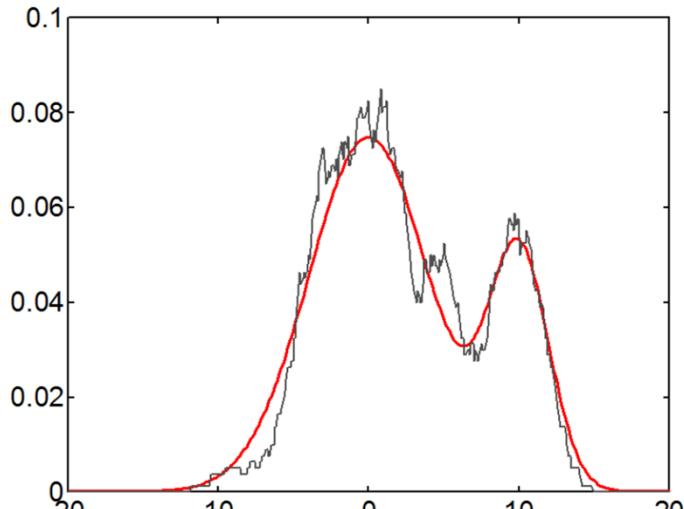
$$\varphi(\mathbf{u}) \geq 0, \int \varphi(\mathbf{u}) d\mathbf{u} = 1$$



Parzen窗方法

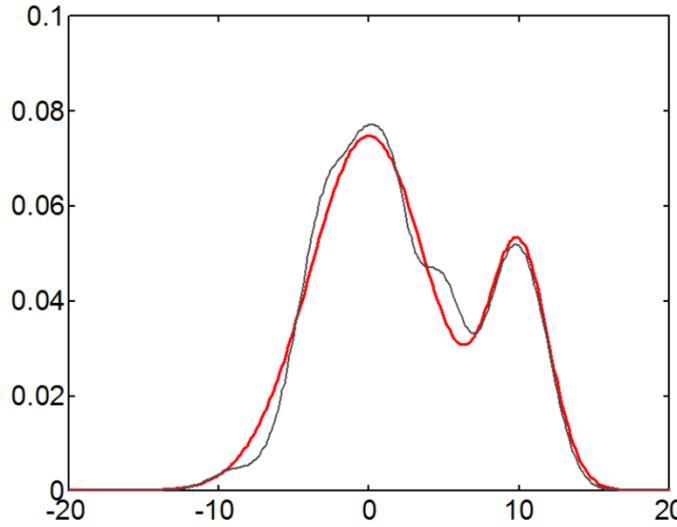
- 矩形窗函数不连续，所得到的概率密度函数呈阶梯型。为保证概率密度估计具有连续性，实际中更多使用连续窗函数，如各向同性的高斯函数：

$$\varphi(\mathbf{x}) = \frac{1}{(\sqrt{2\pi}h_n)^d} \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{2h_n^2}\right)$$



矩形窗函数的估计

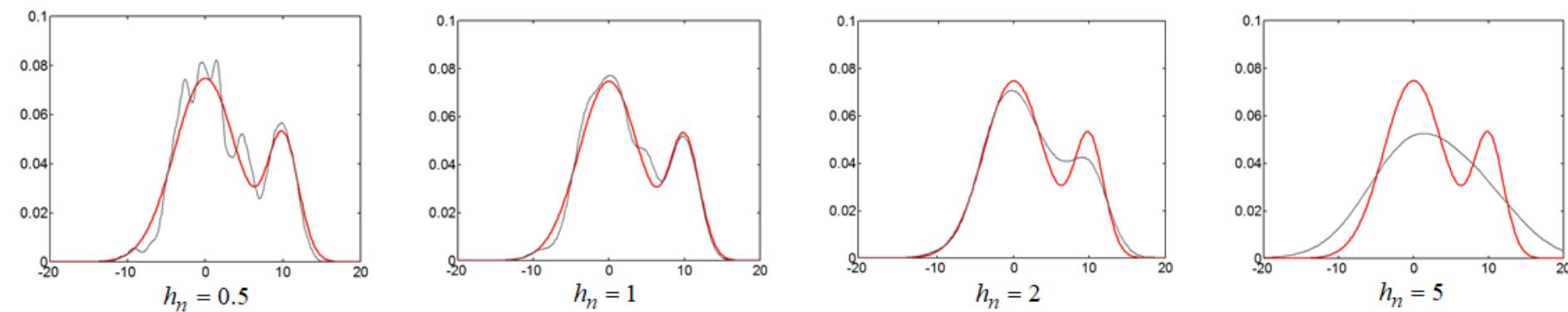
红线：真实概率密度
灰线：概率密度估计



高斯窗函数的估计

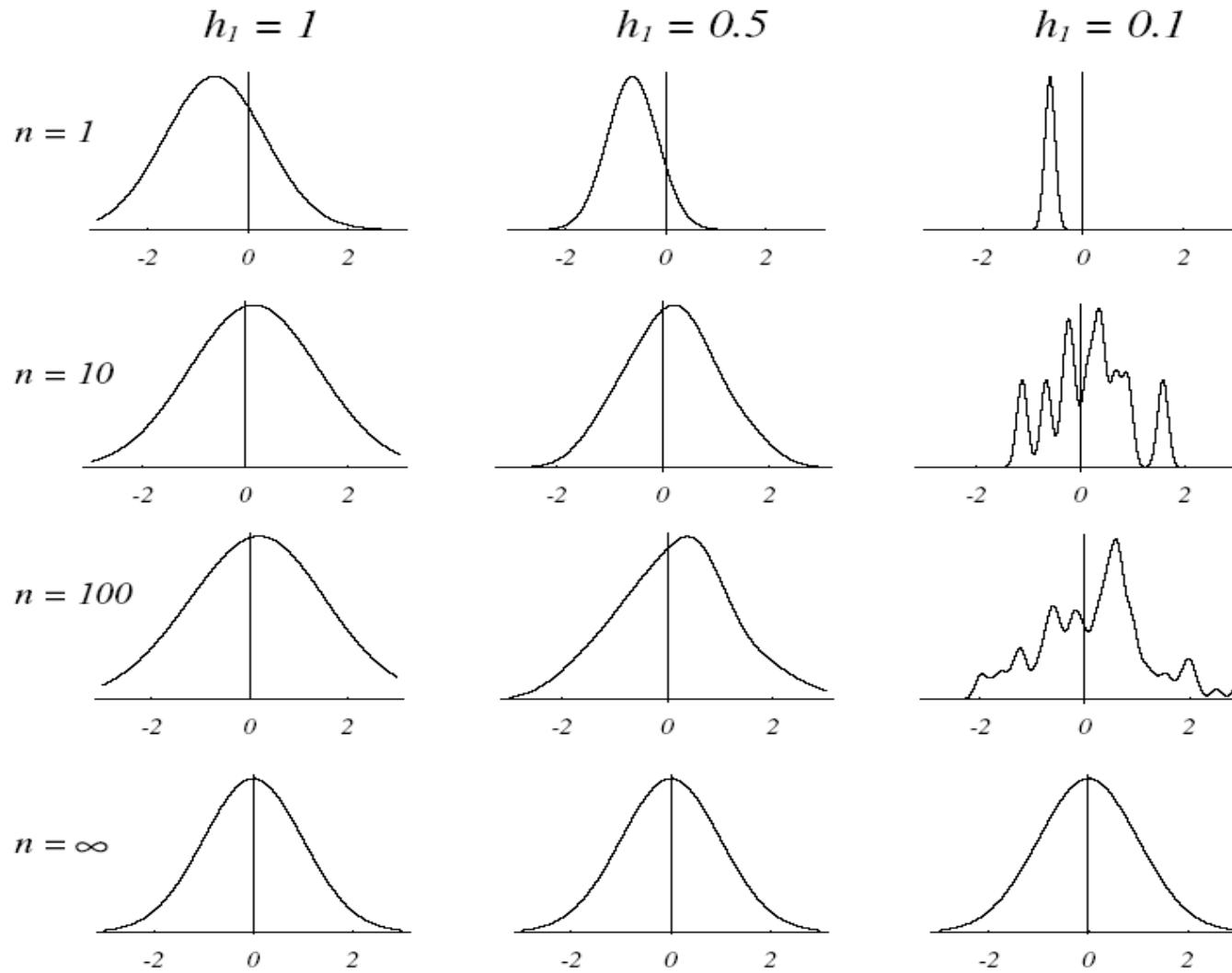
窗函数的宽度对估计的影响

- 高斯窗函数的宽度是高斯函数的标准差 h_n 。
- Parzen窗的宽度对估计结果有很大的影响：
 - 宽度过小，得到的密度曲线抖动过大；
 - 宽度过大，则会使得曲线过于平滑。
- 宽度选择应该与训练样本的数量相适应：样本数多宽度可以小些，样本数少则宽度要大些。



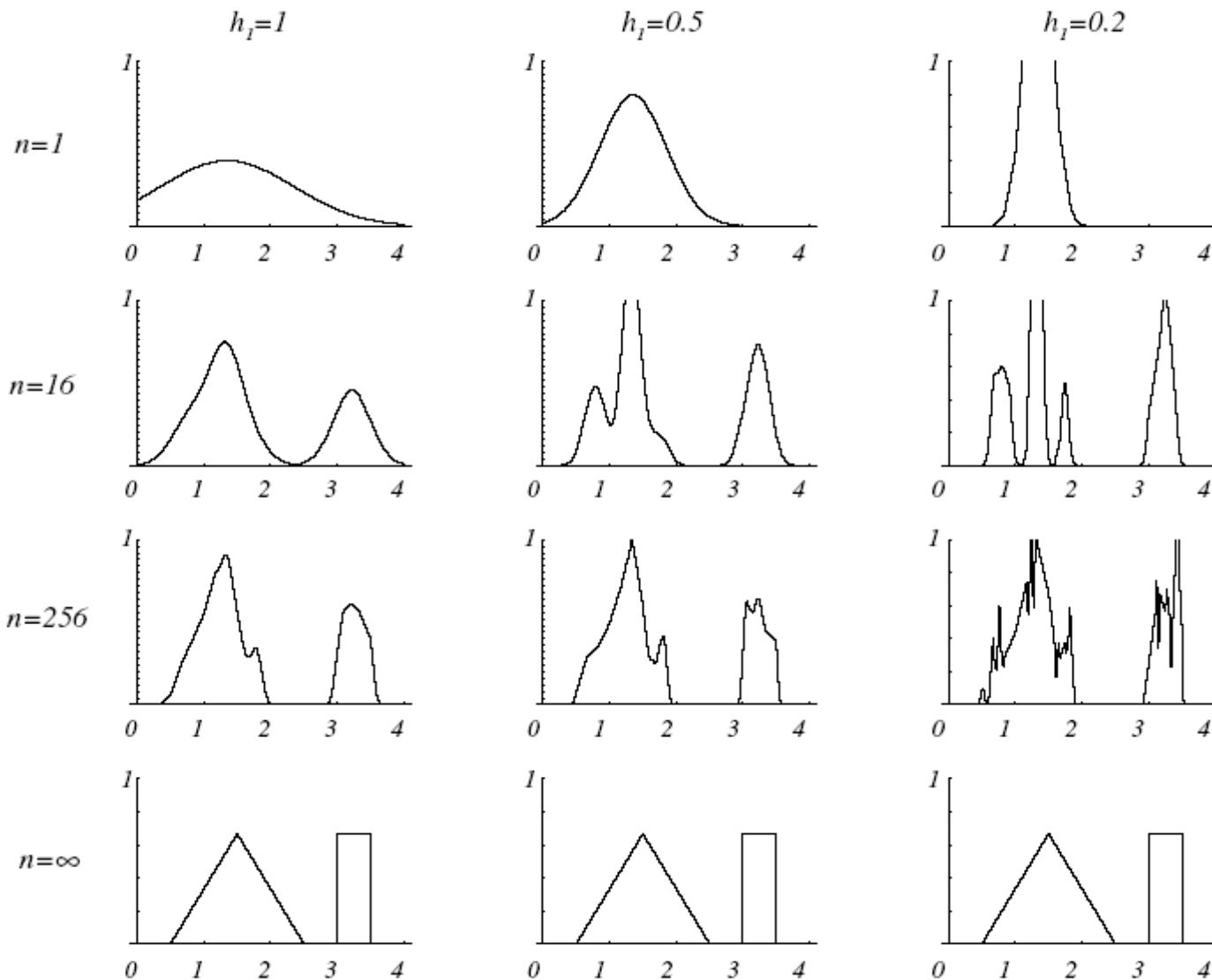
红线：真实概率密度；灰线：概率密度估计

窗函数的宽度对估计的影响



样本数无限时，
不同窗宽的估计一致且收敛

窗函数的宽度对估计的影响



Parzen窗识别算法

- 保存每个类别的所有训练样本；
- 选择窗函数 $\varphi(u)$ ，根据训练样本数 n 设置窗函数宽度 h_n ；
- 使用每类的训练样本，计算待识模式的类条件概率密度：

$$p(\mathbf{x} | \omega_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{1}{V_n} \varphi\left(\frac{\mathbf{x} - \mathbf{x}_j^i}{h_n}\right), \quad i = 1, \dots, c$$

其中 n_i 是第 i 个类别的训练样本数量， \mathbf{x}_j^i 是第 i 个类别的第 j 个训练样本。

- 根据贝叶斯判别准则进行分类。
-

Parzen窗法的优缺点

- Parzen窗法的优点：普遍适应，对规则或不规则分布，单峰或多峰分布都可适用。
- Parzen窗法的缺点：要求样本足够多，才能有较好的估计，计算量大，存储量大。
- 使用Parzen窗法时存在的主要困难：窗宽选择。窗宽太大，得到较平坦的分布，无法反映真实分布变化；窗宽太小，估计变化过大，很不稳定。

近邻估计方法

- Parzen窗法估计的是每个类别的条件概率密度函数 $p(x | \omega_i)$, 而近邻法则是直接估计每个类别的后验概率 $P(\omega_i | x)$ 。
- 将一个体积为 V 的区域 R 放置在待识样本点 x 周围。总训练样本数为 n , 其中 k 个在区域 R 内, 而 k 个样本中有 k_i 个属于 ω_i 类, 则样本 x 与类别 ω_i 同时发生的联合概率密度估计是:

$$p(x, \omega_i) \approx \frac{k_i / n}{V}$$

近邻估计方法

- 根据条件概率公式， ω_i 类后验概率的估计为

$$\begin{aligned}
 P(A | B) = \frac{P(A \text{ and } B)}{P(B)} \quad P(\omega_i | x) &= \frac{p(x, \omega_i)}{p(x)} \\
 &= \frac{p(x, \omega_i)}{\sum_{j=1}^c p(x, \omega_j)} \approx \frac{k_i}{\sum_{j=1}^c k_j} = \frac{k_i}{k}
 \end{aligned}$$

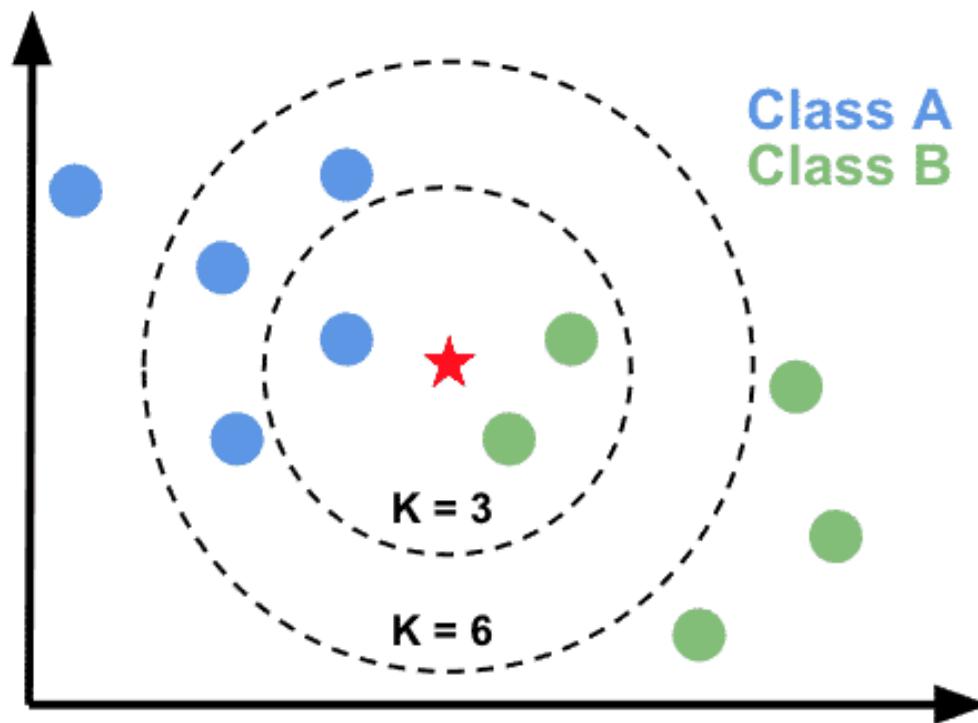
- 使用近邻法进行分类时，根据 k 个样本中属于不同类别样本的多少就可以得到分类结果：

如果 $i = \arg \max_{j=1, \dots, c} k_j$ ，则判别： $x \in \omega_i$

近邻分类算法

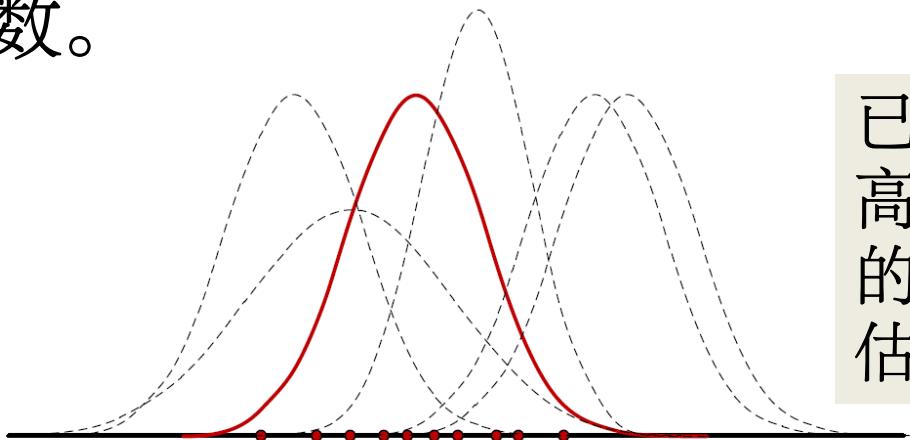
-
- 设置参数 k , 输入待识别样本 x ;
 - 计算 x 与每个训练样本的距离;
 - 选取距离最小的前 k 个样本, 统计其中包含各个类别的样本数 k_i ;
 - 如果 $i = \arg \max_{j=1,\dots,c} k_j$, 则判别 $x \in \omega_i$ 。
-

近邻分类算法



概率密度函数的参数估计方法

- 与非参数估计不同，参数估计方法假设每一类别的概率密度函数的形式已知，但具体参数未知。
- 参数估计方法的核心是估计概率密度函数的分布参数。



已知10个样本来自同一高斯分布，但高斯分布的可能性有很多，需要估计出均值和方差。

- 主要方法包括：**最大似然估计** (Maximum Likelihood Estimation, MLE) 和**贝叶斯估计** (Bayesian Estimation)。

最大似然估计

- 最大似然估计的目标是要从所有可能的分布参数中，寻找最有可能产生出训练样本的分布参数。
- 假设类条件概率密度形式已知，待估计的参数可以表示为参数矢量 θ 。例如一维高斯分布中的参数矢量为 $\theta = (\mu, \sigma^2)^T$ 。
- 将需要估计的概率密度函数表示为参数的形式： $p(x | \theta)$ ，其含义就是在已经确定某个分布参数 θ 的条件下，特征矢量 x 发生的概率密度。

最大似然估计

- 概率和似然的区别：
 - 概率（ probability ）：在已知模型参数 θ 的情况下，对样本 x 发生机会的估计（在参数 θ 确立的模型中产生样本 x 的可能性多大？）；
 - 似然（ likelihood ）：在已观测到样本 x 的前提下，产生该样本的模型参数 θ 的几率（哪哪一个模型参数最有可能产生样本 x ？即是，哪一个参数的似然最大？）。

最大似然估计

- 概率和似然的区别：

Probability	Likelihood
<p>What is the probability that $5 \leq x \leq 10$ given a normal distribution with $\mu = 13$ and $\sigma = 4$?</p> <p>Answer: 0.204</p>	<p>What is the likelihood that $\mu = 13$ and $\sigma = 4$ if you observed a value of $x = 10$?</p> <p>Answer: the likelihood is 0.075</p>
<p>What is the probability that $-1000 < x \leq 1000$ given a normal distribution $\mu = 13$ and $\sigma = 4$?</p> <p>Answer: 1.000</p>	<p>What is the likelihood that $\mu = 15$ and $\sigma = 2$ if you observed a value of $x = 10$?</p> <p>Answer: the likelihood is 0.009</p> <p>Conclusion: if the observed value was 10, it is more likely that the parameters are $\mu = 13$ and $\sigma = 4$.</p>

最大似然估计的主要思想

- 最大似然估计的主要思想：如果在一次观察中一个事件出现了，则我们可以认为这一事件出现的可能性很大。现在，事件 (x_1, \dots, x_N) 在一次观察（从概率总体中抽取一组样本）中居然出现了，则我们认为相应的参数 θ 和 x_1, \dots, x_N 使似然函数 $L(\theta)$ 达到了最大值。
- 观察次数越多，样本越具有代表性，最大似然估计的结果越准确。

最大似然估计

- 已知样本集 D 包含 n 个样本: x_1, \dots, x_n 。由独立同分布假设, 似然函数 (即在某特定 θ 下, 样本集 D 发生的概率) 计算为:

$$L(\theta) = p(D | \theta) = p(x_1, \dots, x_n | \theta) = \prod_{i=1}^n p(x_i | \theta)$$

- 为方便处理, 一般使用对数似然函数:

$$l(\theta) = \ln L(\theta) = \ln \left[\prod_{i=1}^n p(x_i | \theta) \right] = \sum_{i=1}^n \ln p(x_i | \theta)$$

- 参数 θ 的估计问题转化为优化问题来求解:

$$\theta^* = \arg \max_{\theta} l(\theta)$$

最大似然估计

- 例：假设样本满足均值为 μ , 方差为 σ^2 的高斯分布，推导单变量高斯分布参数的最大似然估计。
- 对数似然函数：

$$l(\theta) = \sum_{i=1}^n \ln p(x_i | \theta) = \sum_{i=1}^n \left[-\frac{1}{2} \ln 2\pi - \ln \sigma - \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

- 对数似然函数对 μ 求偏导和极值点：

$$\frac{\partial l(\theta)}{\partial \mu} = \sum_{i=1}^n \left[-\frac{-2(x_i - \mu)}{2\sigma^2} \right] = \frac{1}{\sigma^2} \left(\sum_{i=1}^n x_i - n\mu \right) = 0$$

- 解得均值的估计为： $\mu = \frac{1}{n} \sum_{i=1}^n x_i$

最大似然估计

- 对数似然函数对 σ 求偏导和极值点：

$$\begin{aligned}\frac{\partial l(\theta)}{\partial \sigma} &= \sum_{i=1}^n \left[-\frac{1}{\sigma} - \frac{-2(x_i - \mu)^2}{2\sigma^3} \right] \\ &= \frac{1}{\sigma^3} \left[\sum_{i=1}^n (x_i - \mu)^2 - n\sigma^2 \right] = 0\end{aligned}$$

- 解得方差的估计为：

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

最大似然估计

- 类似地，可以推导出多维高斯分布均值矢量和协方差矩阵的最大似然估计（详见附录D）。
- 样本集合 $D = \{x_1, \dots, x_n\}$ 独立抽样于均值为 μ ，协方差矩阵为 Σ 的高斯分布，则 μ 和 Σ 的最大似然估计分别为

$$\boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$$

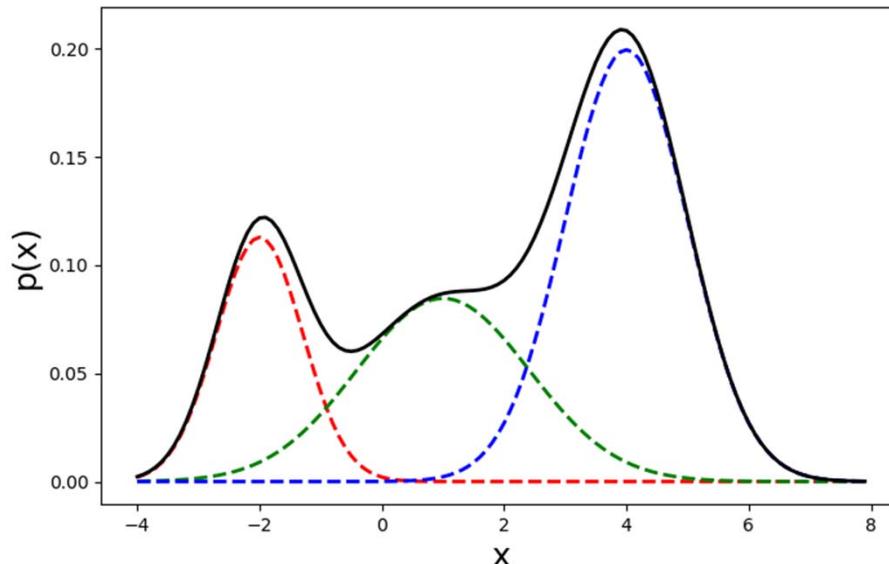
$$\boldsymbol{\Sigma} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T$$

高斯混合模型

- 高斯分布贝叶斯分类器能否适用于任意的模式分类问题？ 否
- 高斯分布只是概率密度函数的一种选择，实际问题中每个类别的样本可能呈现出不同分布形式。
- 在难以获取分布的先验知识的前提下，可以寻找一种“通用”的参数模型以描述任意的分布。

高斯混合模型

- 高斯混合模型 (Gaussian Mixture Model, GMM)：由K个高斯分布的线性组合所构成的，其中的每一个高斯分布也被称作是一个分量。
- 高斯混合模型是一个常用的通用模型，在一定条件下能够以任意的精度逼近任意概率密度函数。



红色、绿色、蓝色分布是三个高斯分量；黑色分布是红绿蓝三个分量之和。

高斯混合模型

- GMM概率密度函数表示为

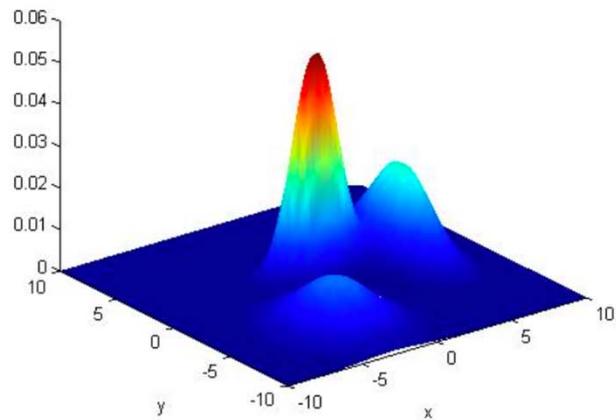
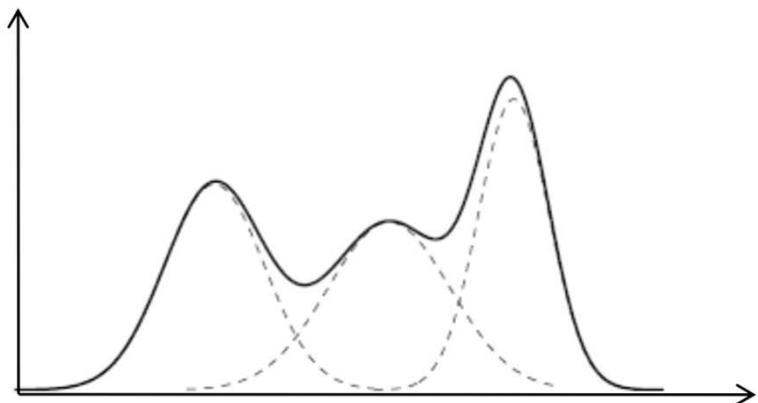
$$p(x|\theta) = \sum_{k=1}^K \alpha_k \mathcal{N}(x; \mu_k, \Sigma_k), \quad \alpha_1, \dots, \alpha_K \geq 0, \quad \sum_{k=1}^K \alpha_k = 1$$

- K 是包含的高斯分布分量数,
- $\mathcal{N}(x; \mu_k, \Sigma_k)$ 表示均值为 μ_k , 协方差矩阵为 Σ_k 的高斯分布密度函数,
- α_k 是第 k 个分量的组合系数,
- 参数矢量 θ 包含每一个分量的组合系数、均值矢量和协方差矩阵中的所有元素。

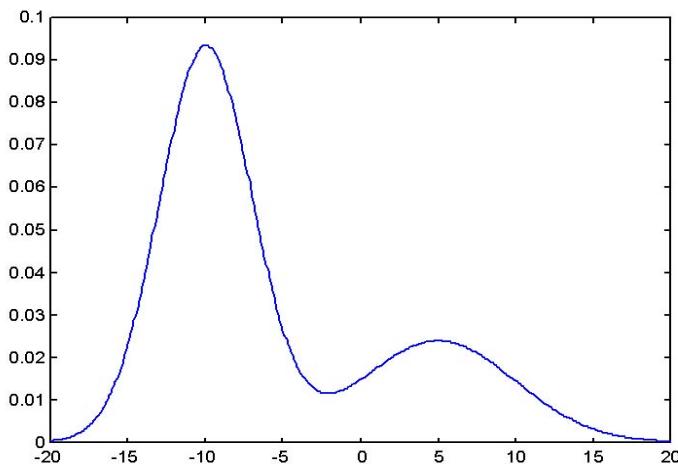
$$\theta = [\alpha_1, \dots, \alpha_K, \mu_1^T, \dots, \mu_K^T, [vec(\Sigma_1)]^T, \dots, [vec(\Sigma_K)]^T]^T$$

高斯混合模型

- 例：一维和二维的GMM



- 如何估计GMM的参数？



$$p(x | \theta) = 0.7 \mathcal{N}(x; -10, 2) + 0.3 \mathcal{N}(x; 5, 3)$$

高斯混合模型参数估计

- 方法1： 使用最大似然法估计。可写出GMM的对数似然函数 $l(\theta)$ 并求出 $l(\theta)$ 关于 θ 中每一个参数的偏导数；但直接令偏导数等于0求极值会导致一个复杂多元超越方程组，很难得到解析解。
- 方法2： 利用最优化方法，如梯度下降法，直接优化求解对数似然函数 $l(\theta)$ 。
- 方法3： 使用**期望最大化算法 (Expectation Maximization, EM)** 估计。

高斯混合模型的产生

- 介绍EM算法前，需要先讨论已知参数的GMM产生样本的过程。

$$p(x|\theta) = \sum_{k=1}^K \alpha_k \mathcal{N}(x; \mu_k, \Sigma_k)$$

- 一个满足已知参数高斯混合模型分布的样本可以按照如下过程产生：
 - 首先以组合系数 α_k 作为先验概率随机地选择一个高斯分量 $\mathcal{N}(x; \mu_k, \Sigma_k)$;
 - 然后根据这个高斯分布 $\mathcal{N}(x; \mu_k, \Sigma_k)$ 的均值矢量和协方差矩阵产生出一个具体的样本。

高斯混合模型的估计

- GMM的参数估计算则是一个相反的过程：已知由GMM产生的样本集 $D = \{x_1, \dots, x_n\}$ ，根据这个样本集估计GMM的参数。
- 存在着两组未知信息：
 - 1) GMM的参数 θ ；
 - 2) 每个样本 x_i 产生自哪一个分量高斯分布？用 $Y = \{y_1, \dots, y_n\}$ 表示这一组待估计信息，其中 $y_i \in \{1, \dots, K\}$ 表示样本 x_i 由第 y_i 个高斯产生。

高斯混合模型的估计

- 已知 Y 估计 θ : 在已知 Y 的情况下, θ 估计如下

$$\alpha_k = \frac{1}{n} \sum_{i=1}^n I(y_i = k) \quad (\text{公式*})$$

$$\boldsymbol{\mu}_k = \sum_{i=1}^n I(y_i = k) \mathbf{x}_i \Bigg/ \sum_{i=1}^n I(y_i = k)$$

$$\boldsymbol{\Sigma}_k = \sum_{i=1}^n I(y_i = k) (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Bigg/ \sum_{i=1}^n I(y_i = k)$$

其中 $I(y_i = k) = \begin{cases} 1, & y_i = k \\ 0, & y_i \neq k \end{cases}$ 为示性函数, 求和式

$\sum_{i=1}^n I(y_i = k)$ 计算的是第 k 个分量产生的样本数。

高斯混合模型的估计

- 已知 θ 估计 Y : 如果知道了GMM的参数 θ , 也可以估计出 Y , 即样本 x_i 由哪个分量高斯产生的。
- 已知 θ 估计 Y 的问题可以看作是 K 个类别的分类问题: 每个类别的条件概率密度函数均为高斯分布, 而类别的先验概率为 $\alpha_1, \dots, \alpha_K$ 。根据最小错误率贝叶斯判别准则, 对 y_i 可做如下估计:

$$y_i = \arg \max_{k=1, \dots, K} \alpha_k \mathcal{N}(x; \mu_k, \Sigma_k) \quad (\text{公式#})$$

高斯混合模型的估计

- Y 和 θ 均未知：如果两组信息均未知，可以使用以下过程对两组信息进行迭代更新：
 - 1) 随机初始化GMM中 K 个高斯分布的参数及组合系数 $\alpha_1, \dots, \alpha_K$ ；
 - 2) 根据公式(#)判别每个训练样本是由哪个分量高斯分布产生的，得到 $Y = \{y_1, \dots, y_n\}$ ；
 - 3) 有了对信息 Y 的估计，可以按照公式(*)重新修正对参数 θ 的估计。
- 循环迭代公式(#)和公式(*)的估计过程，直到两个估计值不再变化（或变化甚微）为止。

高斯混合模型的估计

- 示性函数 $I(y_i = k)$ 是“硬分类”，更合理的“软分类”可计算第 i 个样本由第 k 个高斯产生概率

$$P(y_i = k) = \alpha_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) / \sum_{j=1}^K \alpha_j \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

- 参数矢量 θ 的估计可以更新为：

$$\alpha_k = \frac{1}{n} \sum_{i=1}^n P(y_i = k)$$

$$\boldsymbol{\mu}_k = \sum_{i=1}^n P(y_i = k) \mathbf{x}_i / \sum_{i=1}^n P(y_i = k)$$

$$\boldsymbol{\Sigma}_k = \sum_{i=1}^n P(y_i = k) (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T / \sum_{i=1}^n P(y_i = k)$$

GMM学习算法

- GMM的高斯分量数 K 是预设参数，需要根据具体问题确定。
 - 概率密度函数越复杂需要的 K 越大，然而 K 的增加也会带来模型参数增多，需要的训练样本数量更多。
- 模型其他参数可以随机初始化，但需要保证组合系数 ≥ 0 且相加为1，协方差矩阵对称正定。
- 算法迭代的收敛一般是依据连续两轮迭代中似然函数的变化量确定，当变化量小于某一阈值（收敛精度）时判断算法收敛。

GMM学习算法

- 设置模型中高斯分量的个数 K , 随机初始化参数矢量 $\theta = (\alpha_1, \dots, \alpha_K, \mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K)$, 迭代次数 $t = 0$, 设置收敛精度 η ;
 - 循环: $t \leftarrow t + 1$
 - 计算所有训练样本由每个分量高斯分布产生的概率 $P(y_i = k), i = 1, \dots, n, k = 1, \dots, K$;
 - 重新估计参数 $\theta = (\alpha_1, \dots, \alpha_K, \mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K)$;
 - 计算似然函数值 $L_t(\theta) = p(x_1, \dots, x_n | \theta)$;
 - 直到满足收敛条件: $[L_t(\theta) - L_{t-1}(\theta)] < \eta$ 。
-

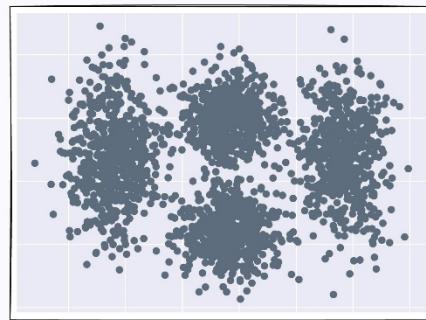
高斯混合模型的讨论

- 类似于GMM，由多个简单概率密度函数的线性组合所构成的复杂分布一般被称作**混合密度函数**。
- 混合密度函数的参数估计同聚类之间存在本质内在联系。事实上，GMM是常用的聚类方法。
- GMM计算的迭代过程可退化为K-均值聚类。
 - 如果假设混合系数相等且每个高斯分布的协方差矩阵为相同的对角矩阵，那么公式(#)只需根据样本与 K 个高斯分布均值矢量之间的距离远近来判别 y_i ，而公式(*) 只需重新估计 K 个均值矢量。

高斯混合模型的讨论

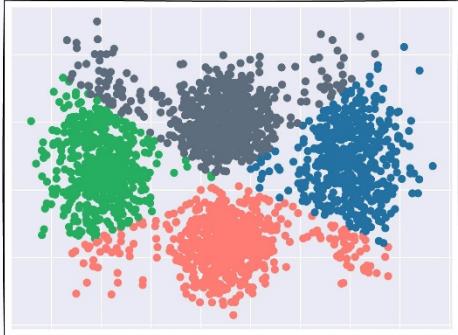
K-means	GMM
主要针对圆形或球形样本进行聚类，对于椭圆数据处理效果不佳	对于较复杂分布的数据，如椭圆数据，都可以有较好效果
对于不均衡的样本类别聚类效果不佳	考虑了各类别权重，适用于非均衡样本
判别式模型，直接在样本空间中寻找最优分界面	产生式模型，从样本分布出发，计算概率分布以求得分类结果
硬聚类，结果属于0-1	软聚类，聚类结果是一个概率分布
计算较简单	计算较复杂

高斯混合模型的讨论

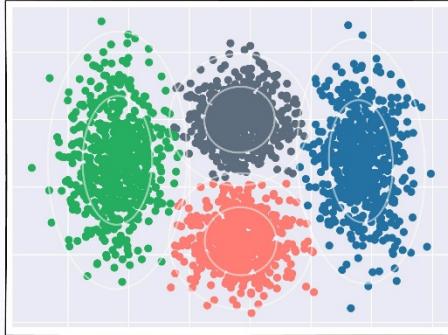


Clusters with
different
variances

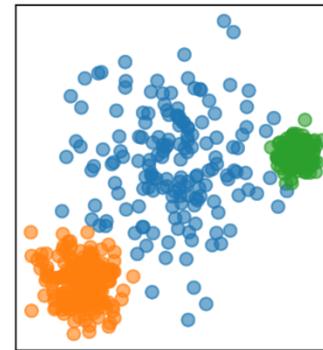
KMeans



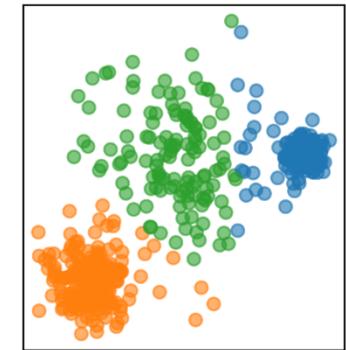
Gaussian Mixture Model



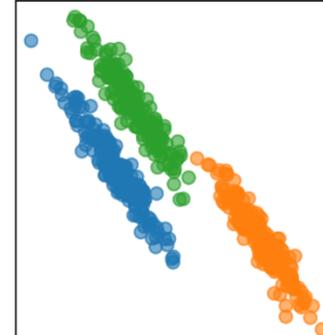
GaussianMixture



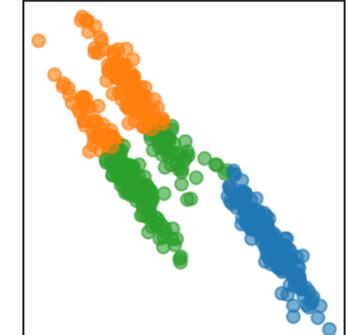
KMeans



GaussianMixture



KMeans



期望最大化算法

- GMM的参数迭代算法是否收敛？是否能够收敛于似然函数的极值点？**是**
- GMM的参数估计算法是一种**期望最大化算法**（Expectation Maximization，EM），且算法的收敛性已被证明。
- EM算法已经被广泛地应用于解决各种复杂概率密度模型的参数估计问题。

期望最大化算法

- 假设样本集由两部分组成： $D = \{X, Y\}$ ，其中 X 为已知可见的数据， Y 是未知隐含的数据。
- 在这种假设条件下重新写出对数似然函数：

$$l(\theta) = \ln p(D | \theta) = \ln p(X, Y | \theta)$$

- 由于 Y 是未知的，所以无法优化对数似然函数求取极值点。下面来考虑 Y 所有可能情况下的对数似然函数 — 期望对数似然函数：

$$Q(\theta) = E_Y [\ln p(X, Y | \theta)] = \int \ln p(X, Y | \theta) p(Y) dY$$

期望最大化算法

- 由于对 Y 取了数学期望， $Q(\theta)$ 的变量中只有 θ 未知。但是对 $Q(\theta)$ 直接优化仍需要知道积分式中的 $p(Y)$ 。
- EM 算法的做法是，首先设置一个参数 θ 的猜测值 θ^g ，在已知 X 和 θ^g 的条件下估计出 Y 发生的概率 $p(Y | X, \theta^g)$ ，用其近似 $p(Y)$ ，得：

$$\text{E步: } Q(\theta; \theta^g) = \int \ln p(X, Y | \theta) p(Y | X, \theta^g) dY$$

其中 $Q(\theta; \theta^g)$ 中的分号表示这是关于 θ 的函数， θ^g 是相关的固定值。

期望最大化算法

- 用 $Q(\theta; \theta^g)$ 替代对数似然函数进行优化：

$$\text{M步: } \theta^* = \arg \max_{\theta} Q(\theta; \theta^g)$$

其中 θ^* 不是极值点，而是改进的猜测值。

- 可以用 θ^* 代替 θ^g : $\theta^g \leftarrow \theta^*$, 重构函数 $Q(\theta; \theta^g)$ 进行优化，这构成了EM算法的一个迭代过程。
- EM算法就是首先随机初始化参数 θ ，然后通过 E步和M步的迭代逐渐优化对 θ 的估计，最后收敛于一个极值点。

期望最大化算法

- 初始化参数 θ^1 , 设置迭代次数 $t = 0$, 设置收敛精度 η ;
 - 循环: $t \leftarrow t + 1$
 - E步: $Q(\theta; \theta^t) = \int \ln p(X, Y | \theta) p(Y | X, \theta^t) dY$;
 - M步: $\theta^{t+1} = \arg \max_{\theta} Q(\theta; \theta^t)$;
 - 直到满足收敛条件: $[Q(\theta; \theta^{t+1}) - Q(\theta; \theta^t)] < \eta$ 。
-

期望最大化算法讨论

- 上表的EM算法是一个形式化的计算过程，需要根据实际问题中的概率密度函数模型，将E步和M步具体化。
- EM算法是收敛的，但只能保证收敛于对数似然函数的一个局部极值点，并不能保证收敛于全局最大值点。具体收敛情况与初始值的设置有关。
 - 使用EM算法时，需要根据具体问题的先验信息来设置适合的参数初始值。
 - 如果缺少先验，可以尝试设置多个不同的初始值，根据算法收敛时的似然函数值的大小选择最优结果。

贝叶斯估计

- 参数估计的贝叶斯观点：
 - 概率密度函数的参数 θ 是一个随机矢量，最大似然估计找到的是 θ 发生可能性最大的值 θ^* ，并认为 x 是由以 θ^* 为参数的概率密度函数产生的，因此得到的概率密度为 $p(x | \theta^*)$ ；
 - 然而随机矢量 θ 虽然是 θ^* 的可能性最大，但它也可能取其他值，简单地以 $p(x | \theta^*)$ 作为概率密度值有失偏颇，应该在考虑 θ 所有发生可能性的条件下计算 x 的概率密度。

贝叶斯估计

- 假设集合 $D = \{x_1, \dots, x_n\}$ 中的训练样本独立地采样自同一个以 θ 为参数的概率密度函数 $p(x | \theta)$ 。
- 最大似然估计**: 认为 θ 是一个确定的未知矢量
 - 用样本集 D 估计出最优参数 θ^* , 然后计算模式 x 的概率密度 $p(x | \theta^*)$ 。
- 贝叶斯估计**: 认为 θ 是一个随机变量, 以一定的概率分布取所有可能的值
 - 学习: 由样本集 D 估计 θ 的分布 $p(\theta | D)$,
 - 分类: 计算在已知 D 的条件下, 模式 x 发生的概率密度 $p(x | D)$ 。

贝叶斯估计

- 分类过程中， $p(x | D)$ 可以估算如下：

$$\begin{aligned} p(x | D) &= \int p(x, \theta | D) d\theta && \text{注: 条件概率符号中的} \\ &\quad \nearrow \text{全概率公式} && \text{“,” 优先级要高于 “|”} \\ p(A) &= \int p(A, B) dB && \\ &\quad \nearrow \text{条件概率公式} && \\ p(A, B) &= p(A | B) p(B) && \\ &\quad \nearrow \text{在已知 } \theta \text{ 的情况下,} \\ &\quad && \quad D \text{ 和 } x \text{ 相互独立 (i.i.d.)} \\ &= \int \underbrace{p(x | \theta)}_{\text{已知的关于 } x \text{ 和 } \theta \text{ 的函}} p(\theta | D) d\theta && \text{(分类公式)} \\ &\quad \downarrow && \\ && \quad \text{未知的在 } D \text{ 条件下 } \theta \\ && \quad \searrow \text{的分布, 需要在学习过} \\ && \quad \text{程中得到} \end{aligned}$$

贝叶斯估计

- 学习过程中， $p(\theta | D)$ 可以估算如下：

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

贝叶斯公式

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

全概率公式

$$p(A) = \int p(A | B)p(B)dB$$

D 中独立同分布的元素 x_i 的联合概率密度

$$= \frac{\prod_{i=1}^n p(x_i | \theta)p(\theta)}{\int \prod_{i=1}^n p(x_i | \theta)\underline{p(\theta)}d\theta}$$

(学习公式)

参数 θ 的先验分布，包含了关于 θ 的先验知识

贝叶斯估计

- 贝叶斯估计的两步计算：
 - 1) 学习过程：根据训练样本集 D 和参数的先验分布 $p(\theta)$ 由（学习公式）计算出参数的后验分布 $p(\theta | D)$ ；
 - 2) 分类过程：将待识模式 x 和参数后验概率 $p(\theta | D)$ 代入（分类公式）计算积分，得到矢量 x 发生的概率密度 $p(x | D)$ 。
- 实际计算过程中，（学习公式）和（分类公式）的积分计算可能非常复杂，很难得到解析解。

贝叶斯估计

- 已知
 - 独立同分布训练样本集 D ;
 - 类条件概率密度函数 $p(x | \theta)$ 的形式，但参数 θ 未知；
 - 参数 θ 的先验概率密度函数 $p(\theta)$;
- 求解：在已有训练样本集 D 的条件下，类条件概率密度函数 $p(x | D)$ 。

高斯分布的贝叶斯估计

- 假设样本集 $D = \{x_1, \dots, x_n\}$ 来自于1维高斯分布 $\mathcal{N}(\mu, \sigma^2)$ ，即 $p(x | \mu) \sim \mathcal{N}(\mu, \sigma^2)$ ，其中方差 σ^2 是已知的，但均值 μ 未知。
- 均值 μ 的先验满足 $p(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$ ，其中 μ_0 和 σ_0^2 均已知。
- 要求：估计均值 μ 的后验概率 $p(\mu | D)$ （学习）并估计 x 的概率密度函数 $p(x | D)$ （分类）。

高斯分布的贝叶斯估计

- (学习过程) 均值 μ 的后验概率估计可得: (详见附录F)

$$p(\mu | D) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_n}{\sigma_n^2}\right)^2\right] = \mathcal{N}(\mu_n, \sigma_n^2)$$

其中 $\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \right) \hat{\mu}_n + \left(\frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \right) \mu_0,$

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n x_i,$$

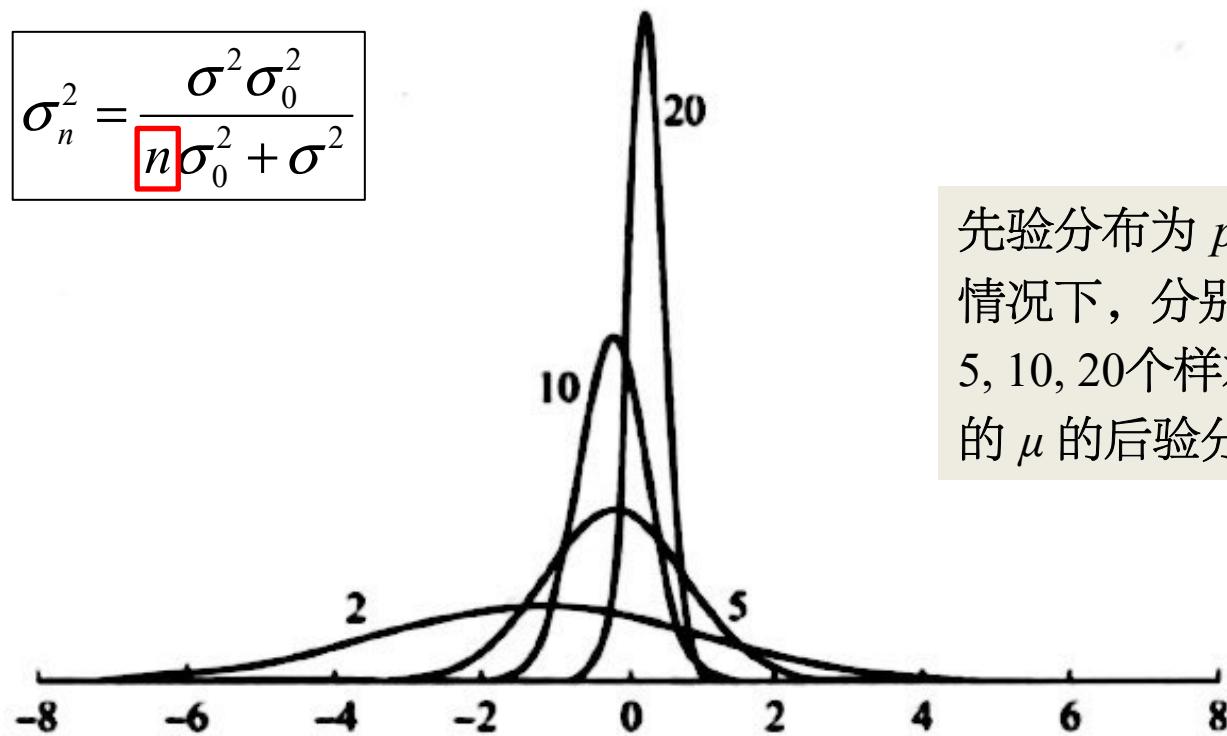
$$\sigma_n^2 = \frac{\sigma^2 \sigma_0^2}{n\sigma_0^2 + \sigma^2}$$

如果 μ 的先验 $p(\mu)$ 是高斯, 那么后验 $p(\mu | D)$ 仍是高斯, 但均值由 μ_0 移动到了 μ_n , 方差由 σ_0^2 变为 σ_n^2 。

高斯分布的贝叶斯估计

- $p(\mu | D)$ 的估计与样本量的关系：随着训练样本数量的增加，后验分布的方差 σ_n^2 逐渐减小，这表明对估计的置信程度在增加， μ 的随机性在减小。

$$\sigma_n^2 = \frac{\sigma^2 \sigma_0^2}{n\sigma_0^2 + \sigma^2}$$



先验分布为 $p(\mu) = \mathcal{N}(0, 20)$ 的情况下，分别使用包含 $n = 2, 5, 10, 20$ 个样本的训练集得到的 μ 的后验分布

高斯分布的贝叶斯估计

- $p(\mu | D)$ 的估计与样本量的关系：当 $n \rightarrow \infty$ 时有如下结果：

$$\lim_{n \rightarrow \infty} \mu_n = \lim_{n \rightarrow \infty} \left[\left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \right) \hat{\mu}_n + \left(\frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \right) \mu_0 \right]$$

$$= \lim_{n \rightarrow \infty} \hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n x_i \quad \longrightarrow$$

说明当样本数量无穷多时，
贝叶斯估计的结果同最大
似然估计是一致的

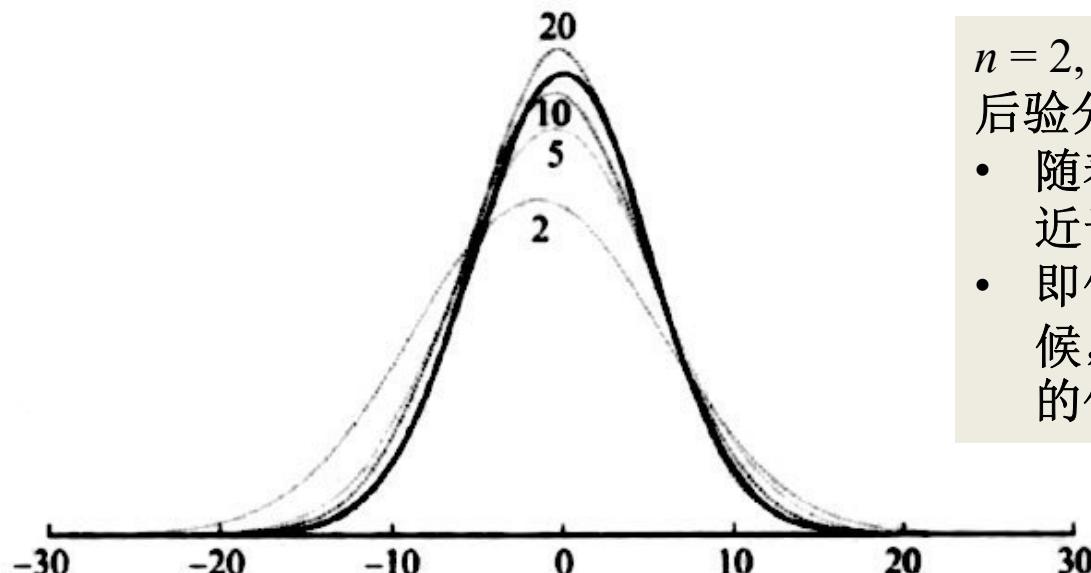
$$\lim_{n \rightarrow \infty} \sigma_n^2 = \lim_{n \rightarrow \infty} \frac{\sigma^2 \sigma_0^2}{n\sigma_0^2 + \sigma^2} = 0 \quad \longrightarrow$$

估计方差为0表明对该估
计的结果是完全确信的，
不存在不确定性。

高斯分布的贝叶斯估计

- (分类过程) x 的概率密度函数估计为: [\(详见附录F\)](#)

$$p(x | D) = \frac{1}{\sqrt{2\pi(\sigma^2 + \sigma_n^2)}} \exp\left[-\frac{1}{2}\left(\frac{x - \mu_n}{\sigma^2 + \sigma_n^2}\right)^2\right]$$
$$= \mathcal{N}(x; \mu_n, \sigma^2 + \sigma_n^2)$$

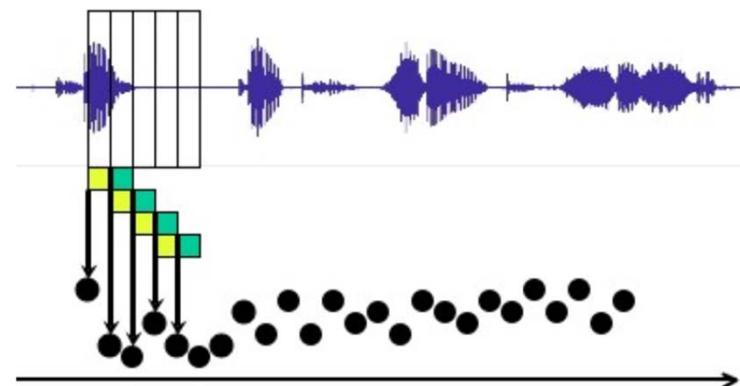


$n = 2, 5, 10, 20$ 的情况下，样本的后验分布 $p(x | D)$ 估计。

- 随着 n 的增多， $p(x | D)$ 越接近于真实分布（图中实线）；
- 即使 $n = 2$ ，训练样本很少的时候，也可以得到一个可以接受的估计结果。

以序列形式出现的模式

- 本课程已介绍的分类方法针对以特征矢量方式描述的模式。但在很多问题中，模式以序列形式出现的，例如视频、语音、基因序列等。如将这序列信号采用特征矢量的方式描述，会损失信号的先后次序变化信息。
- 对于有时间延续性的模式，常将信号划分为一系列短的时间段，在每一个短时间段内提取特征，并将特征连接形成一个特征矢量的序列，有效描述信号变化。

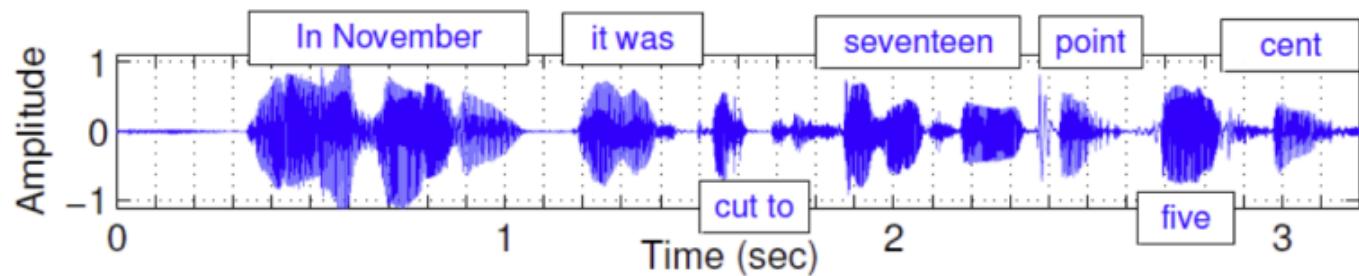


时间序列

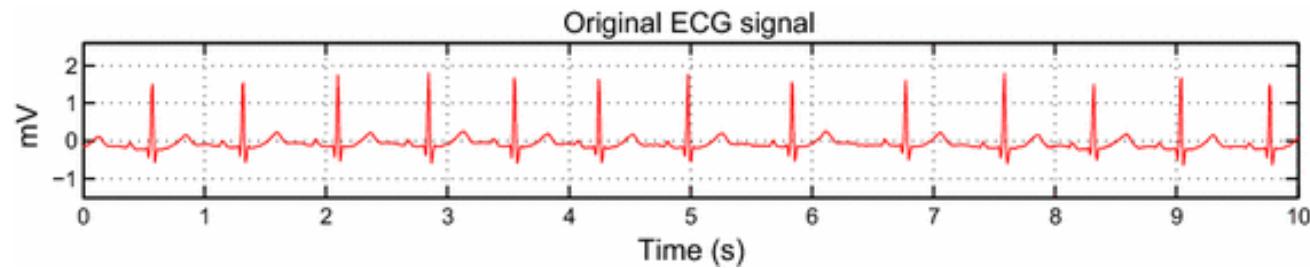
视频



音频



生理信号



以序列形式出现的模式

- 以序列形式出现的模式可以表示为：

$$V^T = v_1, v_2, \dots, v_T$$

其中 T 为序列长度，序列中的元素 v_i 称为时刻 i 的观察值，可以是一个特征矢量。

- 如果希望采用贝叶斯分类器识别以序列形式出现的模式，就需要构建描述序列的概率密度函数，计算每个类别产生出需要识别序列的概率密度。
- 隐含马尔科夫模型（Hidden Markov Model, HMM）是一种常用的序列概率密度描述模型。

马尔可夫模型

- 马尔科夫模型：由若干状态构成，模型当前所处状态只与之前的状态有关，与之后的状态无关。
- 马尔科夫过程：由马尔科夫模型所产生的一类状态转移过程。
- 这里，只讨论一种最简单的马尔科夫模型：离散时间有限状态一阶马尔科夫模型。

一阶马尔科夫模型

- 一阶马尔科夫模型由 M 个状态 $W = \{w_1, \dots, w_M\}$ 构成，在每个时刻 t ，模型处于某个状态 $w(t) \in W$ ，经过 T 个时刻，产生出一个长度为 T 的状态序列 $W^T = w(1), \dots, w(T)$ 。
- 状态转移概率 a_{ij} : 模型在 $t - 1$ 时刻的状态为 w_i ，而在时刻 t 的状态为 w_j 的概率
$$a_{ij} = P(w(t) = w_j | w(t-1) = w_i)$$
- 一阶马尔科夫模型中， a_{ij} 完全由 $t - 1$ 时刻的状态决定，而与时刻 t 无关。

一阶马尔科夫模型

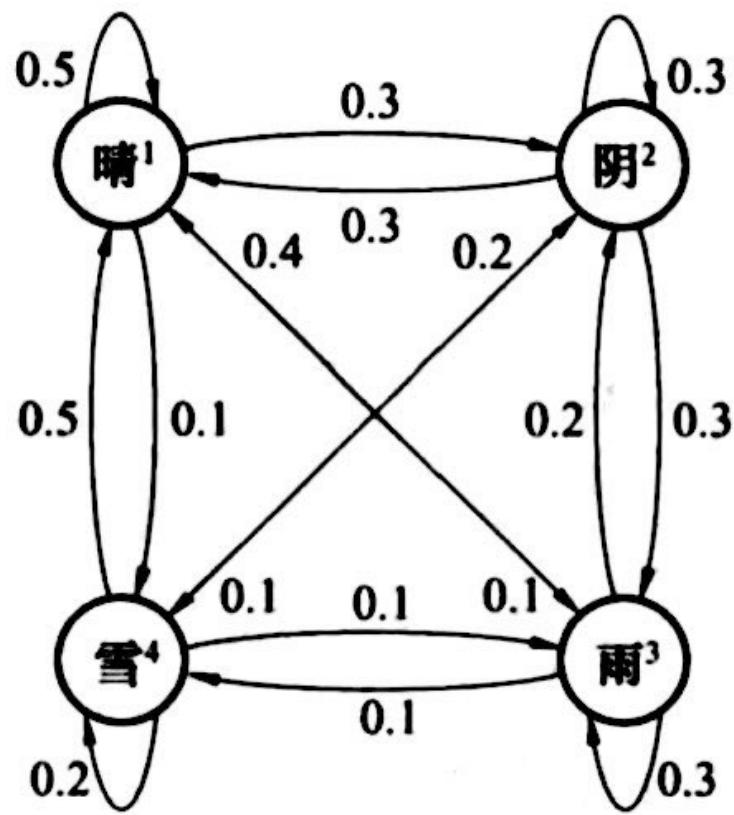
- 定义模型初始于状态 w_i 的概率为 π_i 。
- 完整的一阶马尔科夫模型可以用参数 $\theta = (\pi, A)$ 表示，其中：

$$\pi = (\pi_1, \dots, \pi_M), \quad A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1M} \\ a_{21} & a_{22} & \cdots & a_{11} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MM} \end{bmatrix}$$

且参数应该满足 $\sum_{i=1}^M \pi_i = 1, \sum_{j=1}^M a_{ij} = 1$

一阶马尔科夫模型

- 例：某城市天气变化采用下图所示的一阶马尔科夫模型描述。



- 4种天气状态：晴、阴、雨、雪，编号1~4。
- 每种天气发生的初始概率为 $\pi = (0.5, 0.3, 0.1, 0.1)$
- 状态转移概率矩阵：

$$A = \begin{bmatrix} 0.5 & 0.3 & 0.1 & 0.1 \\ 0.3 & 0.3 & 0.3 & 0.1 \\ 0.4 & 0.2 & 0.3 & 0.1 \\ 0.5 & 0.2 & 0.1 & 0.2 \end{bmatrix}$$

一阶马尔科夫模型

- 模型输出状态序列的概率可以由初始状态概率与各次状态转移概率相乘得到。
- 例如：某一阶马尔科夫模型产生一个特定状态序列 $W^5 = w_3 w_1 w_2 w_2 w_4$ 的概率为：

$$P(W^5 | \theta) = \pi_3 a_{31} a_{12} a_{22} a_{24}$$

- 上例（天气变化）中，前3天晴、后4天下雨的概率是：

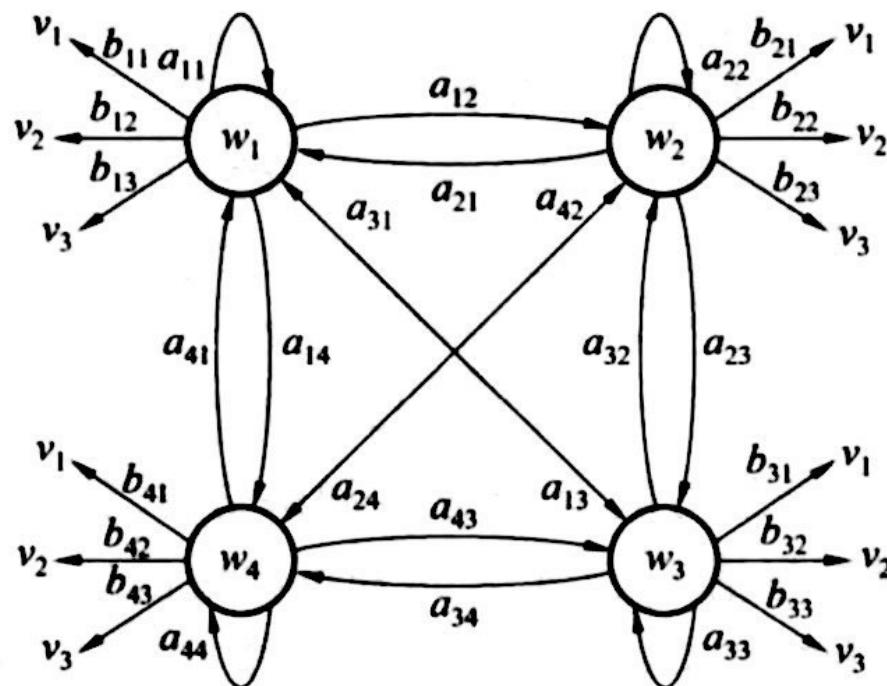
$$\begin{aligned} P(w_1 w_1 w_1 w_3 w_3 w_3) &= \pi_1 a_{11} a_{11} a_{13} a_{33} a_{33} a_{33} \\ &= 0.00016875 \end{aligned}$$

隐含马尔科夫模型

- 隐含马尔科夫模型（HMM）的内部是一个马尔科夫模型，每一时刻都依据概率发生状态转移。但状态转移的过程是观察不到的，所能够观察到的是每一时刻模型根据所处的状态产生的一个“观察值”序列： $V^T = v(1), \dots, v(T)$ 。
- HMM中，不可见的状态是可见的观察值产生的内在原因。
- 隐状态输出的观察值可以是离散值，连续值，也可以是一个矢量。

隐含马尔科夫模型

- 以离散的HMM为例，隐状态可能输出的观察值集合为 $v(t) \in V = \{v_1, \dots, v_K\}$ ，第 i 个隐状态输出第 j 个观察值的概率为 $b_{ij} = P(v_j | w_i)$ 。



隐含马尔科夫模型

- HMM的参数除了初始状态概率 π 和状态转移概率矩阵 A 之外，增加了状态输出概率矩阵 B ：

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1K} \\ b_{21} & b_{22} & \cdots & b_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ b_{M1} & b_{M2} & \cdots & b_{MK} \end{bmatrix}, \text{ 其中 } \sum_{j=1}^K b_{ij} = 1$$

M : 状态数
 K : 可能的观察值数

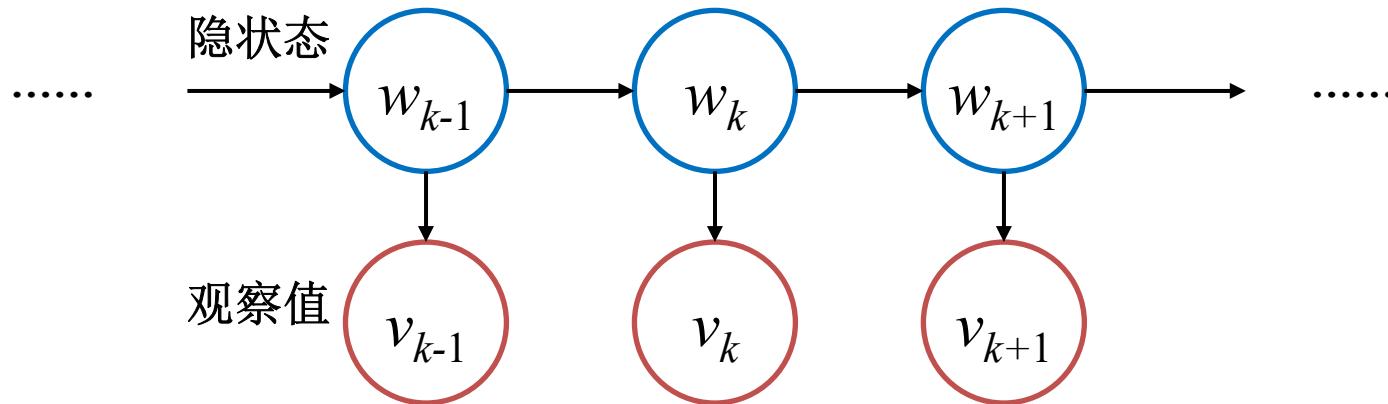
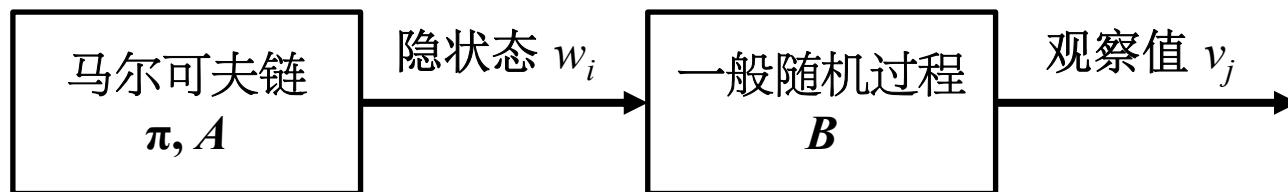
- HMM的完整参数为： $\theta = (\pi, A, B)$ ，其中
 - 初始状态概率 π 是 $M \times 1$ 的矢量
 - 状态转移概率矩阵 A 是 $M \times M$ 的方阵
 - 状态输出概率矩阵 B 是 $M \times K$ 的矩阵

隐含马尔科夫模型

- HMM的状态是不确定或不可见的，只有通过观测序列的随机过程才能表现出来。
- 观察到的事件与状态并不是一一对应，而是通过一组概率分布相联系。
- HMM是一个双重随机过程，两个组成部分：
 - 马尔可夫链（Markov Chain）：描述状态的转移，用转移概率描述。
 - 一般随机过程：描述状态与观察序列之间的关系，用状态输出概率描述。

隐含马尔科夫模型

- HMM的基本结构：



HMM的三个基本问题

- **估值问题**: 已有一个HMM, 其参数已知, 计算这个模型输出特定的观察序列 V^T 的概率;
- **解码问题**: 已有一个HMM, 其参数已知, 计算最有可能输出特定的观察序列 V^T 的隐状态转移序列 W^T ;
- **学习问题**: 已知一个HMM的结构, 其参数未知, 根据一组观察序列对参数进行训练。

估值问题

- **估值问题**: 如何计算一个已知参数的HMM输出特定观察序列 V^T 概率。
 - 例: 已知天气变化可用前例中的马尔可夫模型描述, 某人活动包括{散步、购物、做家务}, 活动和天气间的相关性用状态输出概率表示, 计算连续三天活动依次为散步、做家务和购物的概率。
- 待识模式以观察序列的形式出现, 而每个类别的条件概率(密度) 则由不同的HMM所描述。
- 当输入一个待识模式时, 需要计算每个类别对应的HMM 输出这个观察序列的概率, 然后根据贝叶斯判别准则进行分类。

估值问题

- 一个HMM产生观察序列 V^T 可由下式计算：

$$P(V^T | \theta) = \sum_{r=1}^{r_{\max}} P(V^T, W_r^T | \theta) = \sum_{r=1}^{r_{\max}} P(V^T | W_r^T, \theta) P(W_r^T | \theta)$$

其中，

$r_{\max} = M^T$ 是HMM所有可能的状态转移序列数；

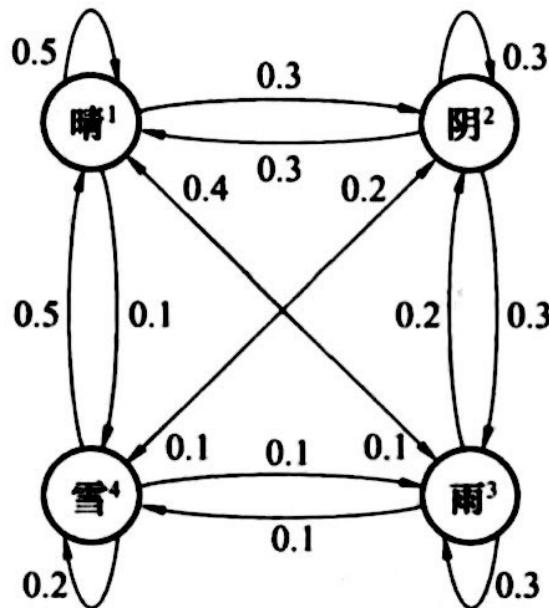
$P(V^T | W_r^T, \theta) = b_{w_r(1)v(1)} b_{w_r(2)v(2)} \cdots b_{w_r(T)v(T)}$ 为状态转移序列 W_r^T 输出观察序列 V^T 的概率；

$P(W_r^T | \theta) = \pi_{w_r(1)} a_{w_r(1)w_r(2)} a_{w_r(2)w_r(3)} \cdots a_{w_r(T-1)w_r(T)}$ 为状态转移序列 W_r^T 发生的概率。

估值问题

- 例：课本p184 【例7.5】（细节略）

某人的活动{散步、购物、做家务}分别按照1-3编号，活动与天气相关性用矩阵 B 表示。计算连续三天活动为散步-做家务-购物的概率。



$$B = \begin{bmatrix} 0.5 & 0.4 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.1 & 0.3 & 0.6 \\ 0.1 & 0.2 & 0.7 \end{bmatrix}$$

估值问题

- 估值问题的计算复杂度为 $O(M^T \times T)$ 。当状态数 M 较多，序列的长度 T 较长时，计算量大。随着序列长度的增加，计算复杂度呈指数增长。
- 事实上，计算过程中有很多重复项只需要计算1次即可，这样就可以有效地减小计算量。据此可提出前向算法。

估值问题前向算法

- 前向算法：在每一个节点上定义 $\alpha_i(t)$ 值，表示 HMM 在第 t 时刻处于第 i 个状态，并且输出序列 $v(1), \dots, v(t)$ 的概率。

$$\alpha_i(1) = \pi_i b_{iv(1)}$$

$$\alpha_i(2) = \left[\sum_{j=1}^M \alpha_j(1) a_{ji} \right] b_{iv(2)}$$

⋮

$$\Rightarrow \alpha_3(3) = \left[\sum_{j=1}^M \alpha_j(2) a_{j3} \right] b_{3v(3)}$$

⋮

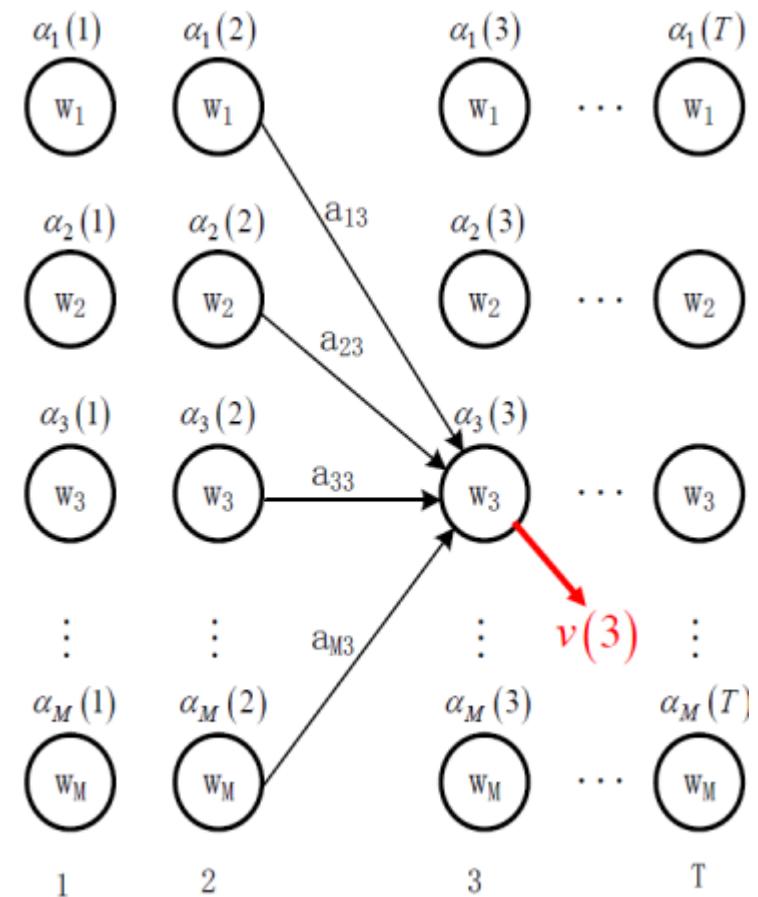
$$\alpha_i(t+1) = \left[\sum_{j=1}^M \alpha_j(t) a_{ji} \right] b_{iv(t+1)}$$

$$i = 1, \dots, M, \quad t = 1, \dots, T-1$$

例：

$$i = 3$$

$$t = 3$$



估值问题前向算法

- 前向算法：计算复杂度为 $O(M^2 \times T)$

 - 初始化： $t = 1$ ；
 - 计算第1列每个节点的 α 值： $\alpha_i(1) = \pi_i b_{iv(1)}, i = 1, \dots, M$ ；
 - 迭代计算第2至 T 列每个节点的 α 值：
$$\alpha_i(t+1) = \left[\sum_{j=1}^M \alpha_j(t) a_{ji} \right] b_{iv(t+1)}, i = 1, \dots, M$$
$$t = t + 1;$$
 - 输出： $P(V^T | \theta) = \sum_{i=1}^M \alpha_i(T)$
-

解码问题

- **解码问题**: 给定已知参数的HMM，计算最有可能产生出特定观察序列 V^T 的状态转移序列。
 - 例：在之前的天气与活动的例子中，如果知道此人连续三天的活动分别为散步、做家务和购物，推测这三天中该城市最有可能的天气状况。
- HMM 中的状态转移过程是不可见的，解码问题实际上关心的是如何根据模型输出的结果（观察序列）来推测产生出该结果的内部机理（状态转移序列）的问题。

解码问题

- 解码问题需要求解的是如下优化问题：

$$W^* = \arg \max_{W^T} P(W^T | V^T, \theta)$$

- 根据贝叶斯公式

$$P(W^T | V^T, \theta) = \frac{P(V^T | W^T, \theta) P(W^T | \theta)}{P(V^T | \theta)}$$

解码问题可以转化为求解如下的优化问题：

$$W^* = \arg \max_{1 \leq r \leq r_{\max}} P(V^T | W_r^T, \theta) P(W_r^T | \theta)$$

其中 W_r^T 取所有可能的 ($r_{\max} = M^T$ 种可能性) 长度为 T 的状态转移序列。

解码问题

- 同估值问题类似，只要计算出 r_{\max} 种可能的状态转移序列发生的概率 $P(W_r^T | \theta)$ ，以及在状态转移序列发生条件下产生出观察序列 V^T 的概率，寻找两项乘积的最大值就可以解决解码问题。
- 例：课本p186 【例7.6】
如果知道某人连续三天的活动是散步-做家务-购物，那么这三天最有可能的天气是什么？
解：计算所有64种连续三天天气的概率和每个天气状态下进行此三项活动的概率，然后比较两项概率乘积的最大值。

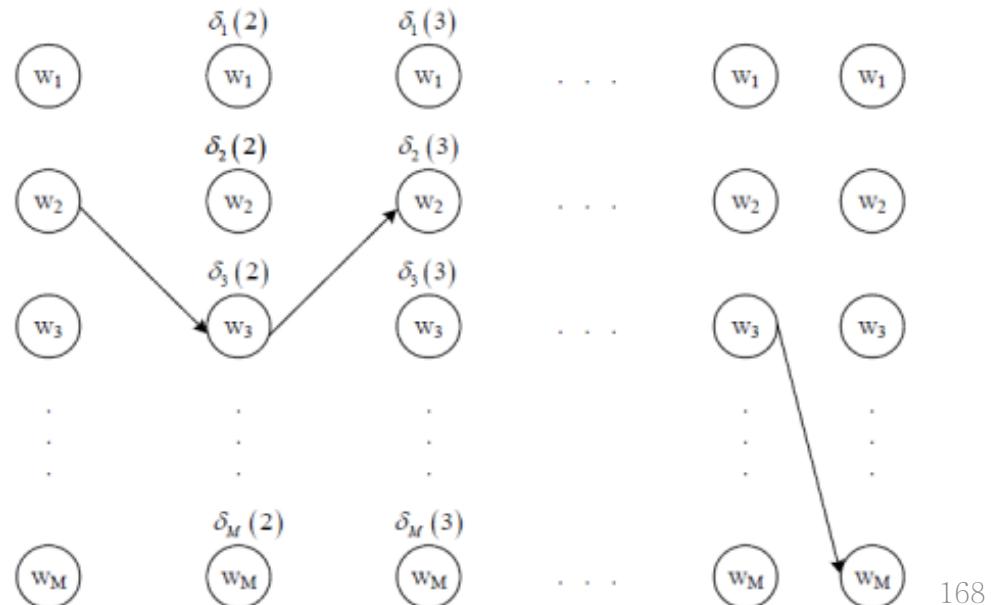
解码问题Viterbi算法

- 解码问题的计算复杂度也为 $O(M^T \times T)$ 。类似地，存在多项式时间复杂度的Viterbi算法。
- Viterbi算法：在每一个节点上定义 $\delta_i(t)$ 值，表示 HMM 在第 t 时刻处于第 i 个状态，并且输出序列 $v(1), \dots, v(t)$ 最优路径的概率值。

$$\delta_i(1) = \pi_i b_{iv(1)}$$

$$\delta_i(t+1) = \max_{1 \leq j \leq M} [\delta_j(t) a_{ji}] b_{iv(t+1)}$$

$$i = 1, \dots, M, \quad t = 1, \dots, T - 1$$



解码问题Viterbi算法

- 迭代计算出全部节点 δ 值之后，只需要在第 T 列找到 δ 值最大的节点，就可以确定最优状态转移序列在第 T 时刻的状态。
- 为了找到 1 至 $T - 1$ 时刻的最优状态，需要在迭代过程的每个节点上记录此节点最优路径上前一时刻的状态， $\varphi_i(t)$ 保存在第 t 时刻 HMM 处于第 i 个状态的最优路径上 $t - 1$ 时刻的状态。
- 在迭代计算 δ 值之后，确定了第 T 时刻的最优状态，可以根据节点的 φ 值回溯出整个的最优状态转移序列。

解码问题Viterbi算法

- Viterbi算法：

- 初始话： $t = 1$ ；
- 计算第1列每个节点的 δ 值：

$$\delta_i(1) = \pi_i b_{iv(1)}, \varphi_i(1) = 0, i = 1, \dots, M ;$$

- 迭代计算第2至 T 列每个节点的 δ 值：

$$\delta_i(t+1) = \max_{1 \leq j \leq M} [\delta_j(t) a_{ji}] b_{iv(t+1)}, i = 1, \dots, M$$

$$\varphi_i(t+1) = \arg \max_{1 \leq j \leq M} [\delta_j(t) a_{ji}], i = 1, \dots, M$$

$$t = t + 1 ;$$

- 最优路径的概率： $P^*(V^T | \theta) = \max_{1 \leq j \leq M} [\delta_j(T)]$

- 回溯最优路径： $w^*(T) = \arg \max_{1 \leq j \leq M} [\delta_j(T)]$

$$w^*(t) = \varphi_{w^*(t+1)}(t+1), t = T-1, \dots, 1$$

学习问题

- **学习问题**: 如何根据一组训练模式的观察序列集合 V 学习HMM参数 θ 的问题。
- HMM 描述的是观察序列发生的概率，因此对它的学习仍然是一个参数估计问题，可以采用最大似然估计的方法求解如下的优化问题：

$$\theta^* = \arg \max_{\theta} P(V | \theta)$$

- 由于状态转移序列是隐含的，因此HMM的参数需要采用EM 算法进行迭代估计。

学习问题

- Baum-Welch 算法：也称为前向后向算法，先设定一个转移概率的初值；然后获得对该初值的一个修正；反复迭代、直到收敛。
- 为计算 $\theta = (\pi, A, B)$ ，定义变量 α , β , γ :
- α 值：估值问题前向算法中定义了 α , $\alpha_i(t-1)$ 表示在 $t-1$ 时刻HMM处于状态 w_i , 并在 $1 \rightarrow t-1$ 期间产生出观察序列 $V^{1 \rightarrow t-1}$ 的概率。
- β 值：类似地在每个节点上还可以定义 β , $\beta_j(t)$ 表示在 t 时刻HMM 处于状态 w_j , 并且在 $t+1 \rightarrow T$ 期间产生出观察序列 $V^{t+1 \rightarrow T}$ 的概率。

学习问题Baum-Welch 算法

- α 值的计算同前向算法一样，由第1列向最后一列迭代：

$$\alpha_i(1) = \pi_i b_{iv(1)}$$

$$\alpha_i(t+1) = \left[\sum_{j=1}^M \alpha_j(t) a_{ji} \right] b_{iv(t+1)}$$

- β 值的计算与 α 类似，只不过是由最后一列向第1列迭代：

$$\beta_j(T) = 1$$

$$\beta_j(t) = \left[\sum_{i=1}^M \beta_i(t+1) a_{ji} \right] b_{jv(t+1)}$$

学习问题Baum-Welch 算法

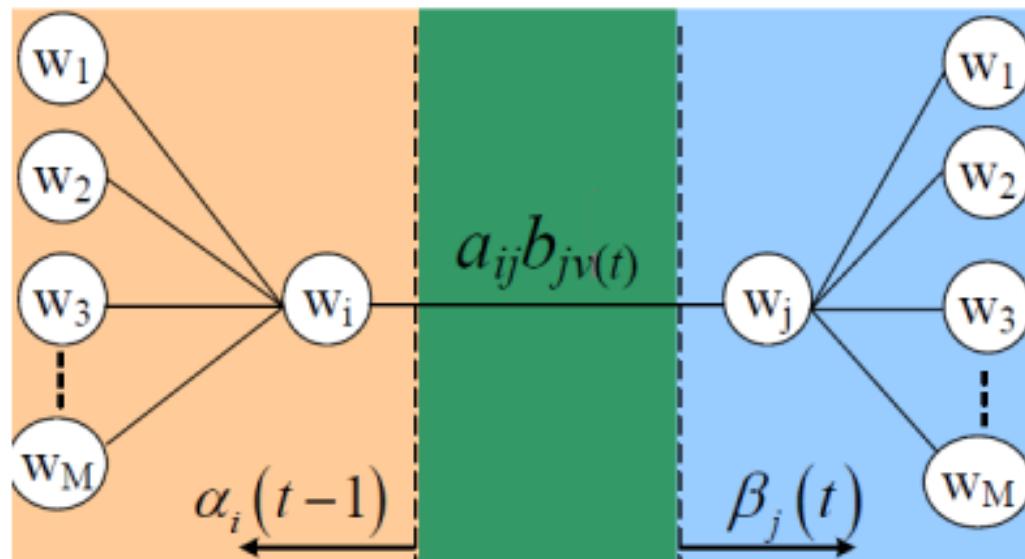
- γ 值: γ 定义在相邻两列任意两个节点之间; $\gamma_{ij}(t)$ 表示当HMM 输出观察序列 V^T 时, 在 $t - 1$ 时刻处于状态 w_i , 在 t 时刻处于状态 w_j 的概率。
- 根据 γ 定义和条件概率公式可得:

条件概率公式
 $p(A, B) = p(A | B)p(B)$

$$\begin{aligned}\gamma_{ij}(t) &= P[w(t-1) = w_i, w(t) = w_j | V^T, \theta] \\ &= \frac{P[w(t-1) = w_i, w(t) = w_j, V^T | \theta]}{P(V^T | \theta)} \\ &= \frac{\alpha_i(t-1)a_{ij}b_{jv(t)}\beta_j(t)}{P(V^T | \theta)}\end{aligned}$$

学习问题Baum-Welch 算法

- $\gamma_{ij}(t) = \frac{\alpha_i(t-1)a_{ij}b_{jv(t)}\beta_j(t)}{P(V^T | \theta)}$ 由三个独立部分组成：



$\alpha_i(t-1)$ 表示在 $t-1$ 时刻处于状态 w_i ，并在 $1 \rightarrow t-1$ 期间产生出观察序列 $V^{1 \rightarrow t-1}$ 的概率

$a_{ij}b_{jv(t)}$ 表示在 t 时刻由状态 w_i ，转移到 w_j ，并且输出 $v(t)$ 的概率

$\beta_j(t)$ 表示在 t 时刻处于状态 w_j ，并且在 $t+1 \rightarrow T$ 期间产生出观察序列 $V^{t+1 \rightarrow T}$ 的概率

学习问题Baum-Welch 算法

- 当 $t = 1$ 时， γ 值的计算公式为：

$$\gamma_j(1) = \frac{\pi_j b_{jv(1)} \beta_j(1)}{P(V^T | \theta)}$$

- 有了 γ 值可以得到模型参数估计的迭代公式。以下依次计算 π, A, B 。
- 根据全概率公式，初始概率 π_i 是在时刻 $t = 1$ 模型处于状态 w_i 的概率，因此 π_i 的迭代估计公式为

$$\pi_i = P[w(1) = w_i | V^T, \theta] = \gamma_i(1)$$

学习问题Baum-Welch 算法

- 在 1 至 T 时刻之间，HMM 由状态 w_i 转移到 w_j 的期望次数为： $\sum_{t=2}^T \gamma_{ij}(t)$ ，而由状态 w_i 转移到任意一个状态的期望次数为： $\sum_{t=2}^T \sum_{k=1}^M \gamma_{ik}(t)$ 。
- 因此 a_{ij} 的迭代估计公式为

$$a_{ij} = \frac{\sum_{t=2}^T \gamma_{ij}(t)}{\sum_{t=2}^T \sum_{k=1}^M \gamma_{ik}(t)}$$

学习问题Baum-Welch 算法

- 在 1 至 T 时刻之间，HMM 在状态 w_i 上输出观察值 v_k 的期望次数为： $\sum_{t=1, v(t)=v_k}^T \sum_{l=1}^M \gamma_{li}(t)$ ，而在状态 w_i 上输出任意观察值的期望次数为： $\sum_{t=1}^T \sum_{l=1}^M \gamma_{li}(t)$ 。
- 因此 b_{ik} 的迭代估计公式为

$$b_{ik} = \frac{\sum_{t=1, v(t)=v_k}^T \sum_{l=1}^M \gamma_{li}(t)}{\sum_{t=1}^T \sum_{l=1}^M \gamma_{li}(t)}$$

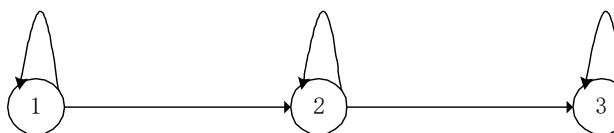
学习问题Baum-Welch 算法

- Baum-Welch 算法的全过程：
 - 初始化HMM参数 θ ，输入训练序列 V^T ；
 - 前向计算 α_i ；
 - 后向计算 β_j ；
 - 计算 γ_{ij} ；
 - 重新估计参数 θ ；
 - 迭代直到满足收敛条件为止。

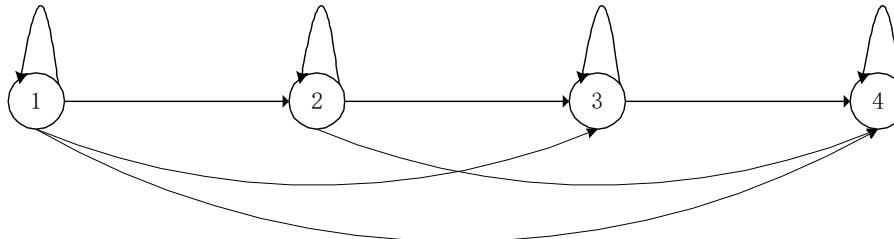
HMM的其它问题

- 连续HMM：在观察序列中每个观察值是一个特征矢量，相应的模型中输出概率就需要用一个概率密度函数描述，其函数形式需要假设，通常使用GMM。
- 模型的拓扑结构：模型结构可以根据实际问题的需要来设计，在初始化状态转移矩阵时，将某些元素设为0即可。

左右模型（语音识别常用）



带跨越的左右模型



本章小结

- 介绍概率密度函数估计的两类方法（非参数估计和参数估计）的特点和比较
- 介绍非参数估计的Parzen窗法的基本原理、识别算法和参数的影响
- 介绍非参数估计的近邻法的基本原理和分类算法
- 介绍参数估计的最大似然估计法的思路和解法

本章小结

- 介绍了高斯混合模型的思想和模型参数估计方法
(期望最大化算法)
- 介绍了期望最大化算法的主要步骤
- 介绍了贝叶斯估计的基本思想和参数估计方法，
并以一维高斯分布均值估计为例进行说明

本章小结

- 介绍了用于处理序列形式模式的马尔可夫模型与隐含马尔可夫模型的基本形式和原理
- 介绍了隐含马尔可夫模型的三个基本问题和对应算法：
 - 估计问题（前向算法）
 - 解码问题（Viterbi算法）
 - 学习问题（ Baum-Welch算法）