模式识别

Pattern Recognition

第5讲 统计分类器及其学习

概率论与统计学习

判别式模型（Discriminative Model）

将$\mathbf{x}$看做特征空间中的点

构建判别函数 $g(\mathbf{x})$ 来决定 $\mathbf{x}$ 属于哪个类别

关键在于计算 $\mathbf{x}$ 与训练样本$\mathbf{x}_i$ 间的相对位置（内积）关系

产生式模型（Generative Model）

将$\mathbf{x}$看做随机变量

根据 $\mathbf{x}$ 属于各类别 ω_i 的概率大小 $P\left(\omega_i \middle| \mathbf{x}\right)$，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 $P\left(\mathbf{x} \middle| \omega_i\right)$

常用的概率表示形式

先验概率： $P(\omega_i)$

后验概率： $P(\omega_i|\mathbf{x})$

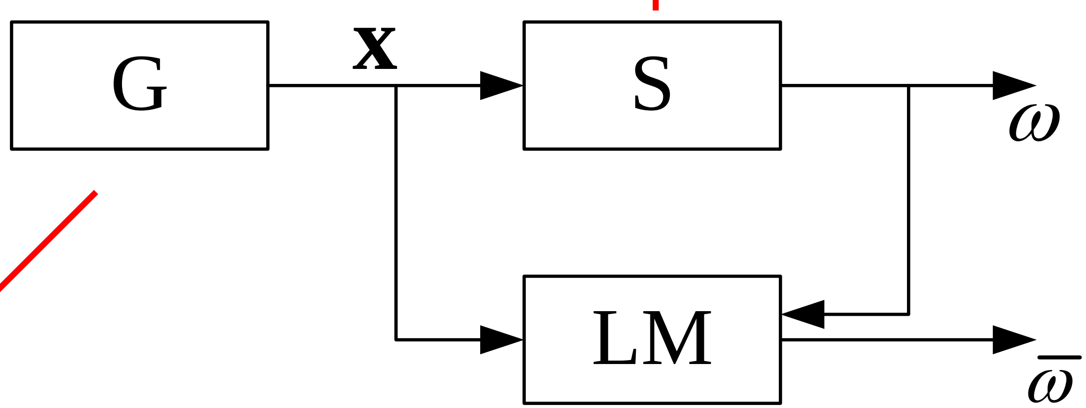
类条件概率： $P(\mathbf{x}|\omega_i)$

贝叶斯公式： $P(\omega_i|\mathbf{x}) = \dfrac{P(\mathbf{x}|\omega_i)P(\omega_i)}{P(\mathbf{x})}$

统计学习模型

S：目标算子（训练器）

将x变换成 ω，存在、不变、未知。



G：数据（实例）发生器

依据某一固定但未知的概率分布函数F(x)，独立同分布地产生向量x

$$F(\mathbf{x},\omega)=F(\mathbf{x})F(\omega|\mathbf{x})$$

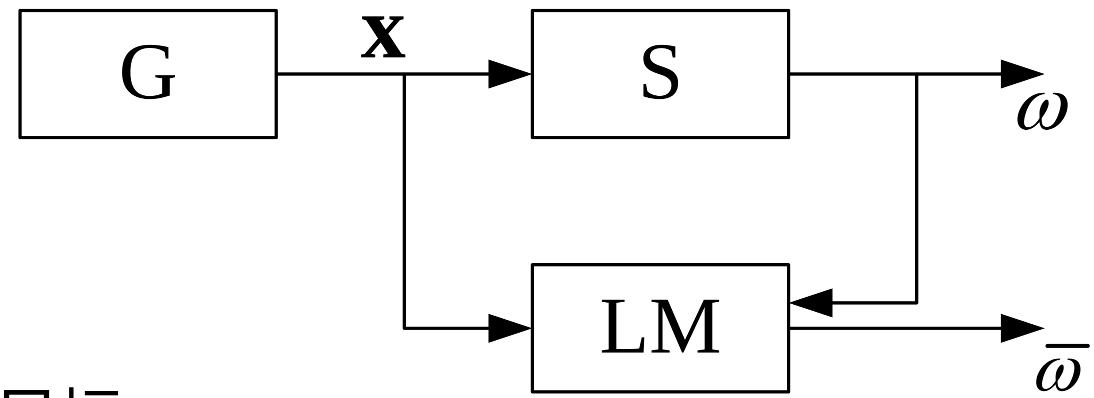
LM：学习机器

观测训练集$(\mathbf{x}_1,\omega_1)\cdots(\mathbf{x}_\ell,\omega_\ell)$

构造算子，预测特定向量$\mathbf{x}_i$的响应ω_i

学习过程：从给定的函数集中寻找一个适当的函数

统计学习模型



学习目标：

1，模仿训练器的算子：

对给定的发生器G，预测训练器S的输出

2，辨识训练器算子：

构造一个非常接近训练器算子S的算子

贝叶斯决策理论

各类别总体的概率分布已知

要决策分类的类别数一定

已知：分类问题有c个类别 $\omega_1, \omega_2, \cdots, \omega_c$

各类别的先验概率 $P(\omega_i)$

各类条件概率密度函数 $P(\mathbf{x}|\omega_i)$

决策：对于特征空间中观测到的向量 $\mathbf{x} = [x_1, x_2, \cdots, x_d]^t$

应该将x分到哪一类?

错误率最小 ⟶ 风险最小

贝叶斯决策理论

- 基于最小错误率的贝叶斯决策
- 基于最小风险的贝叶斯决策
- Neyman-Pearson决策规则: 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$P(\omega_i)$未知的情况下，使最大可能的风险最小化

多类问题最小错误率

- 判别x属于ω_i的错误率：

$$P(error|\mathbf{x}) = \sum_{j \neq i} P(\omega_j|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x})$$

- 判别准则为：

$$i = \underset{1 \le j \le c}{\arg\max} P(\omega_j|\mathbf{x}), \text{ 则：} \mathbf{x} \in \omega_i$$

最大后验概率

贝叶斯最小错误率准则

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

$$g_j(\mathbf{x}) = p(\mathbf{x}|\omega_j)P(\omega_j)$$

$$g_j(\mathbf{x}) = \ln p(\mathbf{x}|\omega_j) + \ln P(\omega_j)$$

Bayes判别准则：

$$i = \arg\max_{1 \le j \le c} g_j(\mathbf{x}) \text{ ，则 } \mathbf{x} \in \omega_i$$

例：

对一大批人进行癌症普查，设ω1类代表患癌症，ω2类代表正常人。已知先验概率：

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

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：

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

现有一人化验结果为阳性，问此人是否患癌症?

最小错误率准则： $g_j(\mathbf{x}) = p(\mathbf{x} | \omega_j) P(\omega_j)$

基于最小风险的贝叶斯决策

判断某人是正常($\omega_{正}$)还是患者($\omega_{患}$),将出现以下情况:

实际类别 ω_i	决策类别 ω_j	风险因子 λ_{ij}
正常人	正常人	0
正常人	患者	0.25
患者	正常人	1
患者	患者	0

何种决策风险最小?

做出正常判决的风险: $R_{正}(\mathbf{x}) = \lambda_{正\to正} P(\omega_{正}|\mathbf{x}) + \lambda_{患\to正} P(\omega_{患}|\mathbf{x})$

做出患病判决的风险: $R_{患}(\mathbf{x}) = \lambda_{正\to患} P(\omega_{正}|\mathbf{x}) + \lambda_{患\to患} P(\omega_{患}|\mathbf{x})$

做出j类判决的风险: $R_j(\mathbf{x}) = \sum_{i=1}^{c} \lambda_{ij} P(\omega_i|\mathbf{x})$

——计算各类决策风险，选则最小风险对应的决策

最小平均风险准则贝叶斯分类器

- 有c个类别$\omega_1, \omega_2, \ldots, \omega_c$, 将$\omega_i$类的样本判别为$\omega_j$类的代价为$\lambda_{ij}$。
- 将未知模式x判别为ω_j类的平均风险为：

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

- 利用Bayes公式:

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

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

- 构造判别函数：

$$g_j(\mathbf{x}) = -\gamma_j(\mathbf{x})$$

例

- 对一大批人进行癌症普查，设ω_1类代表患癌症，ω_2类代表正常人。已知先验概率：

$$P(\omega_1)=0.004, P(\omega_2)=0.996$$

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：

$$P(x=\text{阳性}|\omega_1)=0.96, P(x=\text{阳性}|\omega_2)=0.01$$

判别代价： $\lambda_{11} = 0, \lambda_{22} = 0, \lambda_{12} = 95, \lambda_{21} = 15$

现有一人化验结果为阳性，问此人是否患癌症？

贝叶斯最小风险判别：

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

贝叶斯分类器的其它版本

- 约束一定错误率（风险）：Neyman-Pearson准则；
- 先验概率$P(\omega_i)$未知：极小化极大准则；
- 某些特征缺失的决策：
- 连续出现的模式之间统计相关的决策：

多元正态类条件概率密度

$$p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

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

正态分布的判别函数

贝叶斯判别函数可以写成对数形式：

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

$$\text{其中} p(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right]$$

类条件概率密度函数为正态分布时：

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

情况一： $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}, P(\omega_i) = \frac{1}{c}$ 距离分类器

- 判别函数可以写成:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

$$\Downarrow$$

$$g_i(\mathbf{x}) = -(\mathbf{x}-\boldsymbol{\mu}_i)^t(\mathbf{x}-\boldsymbol{\mu}_i) = -\|\mathbf{x}-\boldsymbol{\mu}_i\|^2$$

- 此分类器称为距离分类器，判别函数可以用待识模式$\mathbf{x}$与类别均值$\boldsymbol{\mu}_i$之间的距离表示:

$$g_i(\mathbf{x}) = -d(\mathbf{x}, \boldsymbol{\mu}_i)$$

情况二： $\boldsymbol{\Sigma}_i=\boldsymbol{\Sigma}$ 线性分类器

判别函数可以写成：

$$g_i(\mathbf{x})=-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)-\frac{d}{2}\ln 2\pi-\frac{1}{2}\ln|\boldsymbol{\Sigma}_i|+\ln P(\omega_i)$$

$$\Downarrow$$

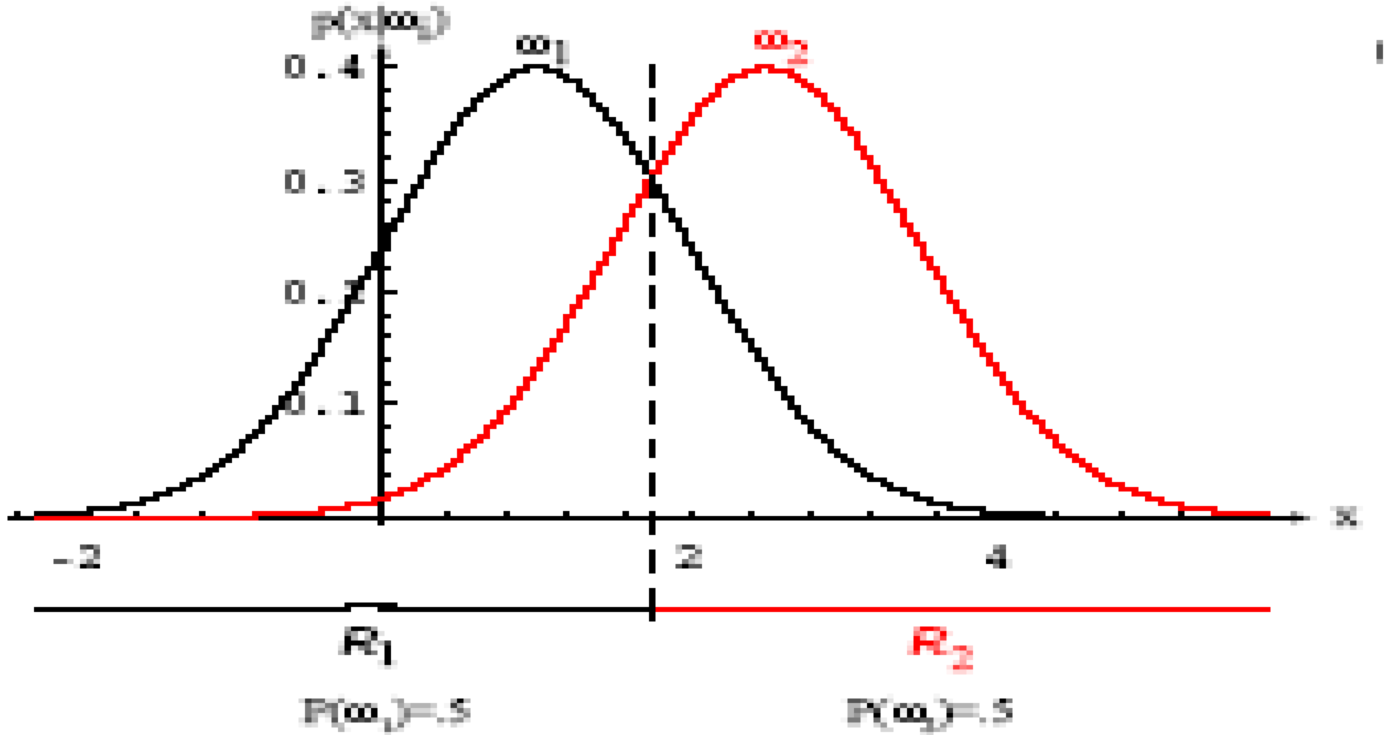
$$g_i(\mathbf{x})=-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)+\ln P(\omega_i)$$

$$\Downarrow$$

$$g_i(\mathbf{x})=\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\mathbf{x}-\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i+\ln P(\omega_i)=\mathbf{w}_i^t\mathbf{x}+w_{i0}$$

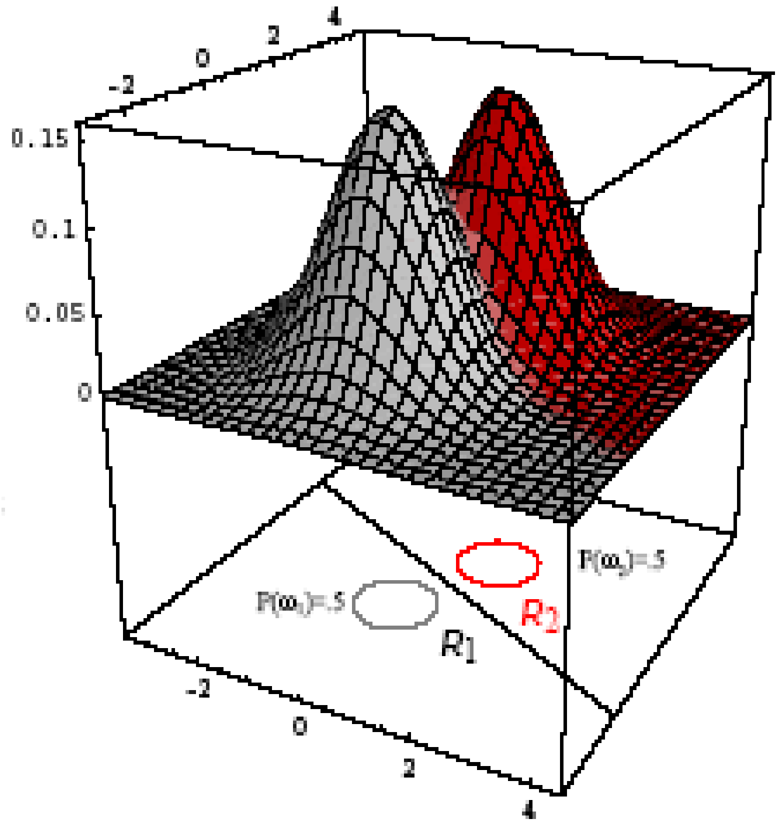
线性分类器

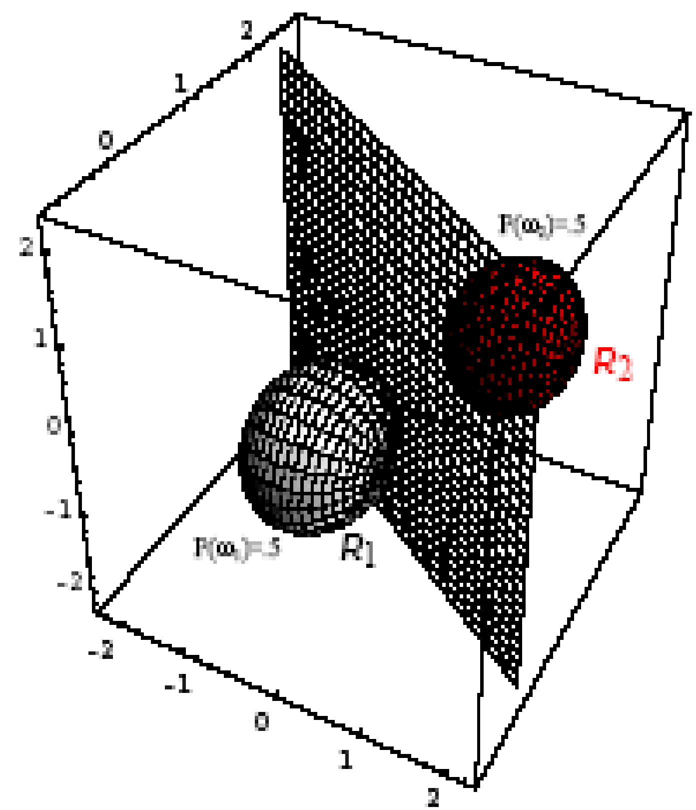
- 两类问题，1维特征，先验概率相同时：



线性分类器

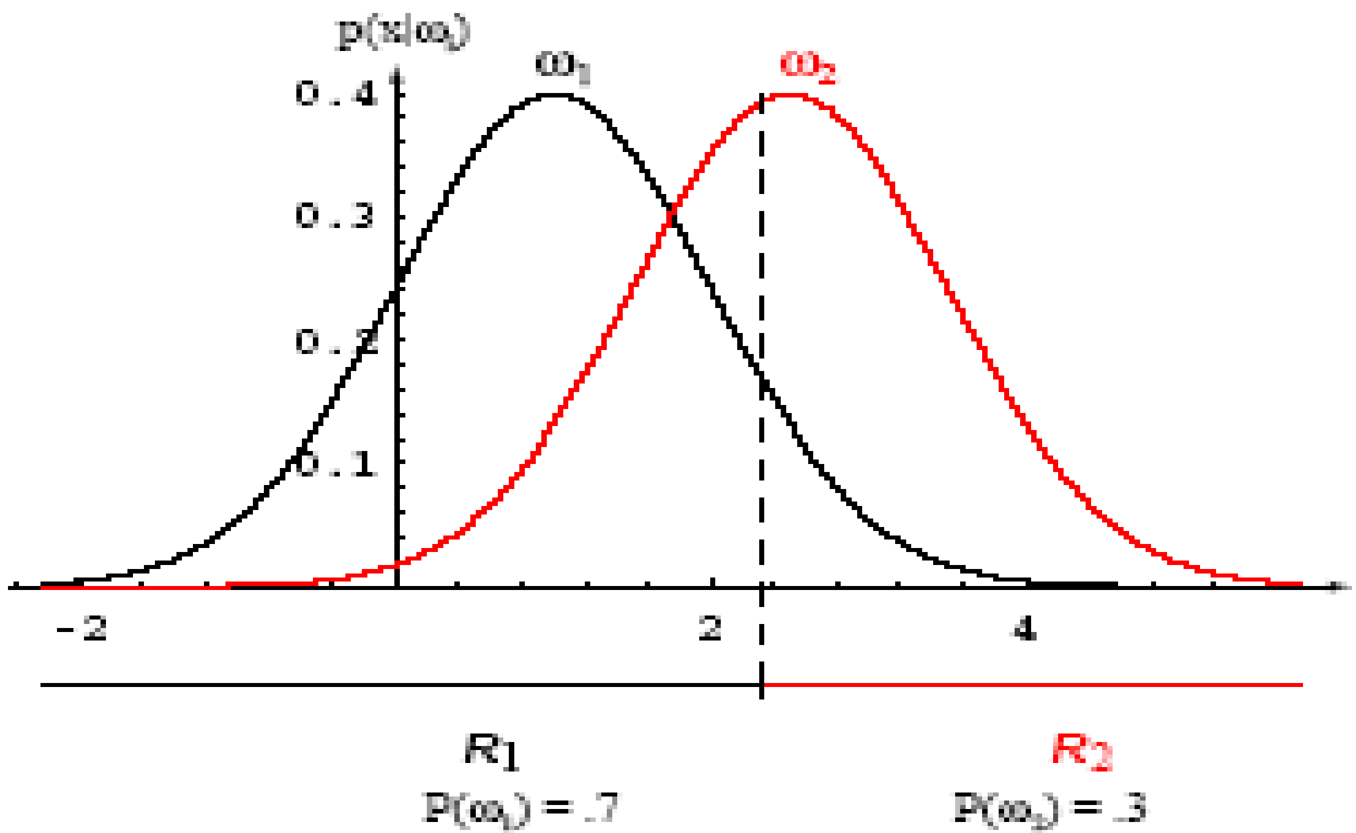
两类问题，高维特征，先验概率相同时：





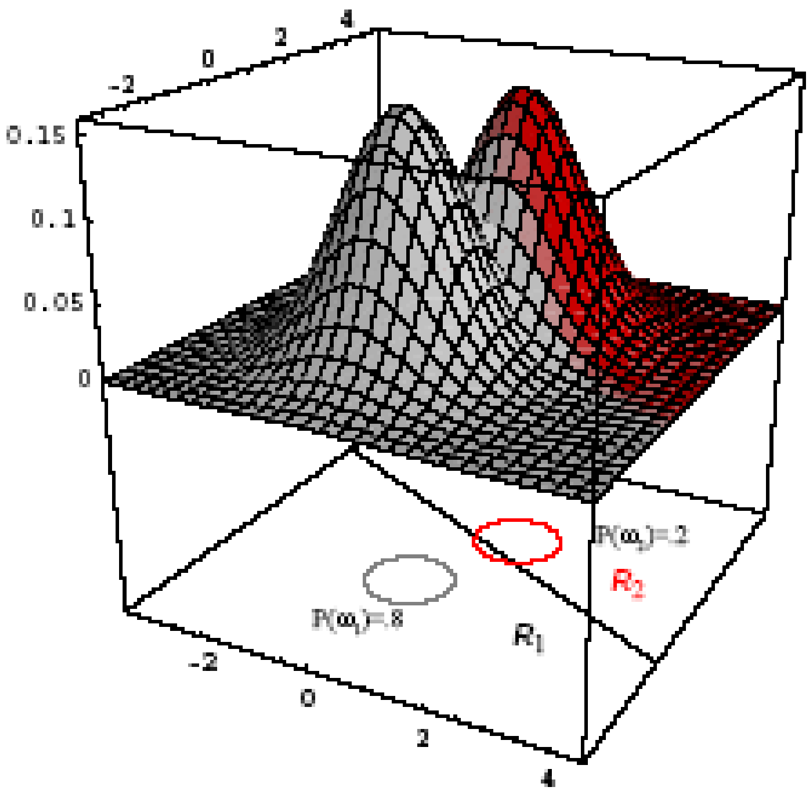
线性分类器

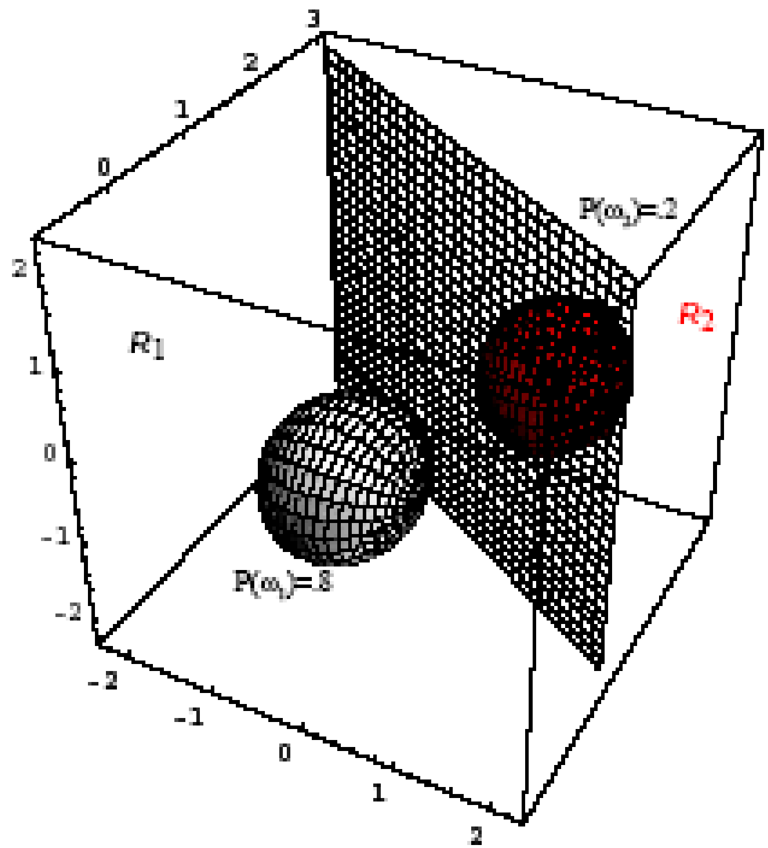
两类问题，1维特征，先验概率不同时：



线性分类器

两类问题，高维特征，先验概率不同时：





情况三：$\boldsymbol{\Sigma}_i$ 任意

判别函数可以写成：

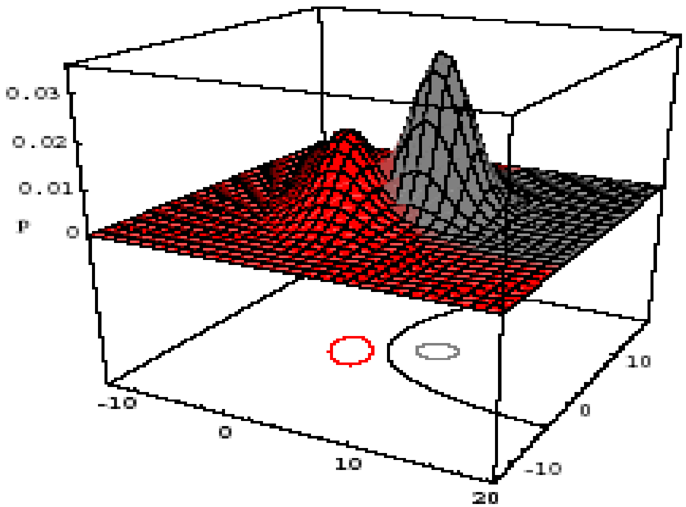
$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

$$\Downarrow$$

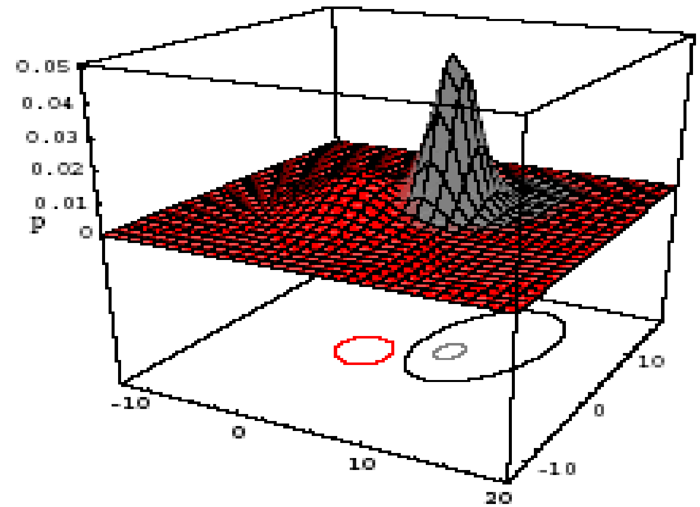
$$g_i(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^t\boldsymbol{\Sigma}_i^{-1}\mathbf{x} + \boldsymbol{\mu}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{x} + \left(-\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)\right)$$

分类界面为2次曲线（面）

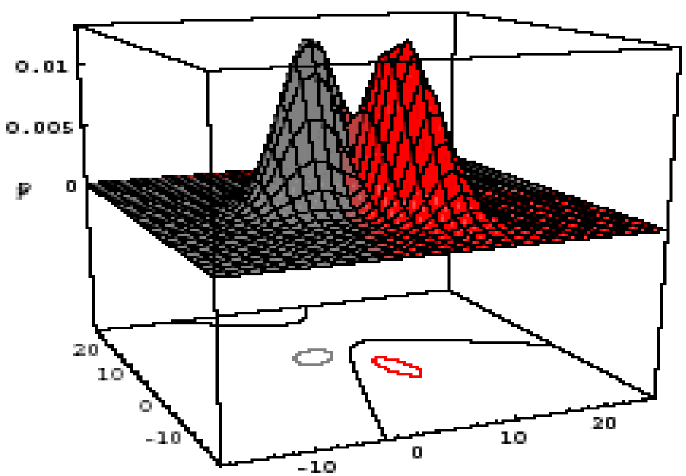
二次分类曲线



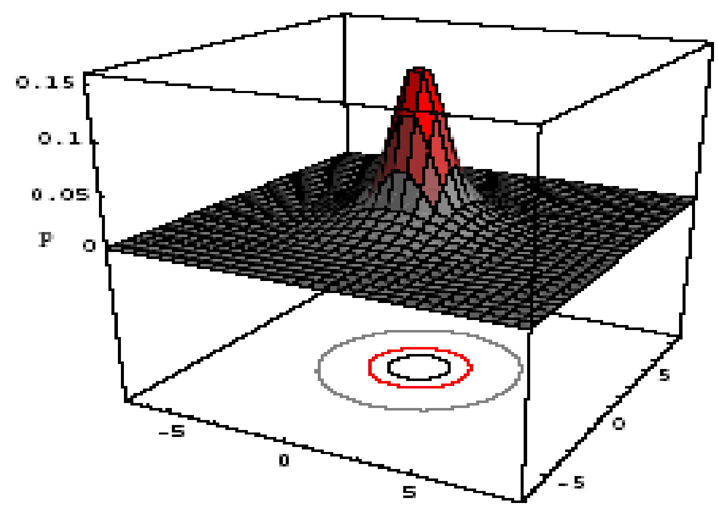
0.03
0.02
0.01
P
0
-10
0
10
20
10
0
-10



0.05
0.04
0.03
0.02
0.01
P
0
-10
0
10
20
10
0
-10

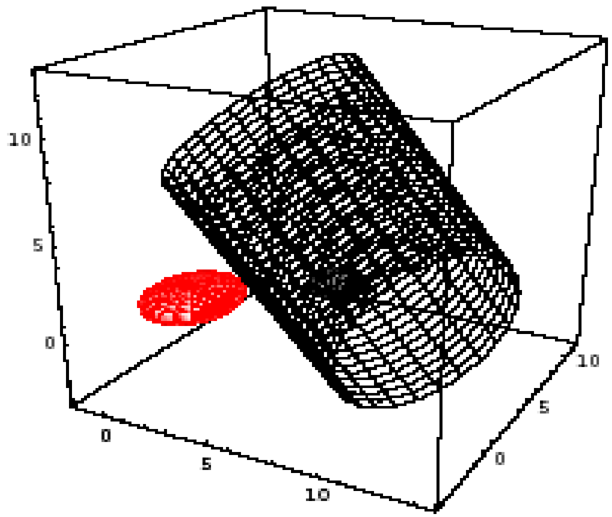


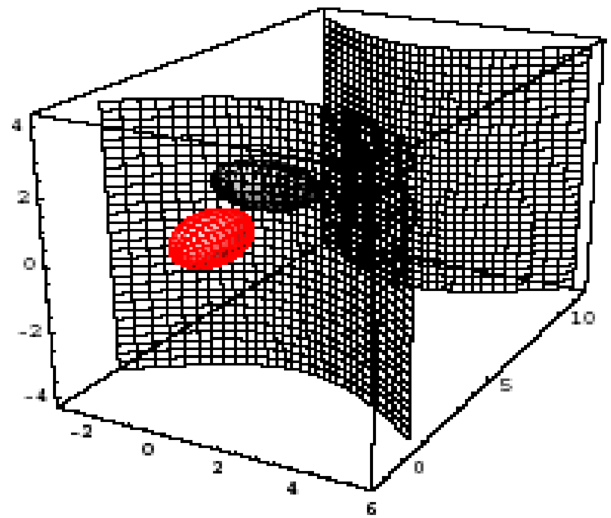
0.01
0.005
P
0
20
10
0
-10
-10
0
10
20

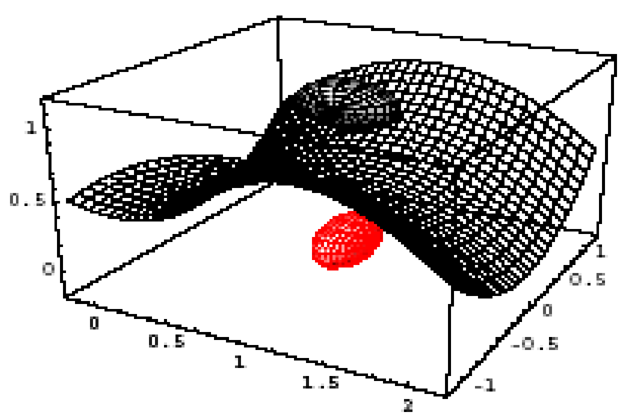


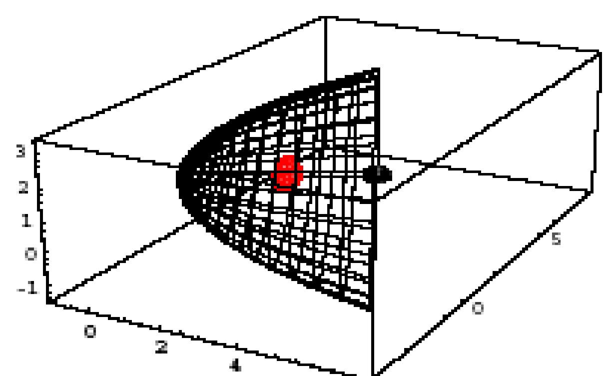
0.15
0.1
0.05
P
0
-5
0
5
5
0
-5

二次分类曲面


10
5
0
0
5
10
0
5
10


4
2
0
-2
-4
-2
0
2
4
6
0
5
10


1
0.5
0
0
0.5
1
1.5
2
-1
-0.5
0
0.5
1


3
2
1
0
-1
0
2
4
6
0
5

例:正态分布二次分类曲面举例

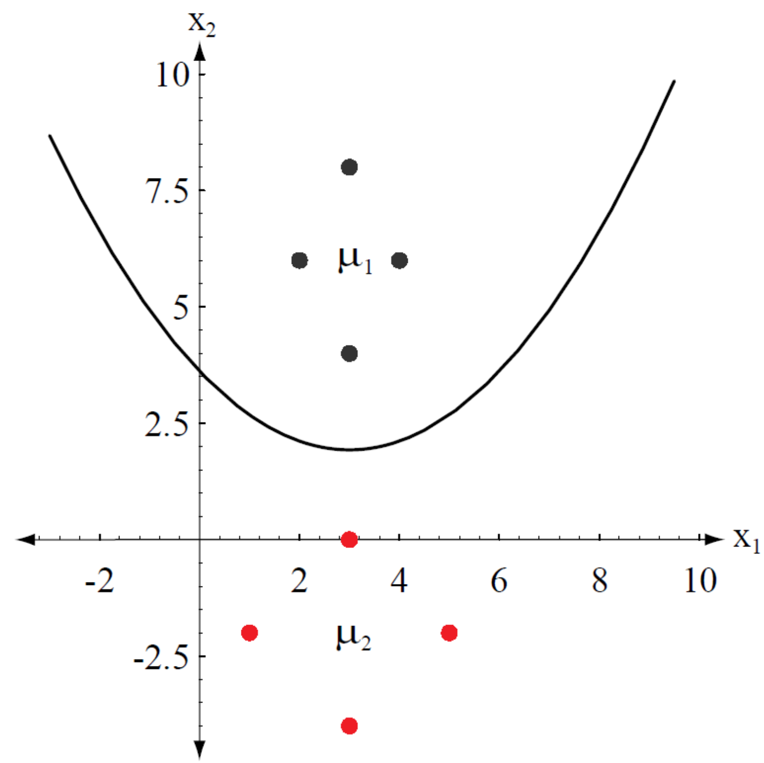
$$g_i(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^t\boldsymbol{\Sigma}_i^{-1}\mathbf{x} + \boldsymbol{\mu}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{x} + \left(-\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right| + \ln P(\omega_i)\right)$$

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

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

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

$$\boldsymbol{\Sigma}_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \quad \boldsymbol{\Sigma}_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}$$



$$g_1(\mathbf{x}) = g_2(\mathbf{x}) \quad \Rightarrow \quad x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2$$

朴素贝叶斯分类器

难估计协方差矩阵：特征的维数较高、训练样本数量较少时，无法有效估计协方差矩阵

假设各维独立：假设各维特征服从相互独立的高斯分布

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

x_j：第j维特征

ω_i：第i类别

对数判别函数为

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

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

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

引言

进行Bayes决策需要事先知道两种知识：

- 各类的先验概率；
- 观测向量的类条件概率密度。

知识的获取（估计）：

- 一些训练数据；
- 对问题的一般性的认识

引言

类的先验概率的估计（较容易）：

- 依靠经验；
- 用训练数据中各类出现的频率估计。
- 用频率估计概率的优点：
  - 无偏性；
  - 相合性；
  - 收敛速度快。

引言

类条件概率密度的估计（非常难）：

- 概率密度函数包含了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函数：

$$p(\mathbf{x}) \geq 0 \qquad \int p(\mathbf{x}) d\mathbf{x} = 1$$

- 问题可以表示为：已有c个类别的训练样本集合D1，D2，…，Dc，求取每个类别的类条件概率密度 $p(\mathbf{x}|\omega_i)$。

引言

概率密度估计的两种主要思路：

参数估计：

根据对问题的一般性的认识，假设随机变量服从某种分布，分布函数的参数通过训练数据来估计。

例如，假定x服从正态分布$N(\mu,\Sigma)$,要估计的参数就是$\theta=(\mu,\Sigma)$

非参数估计：

不用模型，而只利用训练数据本身对概率密度做估计。

-----K近邻分类器

非参数估计的基本思想

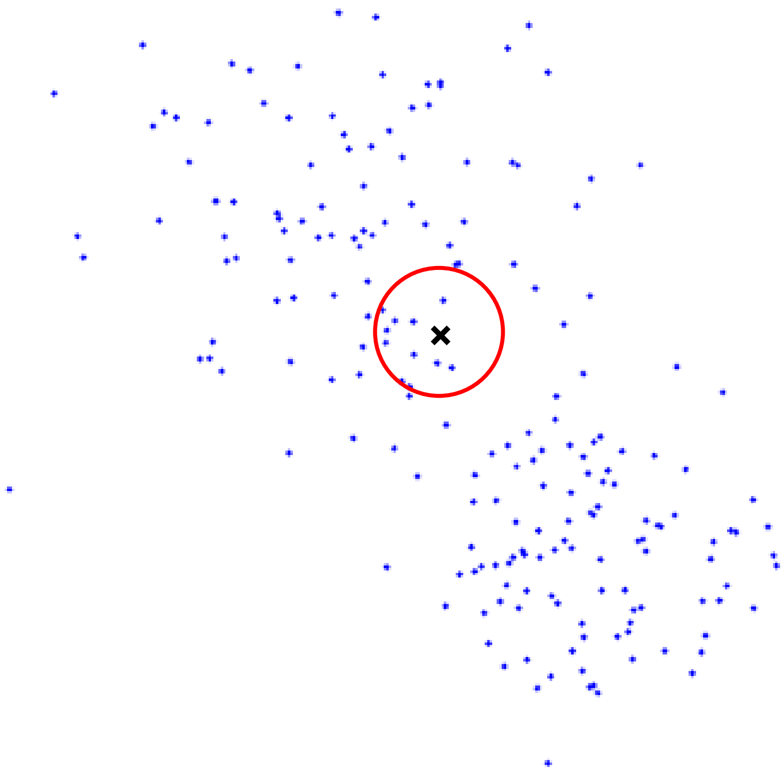
- 令R是包含样本点x的一个区域，其体积为V，设有n个训练样本，其中有k个落在区域R中，则可对概率密度作出一个估计：

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

相当于用**R**区域内的平均性质来作为一点x的估计，是一种数据的平滑。

非参数估计基本思想

根据训练样本，直接估计概率密度



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，不够精确；
过小则可能导致在此区域内无样本点，k=0。

此方法的有效性取决于样本数量的多少，以及区域体积选择的合适。

区域选定的两个途径

- Parzen窗法：区域体积V是样本数n的函数，如：

$$V_n = \frac{1}{\sqrt{n}}$$

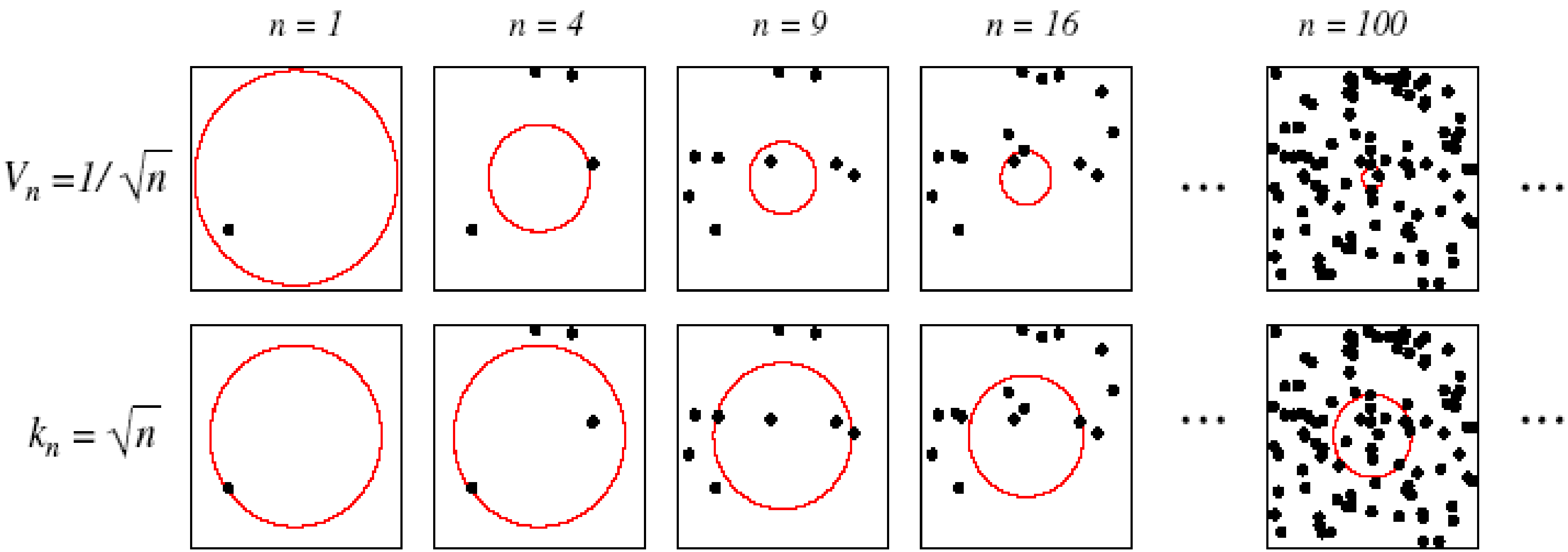
估计类条件概率密度

K-近邻法：落在区域内的样本数k是总样本数n的函数，如：

$$k_n = \sqrt{n}$$

直接估计后验概率

Parzen窗法和K-近邻法



K近邻的估计原理

将一个体积为V的区域放到待识样本点x周围，包含k个训练样本点，其中k_i个属于ω_i类，总的训练样本数为n，则有：

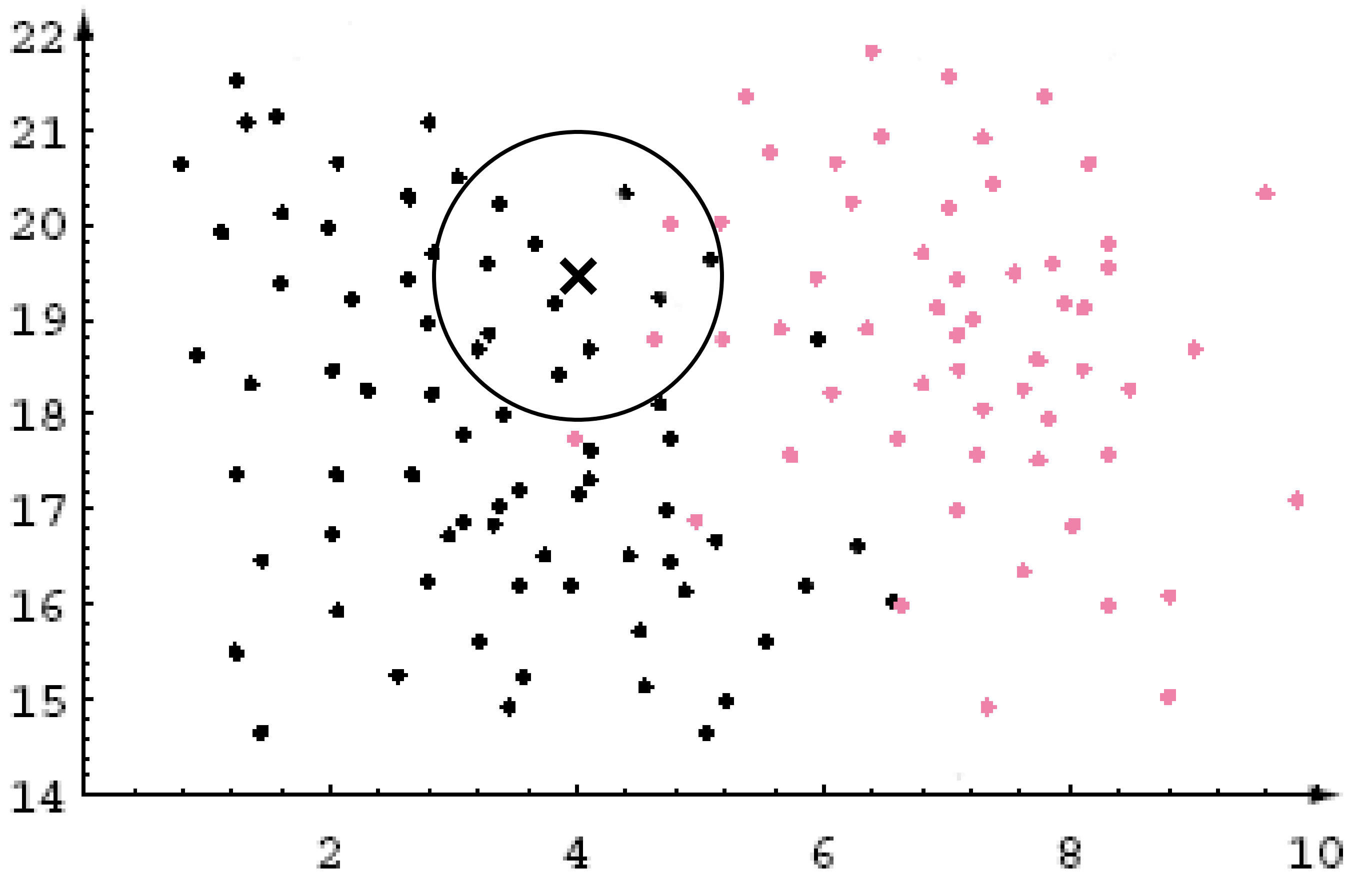
$$p_n(\mathbf{x},\omega_i)=\frac{k_i/n}{V}$$

$$p(\omega_i|\mathbf{x})=\frac{p_n(\mathbf{x},\omega_i)}{p_n(\mathbf{x})}=\frac{p_n(\mathbf{x},\omega_i)}{\sum_{j=1}^{c}p_n(\mathbf{x},\omega_j)}=\frac{k_i}{k}$$

k-近邻分类算法

1. 设置参数k，输入待识别样本$\mathbf{x}$；
2. 计算x与每个训练样本的距离；
3. 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；
4. $class \leftarrow \arg\max_{1 \le i \le c} k_i$

k-近邻分类，k=13



概率密度函数的参数估计方法

预先假设每一个类别的概率密度函数的形式已知，而具体的参数未知；

- 最大似然估计(MLE, Maximum Likelihood Estimation)；
- 贝叶斯估计(Bayesian Estimation)。

似然函数

样本集D中包含n个样本：$\mathbf{x}_1$，$\mathbf{x}_2$， $\ldots$, $\mathbf{x}_n$，样本都是独立同分布的随机变量，样本集D出现的概率为：

$$p\left(D\middle|\boldsymbol{\theta}\right)=p\left(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n\middle|\boldsymbol{\theta}\right)=\prod_{i=1}^{n}p\left(\mathbf{x}_i\middle|\boldsymbol{\theta}\right)$$

定义对数似然函数：

$$l\left(\boldsymbol{\theta}\right)\equiv\ln p\left(D\middle|\boldsymbol{\theta}\right)=\sum_{i=1}^{n}\ln p\left(\mathbf{x}_i\middle|\boldsymbol{\theta}\right)$$

最大似然估计

- 寻找到一个最优矢量 $\hat{\boldsymbol{\theta}}$，使得似然函数 $l(\boldsymbol{\theta})$最大。

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

$$l(\boldsymbol{\theta}) \equiv \ln p(D|\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln p(\mathbf{x}_i|\boldsymbol{\theta})$$

$$\nabla_{\boldsymbol{\theta}} l = \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_i|\boldsymbol{\theta})$$

$$\nabla_{\boldsymbol{\theta}} l = 0$$

正态分布的似然估计

- Gauss分布的参数由均值矢量 $\boldsymbol{\mu}$ 和协方差矩阵 $\boldsymbol{\Sigma}$ 构成，最大似然估计结果为：

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

$$\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\underline{\left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}\right)\left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}\right)^t}$$

设样本集$X=\{x_1,\cdots,x_n\}$服从 **Rayleigh** 分布：

$$p(x|\theta)=\begin{cases}2\theta x e^{-\theta x^2}, & x\ge 0\\ 0, & x<0\end{cases}$$

其中θ是一个未知参数，试推导参数θ的最大似然估计。

构造对数似然函数：

$$l(\theta)=\sum_{i=1}^{n}\ln p(x_i|\theta)=\sum_{i=1}^{n}\left(\ln 2+\ln\theta+\ln x_i-\theta x_i^2\right)$$

求对数似然函数关于θ的极值：

$$\frac{dl(\theta)}{d\theta}=\sum_{i=1}^{n}\left(\frac{1}{\theta}-x_i^2\right)=0$$，解得：$$\theta=\frac{1}{\frac{1}{n}\sum_{i=1}^{n}x_i^2}$$

贝叶斯估计

已有独立同分布训练样本集D；

已知类条件概率密度函数$p(x|\theta)$的形式，但参数θ未知；

已知参数θ的先验概率密度函数$p(\theta)$；

求在已有训练样本集D的条件下，类条件概率密度函数$p(x|D)$。

贝叶斯估计与最大似然估计的差别

最大似然估计：θ是一个确定的未知矢量；

贝叶斯估计：　θ是一个随机变量，θ以一定的概率分布 $p(\theta)$ 所有可能的值

贝叶斯估计的一般理论

类条件概率：

$$p(\mathbf{x}|D)=\int p(\mathbf{x},\boldsymbol{\theta}|D)d\boldsymbol{\theta}$$

$$=\int p(\mathbf{x}|\boldsymbol{\theta},D)p(\boldsymbol{\theta}|D)d\boldsymbol{\theta}=\int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|D)d\boldsymbol{\theta}$$

X与D独立

参数θ的后验分布：

$$p(\boldsymbol{\theta}|D)=\frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}=\frac{\prod_{i=1}^{n}p(\mathbf{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int\prod_{i=1}^{n}p(\mathbf{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

归一化系数

贝叶斯估计的一般理论

学习过程：计算参数的后验分布：

$$p(\boldsymbol{\theta}|D)=\frac{p(D|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{\int p(D|\boldsymbol{\theta})\,p(\boldsymbol{\theta})\,d\boldsymbol{\theta}}=\frac{\prod_{i=1}^{n}p(\mathbf{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int\prod_{i=1}^{n}p(\mathbf{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})\,d\boldsymbol{\theta}}$$

分类过程：将待识模式x和后验概率，计算x发生的概率

$$p(\mathbf{x}|D)=\int p(\mathbf{x},\boldsymbol{\theta}|D)\,d\boldsymbol{\theta}=\int p(\mathbf{x}|\boldsymbol{\theta})\,p(\boldsymbol{\theta}|D)\,d\boldsymbol{\theta}$$

举例：单变量正态分布的贝叶斯估计

已知概率密度函数满足正态分布，其中方差 σ^2已知，均值 μ未知，假设 μ的先验概率满足正态分布，即：

$$p(x|\mu) \sim N(\mu, \sigma^2) \qquad p(\mu) \sim N(\mu_0, \sigma_0^2)$$

在已知训练样本集合$D=\{x_1,x_2,\ldots,x_n\}$的条件下，估计x的概率密度函数：

μ_0 : 对μ的最好先验推测

σ_0^2 : 度量推测的不确定性

$$p(x|D) = \int p(x|\mu)\, p(\mu|D)\, d\mu$$

其中：
$$p(\mu|D) = \underbrace{\frac{p(D|\mu)\, p(\mu)}{\int p(D|\mu)\, p(\mu)\, d\mu}}_{} = \alpha \prod_{i=1}^{n} p(x_i|\mu)\, p(\mu)$$

独立同分布

归一化系数

$$p(\mu|D)=\frac{p(D|\mu)\,p(\mu)}{\int p(D|\mu)\,p(\mu)\,d\mu}=\alpha\prod_{i=1}^{n}p(x_i|\mu)\,p(\mu)$$

$$=\alpha\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)$$

$$=\alpha'\exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2+\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]\right\}$$

$$=\alpha''\exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i+\frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right]$$

$$p(\mu|D)=\alpha''\exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i+\frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right] \tag{1}$$

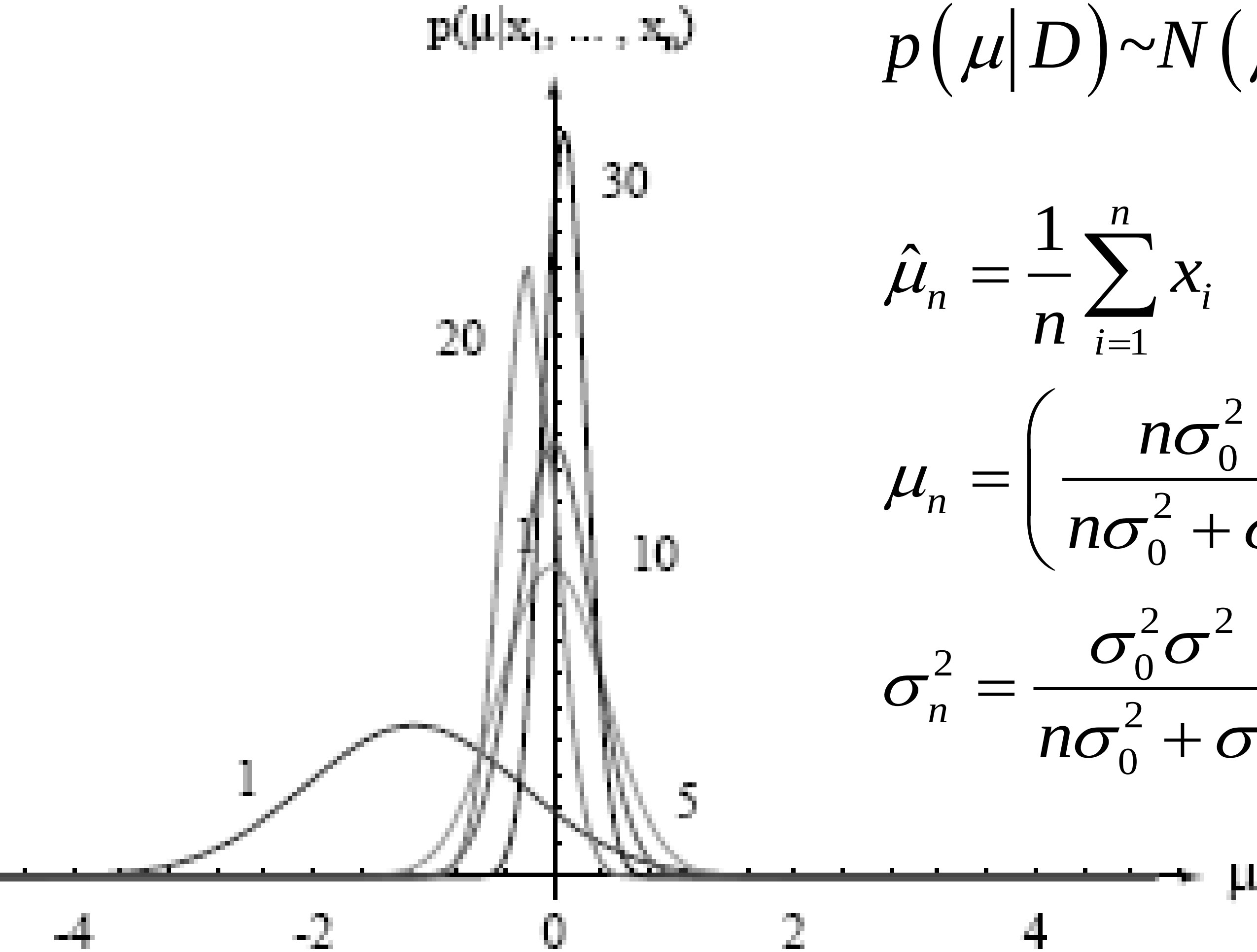
$p(\mu|D)$是μ 的二次函数的指数函数，仍是一个正态密度：

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

联立（1），（2）式，令对应项相等，有：

$$\begin{cases}\dfrac{1}{\sigma_n^2}=\dfrac{n}{\sigma^2}+\dfrac{1}{\sigma_0^2}\\[2ex]\dfrac{\mu_n}{\sigma_n^2}=\dfrac{n}{\sigma^2}\underbrace{\dfrac{1}{n}\sum_{i=1}^{n}x_i}_{\equiv\hat{\mu}_n}+\dfrac{\mu_0}{\sigma_0^2}\end{cases}\Rightarrow\begin{cases}\hat{\mu}_n=\dfrac{1}{n}\sum_{i=1}^{n}x_i\\[2ex]\mu_n=\left(\dfrac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)m_n+\dfrac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0\\[2ex]\sigma_n^2=\dfrac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}\end{cases}$$

均值分布的变化



$$p(\mu|D) \sim N(\mu_n, \sigma_n)$$

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

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

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

类条件概率密度的计算

$$p(x|D) = \int p(x|\mu)\, p(\mu|D)\, d\mu$$

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

$$= \frac{f(\sigma,\sigma_n)}{2\pi\sigma\sigma_n} \exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right]$$

其中，$f(\sigma,\sigma_n) = \int \exp\left[-\frac{1}{2}\frac{\sigma^2+\sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu - \frac{\sigma_n^2 x + \sigma^2\mu_n}{\sigma^2+\sigma_n^2}\right)^2\right] du$

$p(x|D)$是正态分布，均值为μ_n，方差为$\sigma^2+\sigma_n^2$

$$\boxed{\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0 \qquad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}}$$

贝叶斯推断

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)} \propto P(E|H) \cdot P(H)$$

- **H**ypothesis：需要推断的东西；
- **E**vidence：看到的、已知的事情；
- $P(H|E)$：后验概率，看到E之后对H的推断；
- $P(H)$：先验概率，看到E之前对H的推断；
- $P(E|H)$：似然，如果确实是H，E发生的概率；
- $P(E)$：边缘概率，无论如何E都要发生的概率；
  - 与对H的推断无关；

贝叶斯分类

$$y^* = \arg\max_y P(y|\mathbf{x}) = \frac{p(\mathbf{x}|y) \cdot P(y)}{p(\mathbf{x})} \propto p(\mathbf{x}|y) \cdot P(y)$$

- $y \in \{1, \cdots, c\}$：类别属性；
- $\mathbf{x} \in R^d$：待识别模式的特征；
- $P(y)$：先验概率，类别y发生的概率；
- $p(\mathbf{x}|y)$：似然分布，类别y中的模式具有特征$\mathbf{x}$的概率（密度）；

贝叶斯学习

$$p(\boldsymbol{\theta}|\mathcal{X},\boldsymbol{\alpha}) = \frac{p(\mathcal{X}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta};\boldsymbol{\alpha})}{p(\mathcal{X})} \propto p(\mathcal{X}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta};\boldsymbol{\alpha})$$

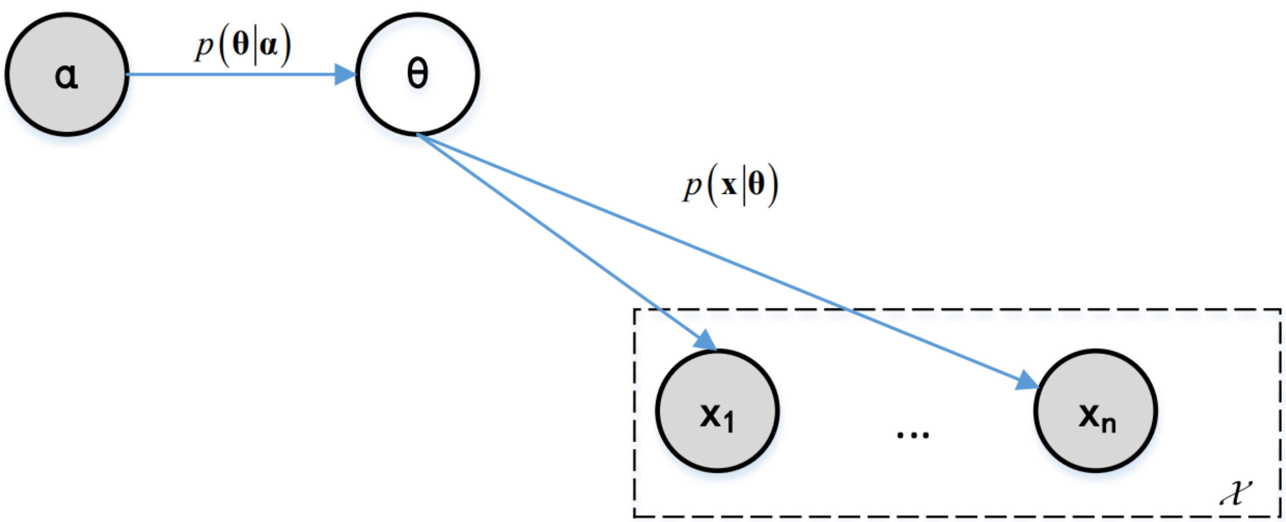
- $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$：样本集，$\mathbf{x}_i \sim p(\mathbf{x};\boldsymbol{\theta})$；
- $\boldsymbol{\theta}$：需要推断的样本集$\mathcal{X}$的抽样分布参数；
- $\boldsymbol{\alpha}$：超参数，先验分布$p(\boldsymbol{\theta})$的参数；

贝叶斯参数估计

- 在概率图模型上，根据可见节点$\boldsymbol{\alpha}$和$\mathcal{X}$推断节点$\boldsymbol{\theta}$的分布：

$$p(\boldsymbol{\theta}, \mathcal{X}, \boldsymbol{\alpha}) = p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\mathcal{X}|\boldsymbol{\theta})$$

$$p(\boldsymbol{\theta}|\mathcal{X}, \boldsymbol{\alpha}) = \frac{p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\mathcal{X}|\boldsymbol{\theta})}{p(\mathcal{X})p(\boldsymbol{\alpha})} \propto p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\mathcal{X}|\boldsymbol{\theta})$$



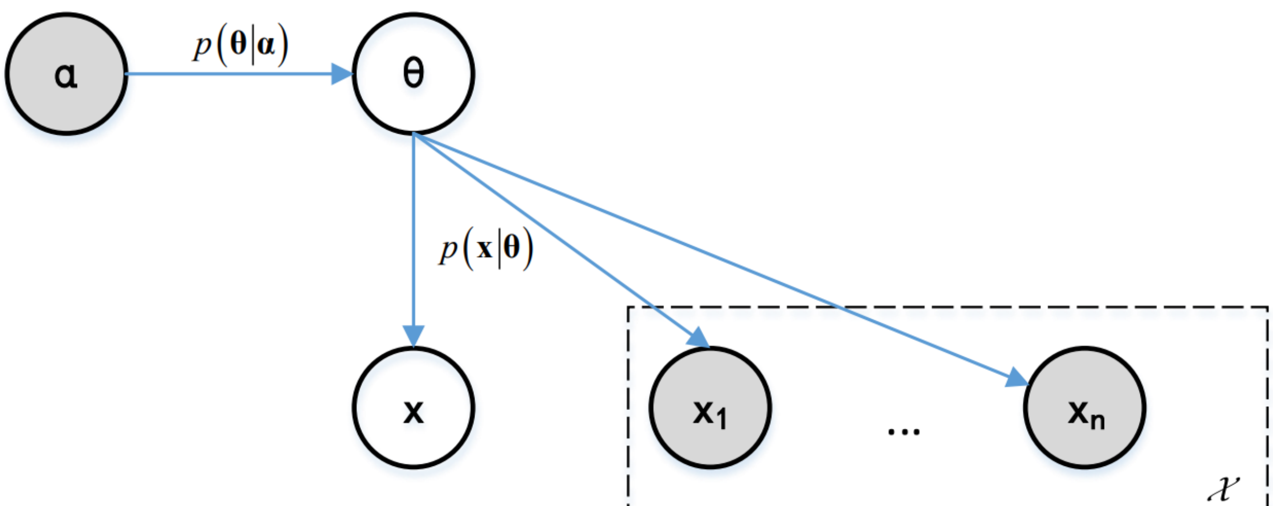
贝叶斯推断

- 在概率图模型上，根据可见节点$\boldsymbol{\alpha}$和$\mathcal{X}$推断节点$\mathbf{x}$的概率：

$$p(\mathbf{x}, \boldsymbol{\theta}, \mathcal{X}, \boldsymbol{\alpha}) = p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\mathcal{X}|\boldsymbol{\theta})$$

$$p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{X}, \boldsymbol{\alpha}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\mathcal{X}|\boldsymbol{\theta})}{p(\mathcal{X})p(\boldsymbol{\alpha})} = p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{X}, \boldsymbol{\alpha})$$

$$p(\mathbf{x}|\mathcal{X}, \boldsymbol{\alpha}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{X}, \boldsymbol{\alpha})d\boldsymbol{\theta}$$



递归贝叶斯学习

$D^n = \left\{ x_1, \cdots x_n \right\}$表示有$n$个样本的样本集

$$p\left(D^n \middle| \boldsymbol{\theta}\right) = p\left(\mathbf{x}_n \middle| \boldsymbol{\theta}\right) p\left(D^{n-1} \middle| \boldsymbol{\theta}\right)$$

$$p\left(\boldsymbol{\theta} \middle| D^n\right) = \frac{p\left(x_n \middle| \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \middle| D^{n-1}\right)}{\int p\left(x_n \middle| \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \middle| D^{n-1}\right) d\boldsymbol{\theta}} \propto p\left(x_n \middle| \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \middle| D^{n-1}\right)$$

随着训练样本的增加，能够产生一系列概率密度函数

$p\left(\boldsymbol{\theta}\right), p\left(\boldsymbol{\theta} \middle| \mathbf{x}_1\right), p\left(\boldsymbol{\theta} \middle| \mathbf{x}_1, \mathbf{x}_2\right), \cdots$

举例：均匀分布的递归贝叶斯学习

一维样本服从均匀分布如下，参数θ值未知

$$p\left(x\middle|\theta\right)\sim U\left(0,\theta\right)=\begin{cases}1/\theta & 0\le x\le\theta\\0 & 其他\end{cases}$$

假设θ服从均匀分布$U\left(0,10\right)$，根据训练样本集$D=\left\{4,7,7,8\right\}$

估计θ及概率密度$p(x)$

$$D^0: 无样本 \quad p\left(\theta\middle|D^0\right)=p\left(\theta\right)=U\left(0,10\right)$$

$$D^1: x_1=4 \quad p\left(\theta\middle|D^1\right)\propto p\left(x\middle|\theta\right)p\left(\theta\middle|D^0\right)=\begin{cases}1/\theta & 4\le\theta\le10\\0 & 其他\end{cases}$$

$$D^2: x_2=7 \quad p\left(\theta\middle|D^2\right)\propto p\left(x\middle|\theta\right)p\left(\theta\middle|D^1\right)=\begin{cases}1/\theta^2 & 7\le\theta\le10\\0 & 其他\end{cases}$$

$$D^1: x_1 = 4 \quad p\left(\theta \middle| D^1\right) \propto p\left(x \middle| \theta\right) p\left(\theta \middle| D^0\right) = \begin{cases} 1/\theta & 4 \le \theta \le 10 \\ 0 & \text{其他} \end{cases}$$

$$D^2: x_2 = 7 \quad p\left(\theta \middle| D^2\right) \propto p\left(x \middle| \theta\right) p\left(\theta \middle| D^1\right) = \begin{cases} 1/\theta^2 & 7 \le \theta \le 10 \\ 0 & \text{其他} \end{cases}$$

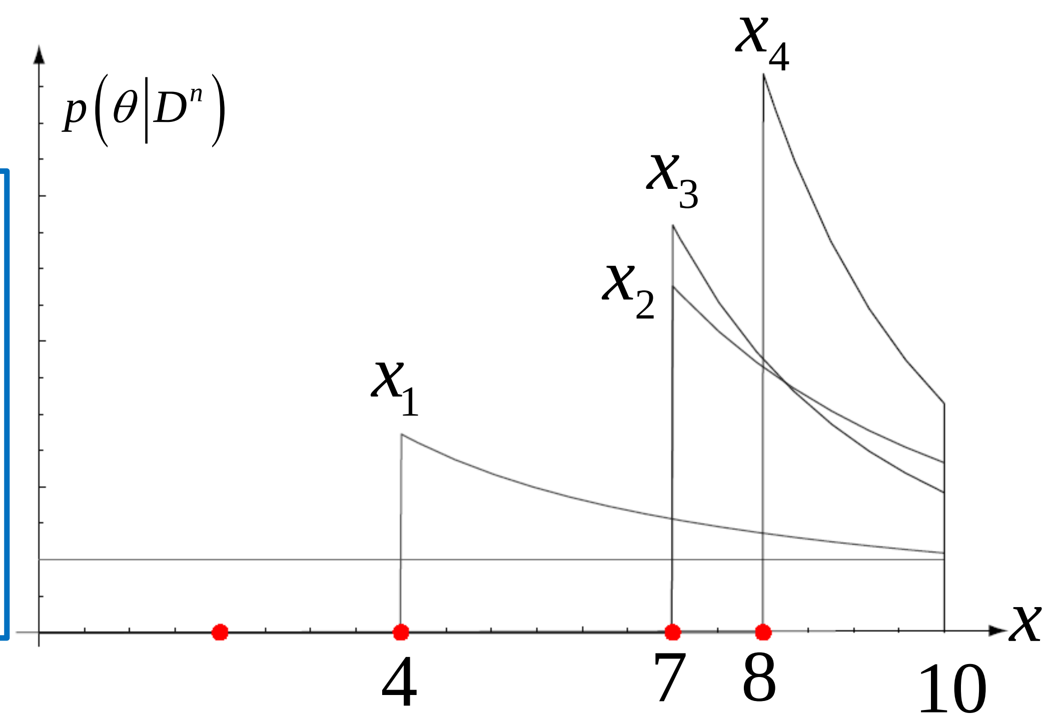
每次递归引入系数$1/\theta$，分布仅对大于最大值的区间非0，即：

$$p\left(\theta \middle| D^n\right) \propto \begin{cases} 1/\theta^n & \max\limits_x \left[D^n\right] \le \theta \le 10 \\ 0 & else \end{cases}$$

训练样本集$D = \{4, 7, 7, 8\}$

⇓

$$p\left(\theta \middle| D^n\right) \propto \begin{cases} 1/\theta^n & 8 \le \theta \le 10 \\ 0 & else \end{cases}$$



$$p(x|\theta)=\begin{cases}1/\theta & 0\le x\le\theta\\ 0 & 其他\end{cases}$$

$$p(\theta|D^n)\propto\begin{cases}1/\theta^n & 8\le\theta\le10\\ 0 & else\end{cases}$$

根据 $p(\theta|D^n)$ ，$p(x|\theta)$ 类条件概率密度的贝叶斯估计：

$$p(\mathbf{x}|D)=\int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|D)d\boldsymbol{\theta}=\int_8^{10}p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|D)d\boldsymbol{\theta}$$

当$0\le x\le 8$时，$p(x|\theta)=1/\theta$

当$\theta<x$时, $p(x|\theta)=0$

当$\theta>x$时, $p(x|\theta)=\dfrac{1}{\theta}$

$$p(\mathbf{x}|D)\propto\int_8^{10}\frac{1}{\theta}\frac{1}{\theta^n}d\boldsymbol{\theta}=\frac{n+1}{8^{n+2}}-\frac{n+1}{10^{n+2}}$$

当$8\le x\le 10$时, $p(x|\theta)=\begin{cases}1/\theta & 0\le x\le\theta\\ 0 & 其他\end{cases}$

$$p(\mathbf{x}|D)\propto\int_8^{x}0\frac{1}{\theta^n}d\boldsymbol{\theta}+\int_x^{10}\frac{1}{\theta}\frac{1}{\theta^n}d\boldsymbol{\theta}=\frac{n+1}{x^{n+2}}-\frac{n+1}{10^{n+2}}$$

当$x<0$或$x>10$时$p(x|\theta)=0$

$$p(\mathbf{x}|D)=0$$

二、不同的x取值，有不同的$p\left(x|\theta\right)$，分两种情况讨论

1、x是0到θ间的均匀分布，

所以当$0\le x\le 8$时，$\theta>8>x$

有$p\left(x|\theta\right)=1/\theta$

$$p\left(\mathbf{x}|D\right)\propto\int_8^{10}\frac{1}{\theta}\frac{1}{\theta^n}d\boldsymbol{\theta}=\frac{n+1}{8^{n+2}}-\frac{n+1}{10^{n+2}}$$

2、当$8<\theta<10$时，要分两种情况：

1）当$\theta<x$时，$p\left(x|\theta\right)=0$

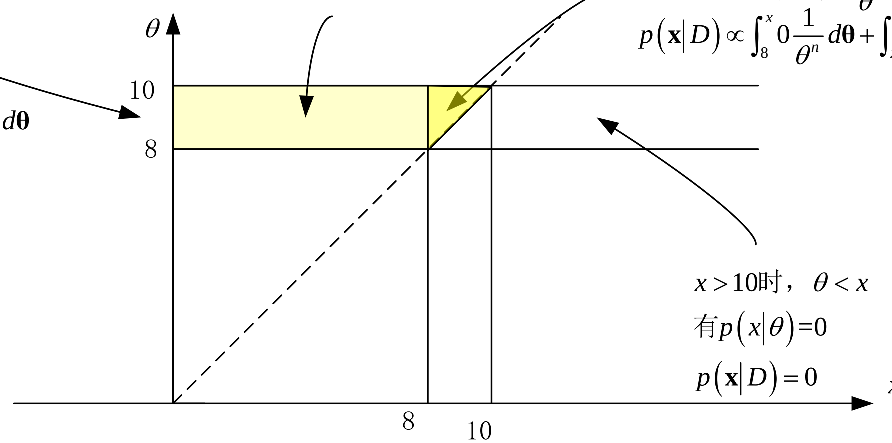
2）当$\theta>x$时，$p\left(x|\theta\right)=\dfrac{1}{\theta}$

$$p\left(\mathbf{x}|D\right)\propto\int_8^x 0\frac{1}{\theta^n}d\boldsymbol{\theta}+\int_x^{10}\frac{1}{\theta}\frac{1}{\theta^n}d\boldsymbol{\theta}=\frac{n+1}{x^{n+2}}-\frac{n+1}{10^{n+2}}$$

一，确定θ的范围

$$p\left(\theta|D^n\right)\propto\begin{cases}1/\theta^n & 8\le\theta\le 10\\ 0 & else\end{cases}$$

$$p\left(\mathbf{x}|D\right)=\int p\left(\mathbf{x}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}|D\right)d\boldsymbol{\theta}=\int_8^{10}p\left(\mathbf{x}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}|D\right)d\boldsymbol{\theta}$$



$x>10$时，$\theta<x$

有$p\left(x|\theta\right)=0$

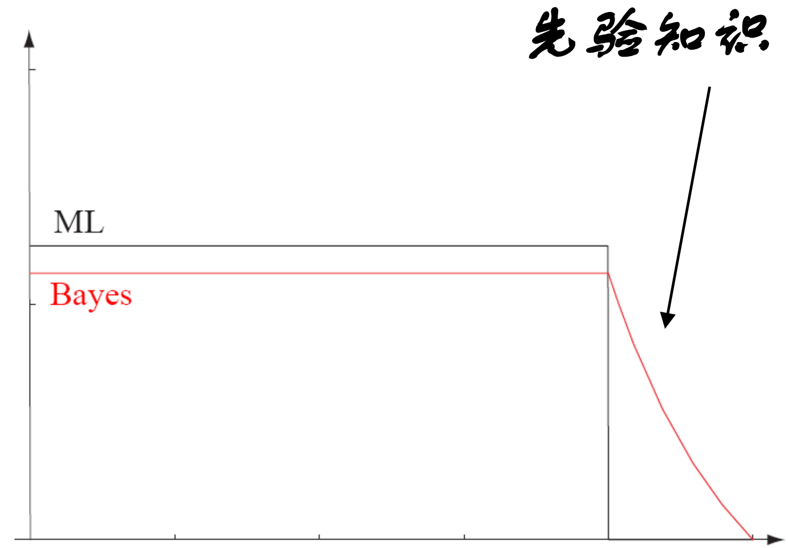
$p\left(\mathbf{x}|D\right)=0$

$$p(x|\theta)=\begin{cases}1/\theta & 0\le x\le\theta\\ 0 & \text{其他}\end{cases}$$

$$p(\theta|D^n)\propto\begin{cases}1/\theta^n & 8\le\theta\le10\\ 0 & else\end{cases}$$

根据 $p(\theta|D^n)$，$p(x|\theta)$ 类条件概率密度的贝叶斯估计：

$$p(\mathbf{x}|D)\propto\begin{cases}\dfrac{n+1}{8^{n+2}}-\dfrac{n+1}{10^{n+2}} & 0\le x\le8\\[2ex] \dfrac{n+1}{x^{n+2}}-\dfrac{n+1}{10^{n+2}} & 8\le x\le10\end{cases}$$



$p(\mathbf{x}|D)$的最大似然估计：

$$\text{似然函数}\, p(D|\boldsymbol{\theta})=\prod_{i=1}^{n}p(x_i|\boldsymbol{\theta})=\frac{1}{\theta^n}$$

$$\max:\frac{1}{\theta^n}\qquad\Rightarrow\qquad\theta=\max\left[D^n\right]$$

$$\text{s.t.:}\quad\theta\ge\max\left[D^n\right]$$

隐Markov模型 (Hidden Markov Model, HMM)

- 有一些模式识别系统处理的是与时间相关的问题，如语音识别，手势识别，唇读系统等；

- 对这类问题采用一个特征矢量序列描述比较方便，这类问题的识别HMM取得了很好的效果。

观察序列

- 信号的特征需要用一个特征矢量的序列来表示：

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

- 其中的v_i为一个特征矢量，称为一个观察值。

ISBN 978-7-5603-4763-9

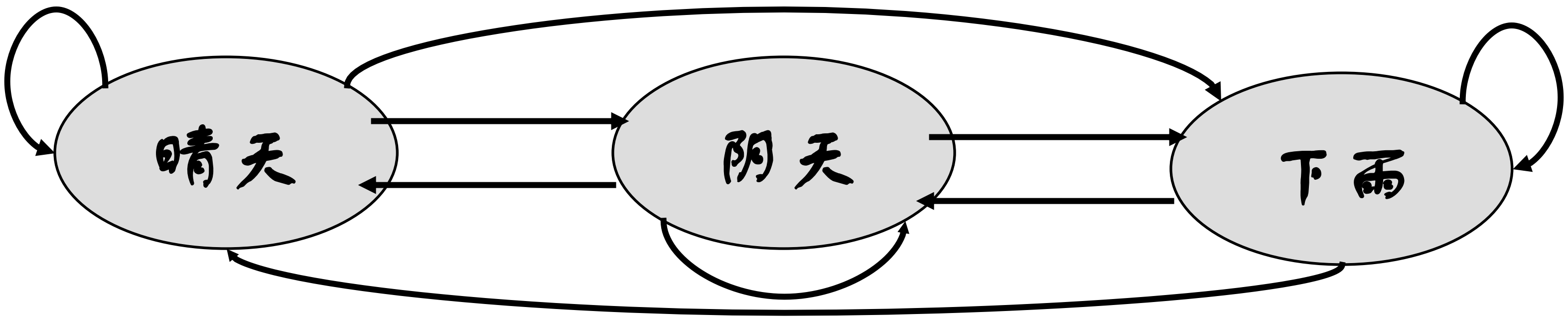
马尔可夫性

- 1870年，俄国有机化学家Vladimir V. Markovnikov第一次提出马尔科夫模型
- 如果一个过程的“将来”仅依赖“现在”而不依赖“过去”，则此过程具有马尔可夫性,或称此过程为马尔可夫过程

$$X(t+1) = f(X(t))$$

ISBN 978-7-5603-4763-9

转移概率



	晴天	阴天	下雨
晴天	0.50	0.25	0.25
阴天	0.375	0.25	0.375
下雨	0.25	0.125	0.625

ISBN 978-7-5603-4763-9

马尔科夫链

- 时间和状态都离散的马尔科夫过程称为马尔科夫链

  记作$\{X_n = X(n), n = 0,1,2,\ldots\}$

  - 在时间集$T_1 = \{0,1,2,\ldots\}$上对离散状态的过程相继观察的结果

- 链的状态空间记做$I = \{a_1, a_2,\ldots\}, a_i \in R$.

- 条件概率$P_{ij}(m, m+n)=P\{X_{m+n} = a_j | X_m = a_i\}$为马氏链在时刻m处于状态$a_i$条件下，在时刻m+n转移到状态$a_j$的转移概率。

ISBN 978-7-5603-4763-9

转移概率矩阵(续)

- 由于链在时刻m从任何一个状态a_i出发，到另一时刻m+n，必然转移到a_1，a_2…，诸状态中的某一个，所以有

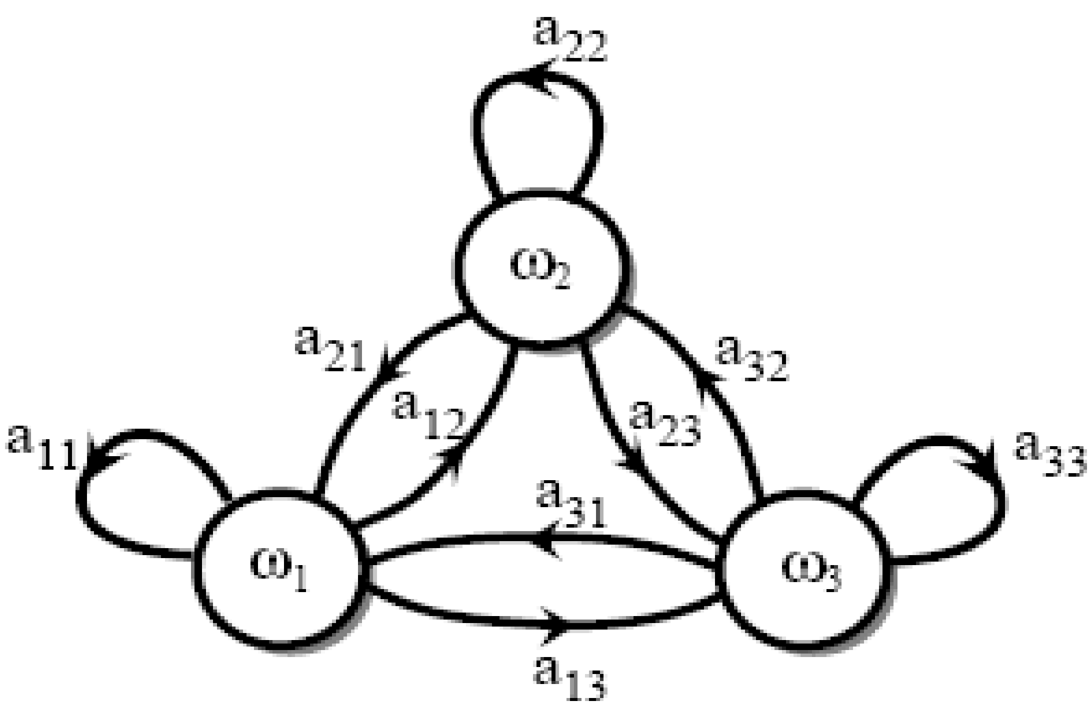
$$\sum_{j=1}^{\infty} P_{ij}(m, m+n) = 1, i = 1, 2, \ldots$$

- 当$P_{ij}(m,m+n)$与m无关时，称马尔科夫链为齐次马尔科夫链，通常说的马尔科夫链都是指齐次马尔科夫链。

ISBN 978-7-5603-4763-9

一阶Markov模型

- 一阶Markov模型由M个状态构成，在每个时刻t，模型处于某个状态w(t)，经过T个时刻，产生出一个长度为T的状态序列W^T=w(1),…,w(T)。



ISBN 978-7-5603-4763-9

一阶Markov模型的状态转移

- 模型在时刻 t 处于状态 w_j 的概率完全由 t-1时刻的状态 w_i 决定，而且与时刻 t 无关，即：

$$P\left(w(t)\middle|W^T\right)=P\left(w(t)\middle|w(t-1)\right)$$

$$P\left(w(t)=\omega_j\middle|w(t-1)=\omega_i\right)=a_{ij}$$

Markov模型的初始状态概率

- 模型初始于状态w_i的概率用 π_i 表示。
- 完整的一阶Markov模型可以用参数 $\boldsymbol{\theta} = (\boldsymbol{\pi}, \mathbf{A})$ 表示，其中：

$$\boldsymbol{\pi} = (\pi_1, \cdots, \pi_M)$$

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1M} \\ a_{21} & a_{22} & \cdots & a_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MM} \end{bmatrix}$$

一阶Markov模型输出状态序列的概率

- 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：$W^5=w_1, w_1, w_3, w_1, w_2$，则模型输出该序列的概率为：

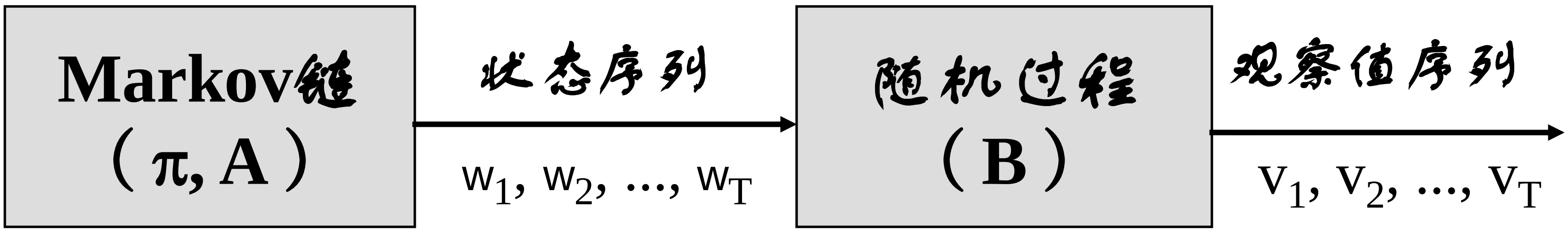
$$P\left(W^5\right)=\pi_1 a_{11} a_{13} a_{31} a_{12}$$

HMM概念

- 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：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间的关系，用观察值概率描述。

ISBN 978-7-5603-4763-9

HMM组成



HMM的组成示意图

ISBN 978-7-5603-4763-9

一阶隐含Markov模型

- 隐含Markov模型中，状态是不可见的，在每一个时刻t，模型当前的隐状态可以输出一个观察值。

- 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。

ISBN 978-7-5603-4763-9

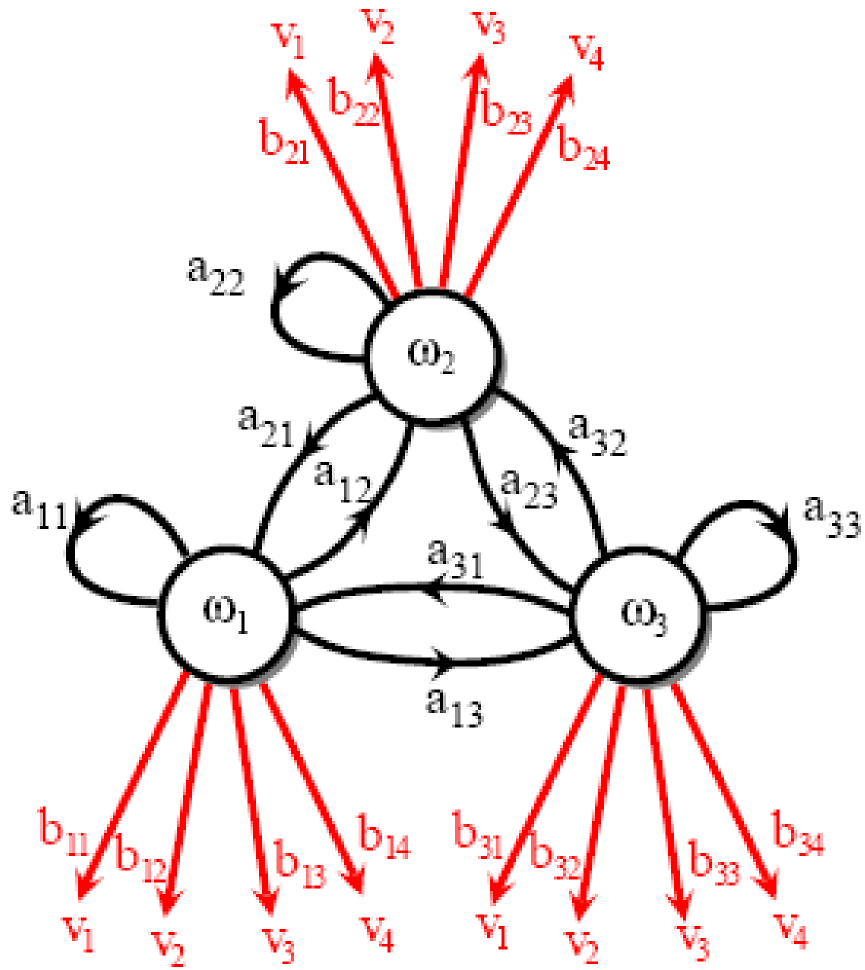
HMM的工作原理

- HMM的内部状态转移过程同Markov模型相同，在每次状态转移之后，由该状态输出一个观察值，只是状态转移过程无法观察到，只能观察到输出的观察值序列。

- 以离散的HMM为例，隐状态可能输出的观察值集合为$\{v_1, v_2, \ldots, v_K\}$，第i个隐状态输出第k个观察值的概率为$b_{ik}$。

- 例如：T=5时，可能的观察序列$V^5=v_3v_2v_3v_4v_1$

HMM的工作过程



ISBN 978-7-5603-4763-9

HMM假设

对于一个随机事件，有一个观察值序列：$v_1,\ldots,v_T$

该事件隐含着一个状态序列：$w_1,\ldots,w_T$

假设1：马尔可夫假设（状态构成一阶马尔可夫链）

$$p(w_i \mid w_{i-1}\ldots w_1) = p(w_i \mid w_{i-1})$$

假设2：不动性假设（状态与具体时间无关）

$$p(w_i \mid w_{i-1}) = p(w_j \mid w_{j-1}) \qquad \text{对任意}i, j\text{成立}$$

假设3：输出独立性假设（输出仅与当前状态有关）

$$p(v_1,\cdots,v_T \mid w_1,\cdots,w_T) = \prod_{t=1} p(v_t \mid w_t)$$

HMM的参数表示

$$\boldsymbol{\theta} = \left(\boldsymbol{\pi}, \mathbf{A}, \mathbf{B}\right)$$

M个状态，K个可能的输出值。

初始概率：$\boldsymbol{\pi}$，包括M个元素。

$$\boldsymbol{\pi} = \left(\pi_1, \cdots, \pi_M\right)^T \qquad \pi_i: \text{第一个时刻处于状态 } w_i \text{ 的概率}$$

状态转移矩阵：$\mathbf{A}$，M*M方阵；

$$\mathbf{A} = \left(a_{ij}\right)_{M\times M} \qquad a_{ij} = P\left(w(t) = w_j \middle| w(t-1) = w_i\right)$$

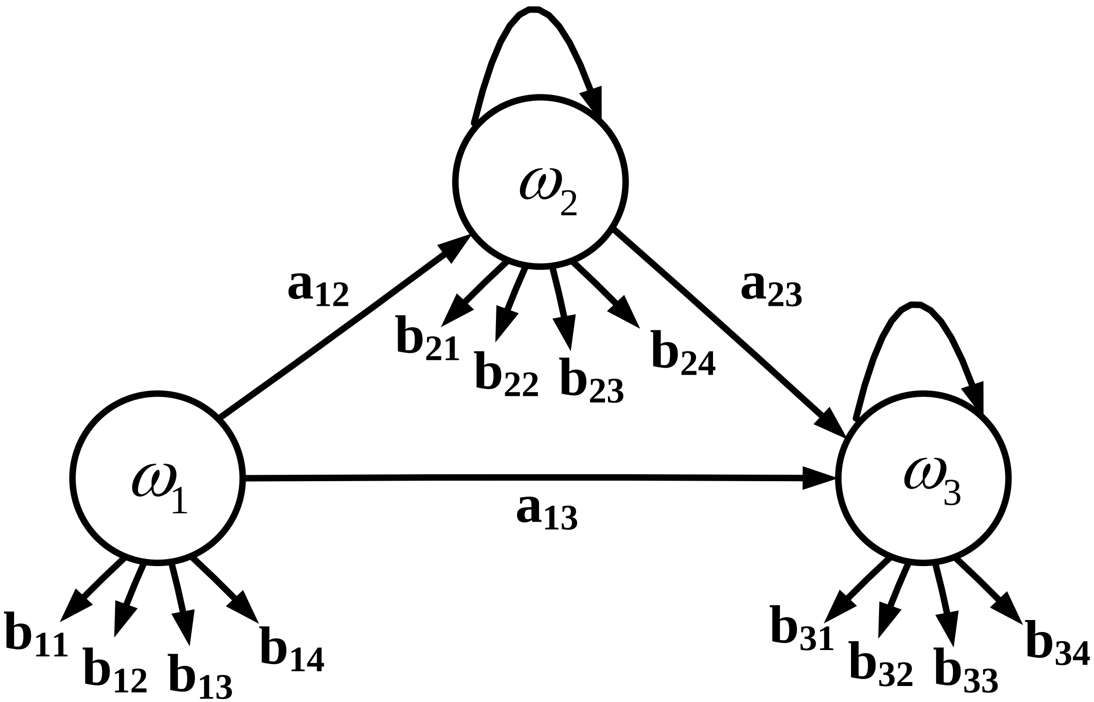
状态输出概率：B，M*K矩阵；

$$\mathbf{B} = \left(b_{ij}\right)_{M\times K} \qquad b_{ij} = P(v_j | w_i)$$

ISBN 978-7-5603-4763-9

HMM示例

如图HMM模型，初始概率：$\pi_1 = 1\ \pi_2 = 0\ \ \pi_3 = 0$ 状态转移概率矩阵：



$$\mathbf{A} = \begin{bmatrix} 0 & 0.3 & 0.7 \\ 0 & 0.6 & 0.4 \\ 0 & 0 & 1 \end{bmatrix}$$

状态输出概率矩阵：

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

1，请列出所有可能输出序列 $V = v_2 v_4 v_4 v_1$ 的状态转移序列。

2，分别计算由每一个状态转移序列输出观察序列V的概率

3，计算最有可能输出观察序列V的状态转移序列

ISBN 978-7-5603-4763-9

HMM的三个核心问题

- **估值问题：已有一个HMM模型，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模型，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模型的结构，其参数未知，根据一组训练序列对参数进行训练；

估值问题

HMM模型产生观察序列 V^T 可以由下式计算：

$$P\left(V^T \middle| \boldsymbol{\theta}\right) = \sum_{r=1}^{r_{\max}} P\left(V^T \middle| W_r^T\right) P\left(W_r^T \middle| \boldsymbol{\theta}\right)$$

- $r_{\max}=M^T$ 为HMM所有可能的状态转移序列数；
- $P\left(V^T \middle| W_r^T\right)$ 为状态转移序列 W_r^T 输出观察序列 V^T 的概率；
- $P\left(W_r^T \middle| \boldsymbol{\theta}\right)$ 为状态转移序列 W_r^T 发生的概率。

□计算复杂度：$O\left(M^T \times T\right)$

ISBN 978-7-5603-4763-9

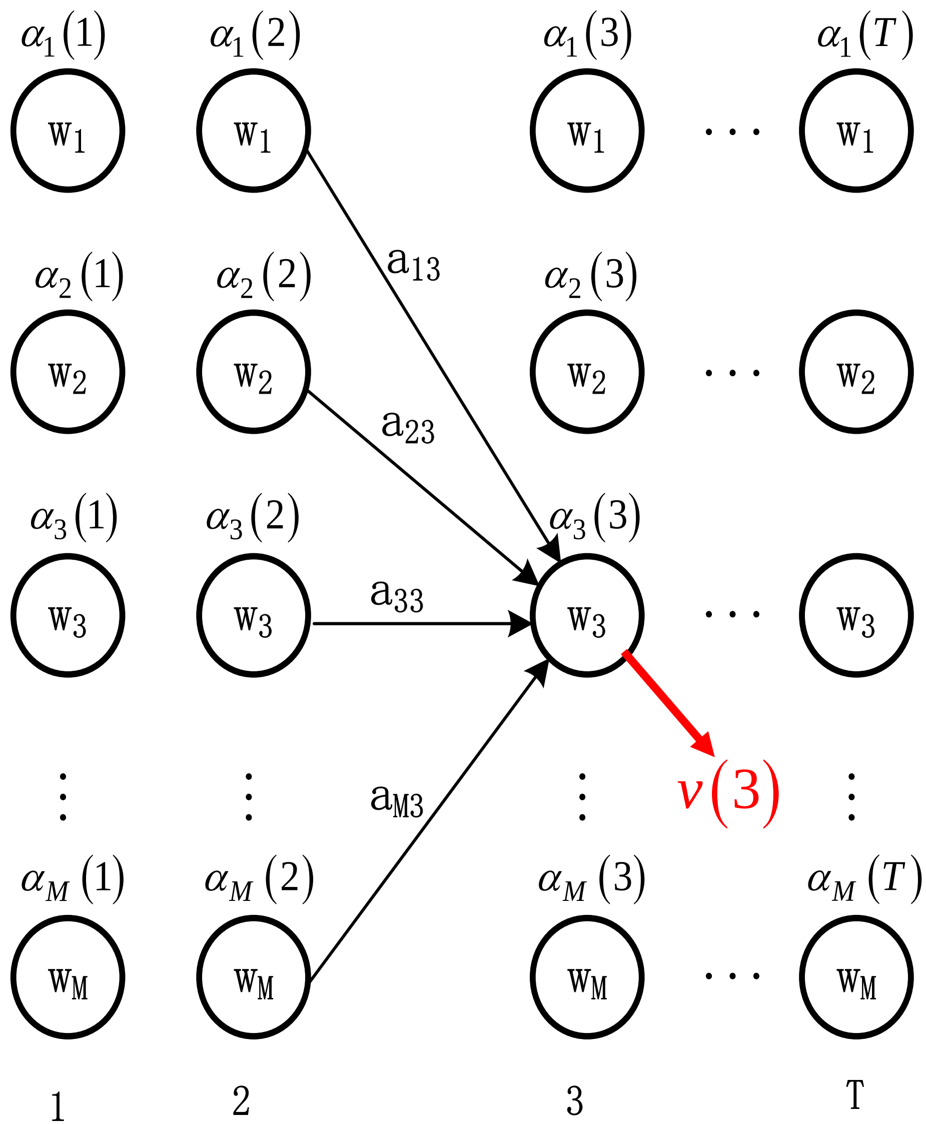
HMM前向算法

$\alpha_i(t)$：t 时刻，位于隐状态 ω_i 产生前 t 个可见符号的概率

$$\alpha_i(1)=\pi_i b_i\left[v(1)\right]$$

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

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



下标 i 表示隐状态

$b_{iv(t)}$：隐状态为i，输出为$v(t)$的概率

ISBN 978-7-5603-4763-9

HMM的前向算法

前向算法

- 初始化 $t=1$；
- 计算第1列每个节点的 α 值：$\alpha_i(1)=\pi_i b_i\left[v(1)\right]$ ；
- 迭代计算t=2至 T 列每个节点的 α 值：

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

- 输出：

$$P\left(V^T\middle|\boldsymbol{\theta}\right)=\sum_{i=1}^{M}\alpha_i(T)$$

计算复杂度：$O\left(M^2T\right)$

哈尔滨工业大学 计算机学院
模式识别与智能系统研究中心
ISBN 978-7-5603-4763-9

HMM的后向算法

$\beta_i(t)$：t 时刻，位于隐状态 ω_i 产生V^T的t时刻值后的T-t个可见符号的概率

初始化：$\beta_i(T)=1, i=1,\cdots M$

迭代计算：

$$\beta_i(t)=\left[\sum_{j=1}^{M}\beta_j(t+1)a_{ji}\right]b_j\left(v(t+1)\right), i=1,\cdots,M$$

结束输出：

$$P\left(V^T\middle|\boldsymbol{\theta}\right)=\sum_{i=1}^{M}\beta_i(T)$$

计算复杂度：$O\left(M^2T\right)$

ISBN 978-7-5603-4763-9

前向算法与后向算法的结合

将V^T的在T′时刻分为V1，V2两个序列，有：

$$P\left(V^T \middle| \boldsymbol{\theta}\right) = \sum_{i=1}^{M} \alpha_i\left(T'\right)\beta_i\left(T'\right)$$

$$P\left(V^T \middle| \boldsymbol{\theta}\right) = P\left(V_1, V_2 \middle| \boldsymbol{\theta}\right) = \sum_{i=1}^{M} P\left(V_1, V_2, \omega(T') = i \middle| \boldsymbol{\theta}\right)$$

$$= \sum_{i=1}^{M} P\left(V_1, \omega(T') = i \middle| \boldsymbol{\theta}\right) P\left(V_2, \omega(T') = i \middle| \boldsymbol{\theta}\right)$$

$$= \sum_{i=1}^{M} \alpha_i\left(T'\right)\beta_i\left(T'\right)$$

HMM的三个核心问题

- 估值问题：已有一个HMM模型，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模型，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模型的结构，其参数未知，根据一组训练序列对参数进行训练；

哈尔滨工业大学 计算机学院
模式识别与智能系统研究中心
ISBN 978-7-5603-4763-9

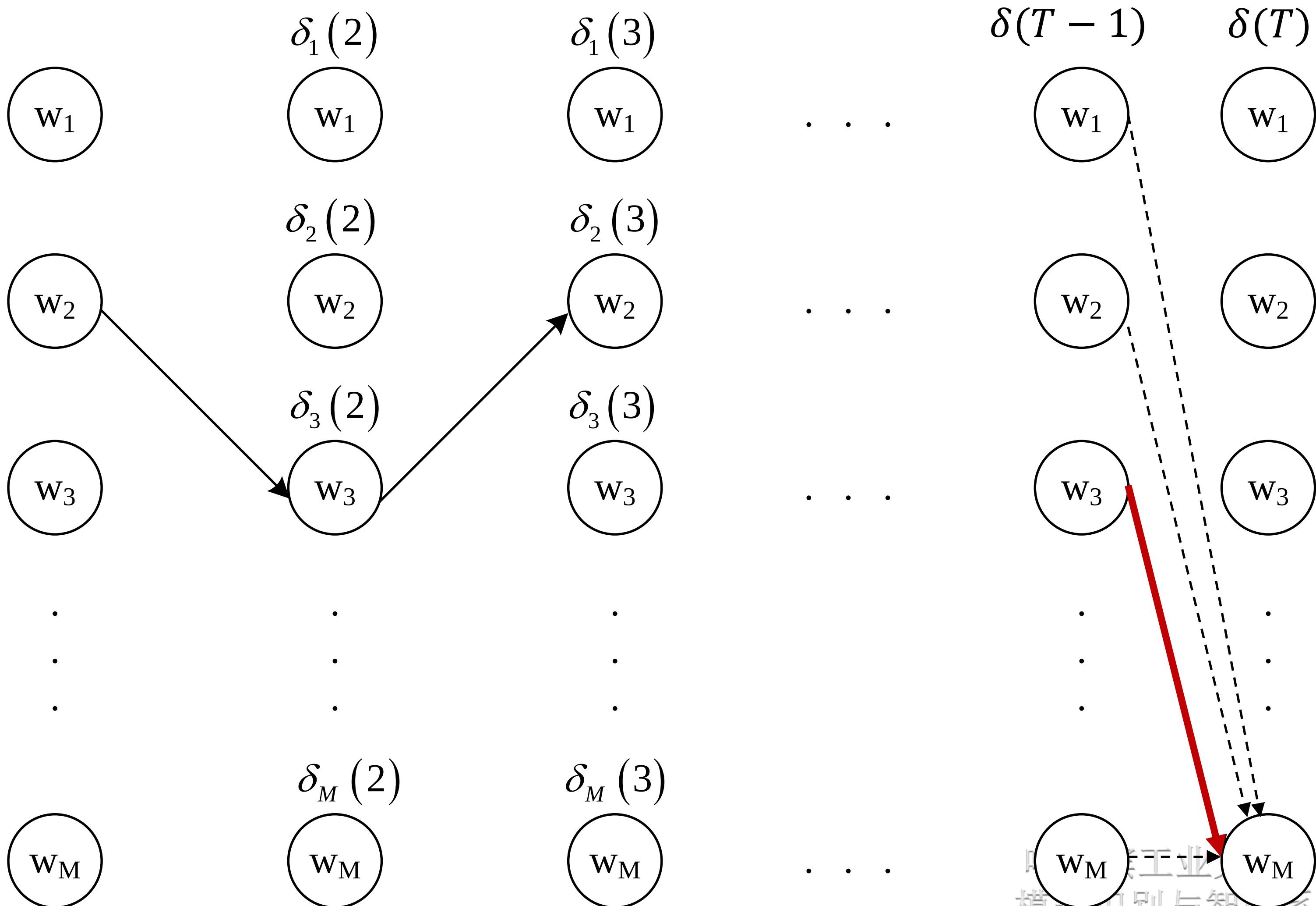
解码问题

- 解码问题的计算同估值问题的计算类似，最直观的思路是遍历所有的可能状态转移序列，取出最大值，计算复杂度为：$O(M^T T)$。

- 同样存在着优化算法：Viterbi算法。

Viterbi算法图示

有向无环图，动态规划



Viterbi算法

$\delta_i(t)$：t 时刻，位于隐状态 ω_i 产生V^T的前t个可见符号的最大概率

$$\delta_i(1)=\pi_i b_i(v(1))$$

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

建立一个矩阵Φ，其元素 $\varphi_i(t)$保存（第t步为第i个状态时）在第t-1步的最优状态。

$$\varphi_i(t+1)=\arg\max_{1\le j\le M}\left[\delta_j(t)a_{ji}\right]$$

ISBN 978-7-5603-4763-9

Viterbi算法

1. 初始化： $\delta_i(1)=\pi_i b_i\left(v(1)\right), i=1,\cdots M,\quad \phi_1\left(i\right)=0$

2. 迭代计算：

$$\phi_i\left(t+1\right)=\arg\max_{1\le j\le M}\left[\delta_j\left(t\right)a_{ji}\right]$$

$$\delta_i\left(t+1\right)=\max_{1\le j\le M}\left[\delta_j\left(t\right)a_{ji}\right]b_i\left(v\left(t+1\right)\right), i=1,\cdots,M,$$

3. 结束：

$$P^*\left(V^T\middle|\boldsymbol{\theta}\right)=\max_{1\le j\le M}\left[\delta_j\left(T\right)\right],\; w^*\left(T\right)=\arg\max_{1\le j\le M}\left[\delta_j\left(T\right)\right]$$

4. 路径回朔： $w^*\left(t\right)=\phi_{w^*(t+1)}\left(t+1\right)$

ISBN 978-7-5603-4763-9

Viterbi算法

$$\delta_i(t+1) = \max_{1\le j\le M}\left[\delta_j(t)a_{ji}\right]b_i\left(v(t+1)\right), i = 1,\cdots,M,$$

$$\ln\delta_i(t+1) = \max_{1\le j\le M}\ln\left\{\left[\delta_j(t)a_{ji}\right]b_i\left(v(t+1)\right)\right\}$$

$$= \max_{1\le j\le M}\left\{\underbrace{\ln\delta_j(t)}_{=\rho_j(t)} + \ln a_{ji} + \ln b_i\left(v(t+1)\right)\right\}$$

$$\rho(t+1) = \max_{1\le j\le M}\left\{\rho_j(t) + \ln a_{ji} + \ln b_i\left(v(t+1)\right)\right\}$$

计算复杂度：$O\left(M^2T\right)$

ISBN 978-7-5603-4763-9

作业：HMM应用举例

假设你有一个朋友在外地，每天做三种活动之一——Walk, Shop, Clean。从事活动的概率与天晴、下雨有关，天气与运动的关系及天气间转换的关系如下表：

	Rainy	Sunny
Walk	0.1	0.6
Shop	0.4	0.3
Clean	0.5	0.1

下雨散步的可能性是0.1。

	Rainy	Sunny
Rainy	0.7	0.3
Sunny	0.4	0.6

从行到列：从今天是晴天而明天就开始下雨的可能性是0.4 。

第一天的天气有0.6的概率是Rainy，有0.4概率是Sunny。如果连续三天，你发现你的朋友的活动是：Walk->Shop->Clean；那么，如何判断你朋友那里这几天的天气是怎样的?

ISBN 978-7-5603-4763-9

HMM的三个核心问题

- 估值问题：已有一个HMM模型，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模型，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模型的结构，其参数未知，根据一组训练序列对参数进行训练；

ISBN 978-7-5603-4763-9

HMM的学习问题

已知一组观察序列(训练样本集合)：

$$V=\left\{V_1^{T_1},V_2^{T_2},\cdots,V_n^{T_n}\right\}$$

如何确定最优的模型参数θ，使得模型产生训练集合V的联合概率最大

$$\hat{\boldsymbol{\theta}}=\max_{\boldsymbol{\theta}} P\left(V\middle|\boldsymbol{\theta}\right)$$

这同样是一个最大似然估计问题，需要采用EM算法。

ISBN 978-7-5603-4763-9

HMM学习问题

Baum-Welch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。

——广义EM算法在HMM中的具体实现

$$a_{ij}=\frac{\text{从状态}i\text{跳转到状态}j\text{的概率}}{\text{从状态}i\text{跳出的概率}}\qquad b_{ik}=\frac{\text{从状态}i\text{输出观测}k\text{的概率}}{\text{跳转到状态}i\text{的概率}}$$

设t时刻，从 i 跳转到 j 的概率为$\gamma_{ij}(t)$，观测长度为T的训练序列，对t求和：

如何计算?

$$a_{ij}=\frac{\sum_{t=2}^{T}\gamma_{ij}(t)}{\sum_{t=2}^{T}\sum_{k=1}^{M}\gamma_{ik}(t)}\qquad b_{ik}=\frac{\sum_{t=1,v(t)=v_k}^{T}\sum_{l=1}^{M}\gamma_{li}(t)}{\sum_{t=1}^{T}\sum_{l=1}^{M}\gamma_{li}(t)}$$

ISBN 978-7-5603-4763-9

t 时刻，从状态 i 跳转到 j 的概率 $\gamma_{ij}(t)$

模型参数为θ，观测到序列V^T的条件下，t-1 时刻处于ω_i，t 时刻处于ω_j的概率

$$\gamma_{ij}(t)=P\left[\omega(t-1)=\omega_i,\omega(t)=\omega_j \middle| V^T,\boldsymbol{\theta}\right]$$

$$=\frac{P\left[\omega(t-1)=\omega_i,\omega(t)=\omega_j,V^T \middle| \boldsymbol{\theta}\right]}{P\left(V^T \middle| \boldsymbol{\theta}\right)}=\frac{\alpha_i(t-1)a_{ij}b_{jv(t)}\beta_j(t)}{P\left(V^T \middle| \boldsymbol{\theta}\right)}$$

三个独立子事件:

t-1时刻处于状态 ω_i，从1到t-1产生序列 $V^{1\to t-1}$,概率为 $\alpha_i(t-1)$

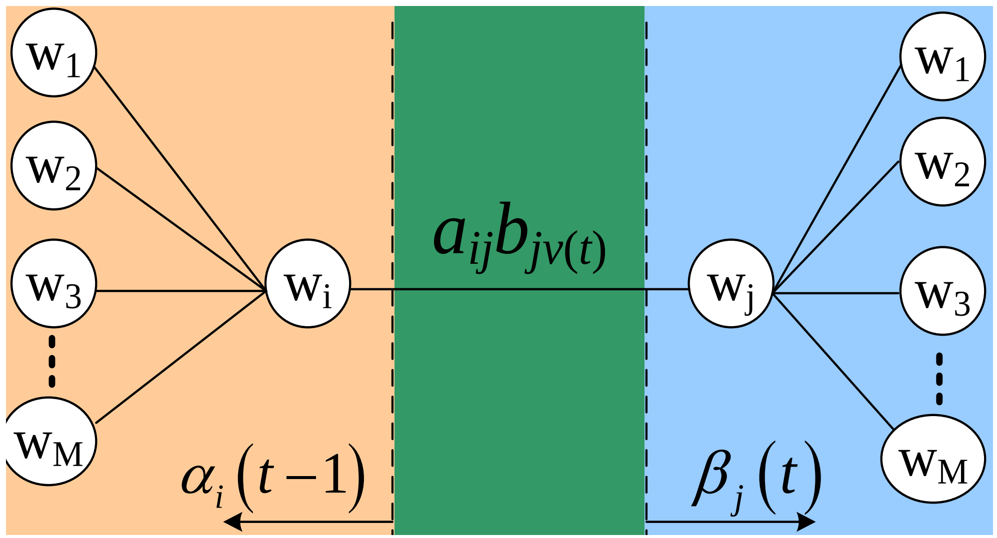
t 时刻处于状态ω_j，从$t+1$ 到 **T** 产生序列$V^{t+1\to T}$ ，概率为$\beta_j(t)$

在t时刻由状态ω_i转移到ω_j，概率为$a_{ij}b_{jv(t)}$

ISBN 978-7-5603-4763-9

参数为θ,观测到序列V^T的条件下,$t-1$ 时刻处于ω_i,t 时刻处于ω_j 的概率

$$\gamma_{ij}(t)=\frac{P\left[\omega(t-1)=\omega_i,\omega(t)=\omega_j,V^T\middle|\boldsymbol{\theta}\right]}{P\left(V^T\middle|\boldsymbol{\theta}\right)}=\frac{\alpha_i(t-1)a_{ij}b_{jv(t)}\beta_j(t)}{P\left(V^T\middle|\boldsymbol{\theta}\right)}$$



$\alpha_i(t-1)$：在 t-1时刻处于状态 ω_i，从1到t-1时刻之间产生序列的概率；

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

$\beta_j(t)$：在 t 时刻处于状态ω_j，从t+1 到 T 时刻之间产生序列的概率；

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

状态转移概率的估计

输出观察序列V^T时，在t-1时刻HMM处于ω_i状态，在时刻t处于ω_j状态的概率：

$$\gamma_{ij}(t)=\frac{\alpha_i(t-1)a_{ij}b_j(v(t))\beta_j(t)}{P(V^T|\boldsymbol{\theta})}$$

ω_i到ω_j的预期数： $\sum_{t=1}^{T}\gamma_{ij}(t)$

从 ω_i的任何转移的总预期数： $\sum_{t=1}^{T}\sum_{k}\gamma_{ik}(t)$

从ω_i到ω_j的 转移的概率估计： $\hat{a}_{ij}=\frac{\sum_{t=1}^{T}\gamma_{ij}(t)}{\sum_{t=1}^{T}\sum_{k}\gamma_{ik}(t)}$

ISBN 978-7-5603-4763-9

初始概率的估计

输出观察序列V^T时，在t=1时刻HMM处于ω_j状态的概率：

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

π_i 的迭代公式为：

$$\pi_i = P\left[w(1) = w_i \middle| V^T, \boldsymbol{\theta}\right] = \gamma_i(1)$$

ISBN 978-7-5603-4763-9

观察概率的估计

输出观察序列V^T时，在t-1时刻HMM处于ω_i状态，在时刻t处于ω_j状态的概率：

$$\gamma_{ij}(t)=\frac{\alpha_i(t-1)a_{ij}b_j(v(t))\beta_j(t)}{P(V^T|\boldsymbol{\theta})}$$

ω_i 上观察到 v_k 的预期数： $\sum\limits_{t=1,v(t)=v_k}^{T}\sum\limits_{l=1}^{M}\gamma_{li}(t)$

到 ω_i的任何转移的总预期数： $\sum\limits_{t=1}^{T}\sum\limits_{l=1}^{M}\gamma_{li}(t)$

从 ω_i 到b_k的 观测概率估计：

$$\hat{b}_{ik}=\frac{\sum\limits_{t=1,v(t)=v_k}^{T}\sum\limits_{l=1}^{M}\gamma_{li}(t)}{\sum\limits_{t=1}^{T}\sum\limits_{l=1}^{M}\gamma_{li}(t)}$$

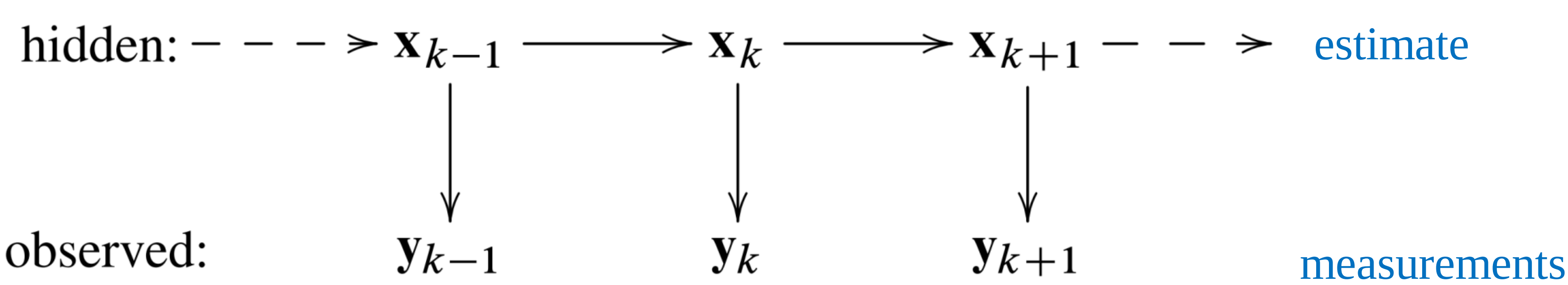
估计问题建模

状态空间模型

$$\begin{cases} x_k = f\left(x_{k-1}, u_k\right) + w_k \\ y_k = h\left(x_k\right) + v_k \end{cases}$$

概率状态空间模型

$$\begin{cases} x_k \sim p\left(x_k \middle| x_{0:k-1}, y_{1:k-1}\right) \\ y_k \sim p\left(y_k \middle| x_{0:k}, y_{1:k-1}\right) \end{cases}$$



估计问题建模

$$\begin{tikzcd}
\text{hidden:} \arrow[r, dashed] & \mathbf{x}_{k-1} \arrow[r] \arrow[d] & \mathbf{x}_k \arrow[r] \arrow[d] & \mathbf{x}_{k+1} \arrow[r, dashed] \arrow[d] & {} \\
\text{observed:} & \mathbf{y}_{k-1} & \mathbf{y}_k & \mathbf{y}_{k+1} &
\end{tikzcd}$$

概率模型简化

$$\begin{cases} x_k \sim p\left(x_k \middle| x_{0:k-1}, y_{1:k-1}\right) \\ y_k \sim p\left(y_k \middle| x_{0:k}, y_{1:k-1}\right) \end{cases} \xRightarrow{\text{HMM}} \begin{cases} x_k \sim p\left(x_k \middle| x_{k-1}\right) \\ y_k \sim p\left(y_k \middle| x_k\right) \end{cases}$$

全状态 $x_{0:k}$ 的后验分布

$$p\left(x_{0:k} \middle| y_{1:k}\right) = \frac{p\left(y_{1:k} \middle| x_{0:k}\right) p\left(x_{0:k}\right)}{p\left(y_{1:k}\right)} \propto p\left(y_{1:k} \middle| x_{0:k}\right) p\left(x_{0:k}\right)$$

$x_{0:k}$ 的贝叶斯估计： $\underset{x_{0:k}}{\arg\max}\, p\left(y_{1:k} \middle| x_{0:k}\right) p\left(x_{0:k}\right)$

很多实际任务只关心某一时刻的状态 x_k

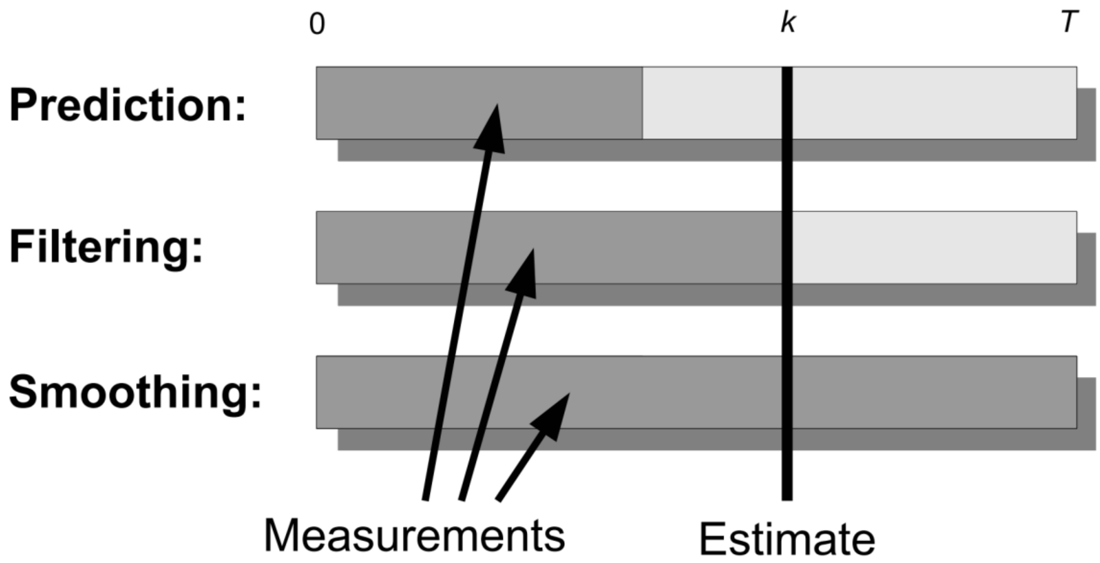
针对k时刻的状态估计

预测（prediction），滤波（filtering）和平滑（smoothing）

$$p\left(x_{k+n} \middle| y_{1:k}\right)$$

$$p\left(x_k \middle| y_{1:k}\right)$$

$$p\left(x_k \middle| y_{1:T}\right)$$



贝叶斯滤波——递归贝叶斯估计

已知K-1时刻状态

$$p\left(x_{k-1} \mid y_{1:k-1}\right)$$

预测K时刻状态(1步转移概率，prior）

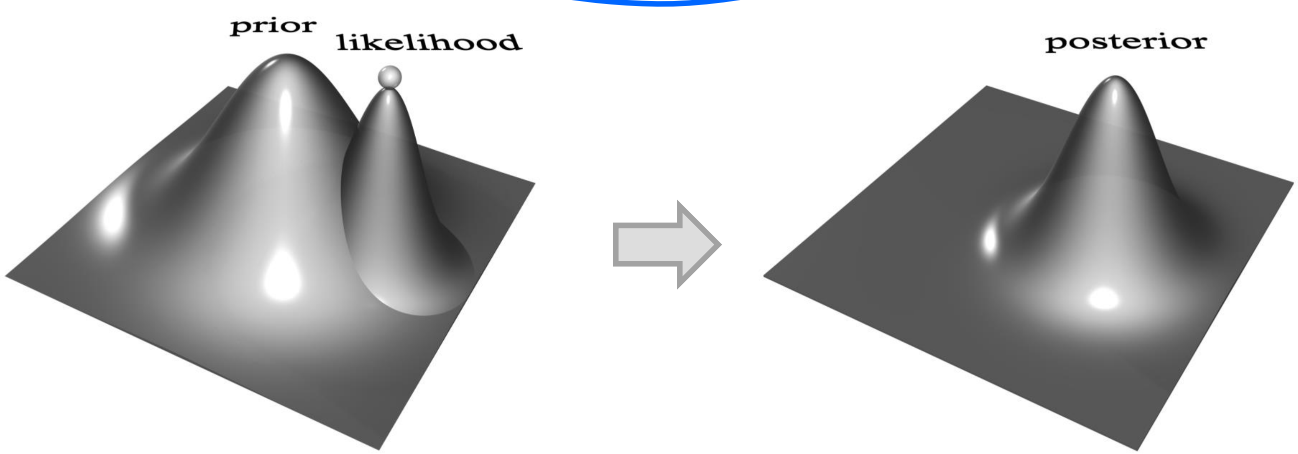
$$p\left(x_k \mid y_{1:k-1}\right) = \int p\left(x_k, x_{k-1} \mid y_{1:k-1}\right) dx_{k-1}$$

获得K时刻测量likelihood

$$p\left(y_k \mid x_k\right)$$

更新k时刻状态估计，posterior

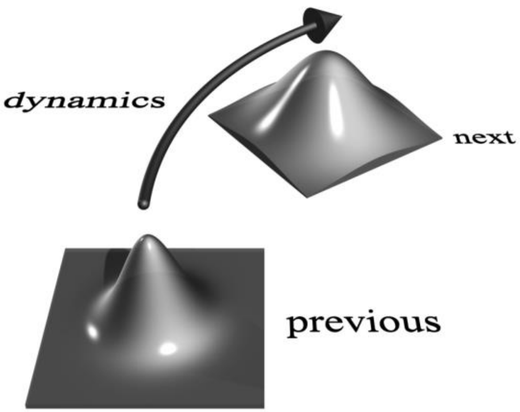
$$p\left(x_k \mid y_{1:k}\right) = \frac{p\left(x_k, y_{1:k}\right)}{p\left(y_{1:k}\right)} = \frac{1}{Z_k} p\left(y_k \mid x_k\right) p\left(x_k \mid y_{1:k-1}\right)$$



ISBN 978-7-5603-4763-9

贝叶斯滤波

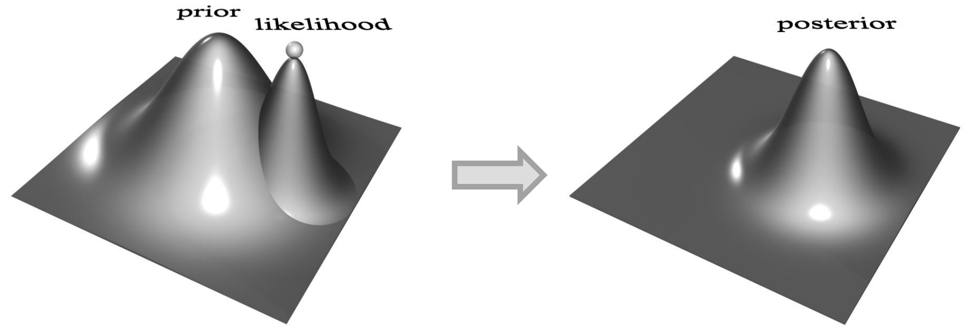
状态预测： $p\left(x_k \middle| y_{1:k\text{-}1}\right)$为联合分布$p\left(x_k, x_{k\text{-}1} \middle| y_{1:k\text{-}1}\right)$对$x_{k\text{-}1}$积分



$$
\begin{aligned}
& p\left(x_k \middle| y_{1:k\text{-}1}\right) \\
&= \int p\left(x_k, x_{k\text{-}1} \middle| y_{1:k\text{-}1}\right) dx_{k-1} \\
&= \int \underline{p\left(x_k \middle| x_{k\text{-}1}, y_{1:k\text{-}1}\right)}\, p\left(x_{k\text{-}1} \middle| y_{1:k\text{-}1}\right) dx_{k-1} \\
&= \int \underline{p\left(x_k \middle| x_{k\text{-}1}\right)}^{\text{HMM}}\, p\left(x_{k-1} \middle| y_{1:k\text{-}1}\right) dx_{k-1}
\end{aligned}
$$

ISBN 978-7-5603-4763-9

贝叶斯滤波



测量更新

$$\underbrace{p\left(x_k \mid y_{1:k}\right)}_{\text{posterior}}=\frac{p\left(x_k, y_{1:k}\right)}{p\left(y_{1:k}\right)}$$

$$=\frac{p\left(y_k \mid x_k\right) p\left(x_k \mid y_{1:k-1}\right) p\left(y_{1:k-1}\right)}{p\left(y_{1:k-1}\right)\int p\left(y_k \mid x_k\right) p\left(x_k \mid y_{1:k-1}\right) dx_k}$$

$$=\frac{1}{\mathrm{Z}_k} \underbrace{p\left(y_k \mid x_k\right)}_{\text{likelihood}} \underbrace{p\left(x_k \mid y_{1:k-1}\right)}_{\text{prior}}$$

分子

$$p\left(x_k, y_{1:k}\right)$$

$$=p\left(y_k \mid x_k, y_{1:k-1}\right) p\left(x_k, y_{1:k-1}\right)$$

HMM

$$=p\left(y_k \mid x_k\right) p\left(x_k \mid y_{1:k-1}\right) p\left(y_{1:k-1}\right)$$

分母

$$p\left(y_{1:k}\right)$$

$$=\int p\left(x_k, y_{1:k}\right) dx_k$$

$$=\int p\left(y_k \mid x_k\right) p\left(x_k \mid y_{1:k-1}\right) p\left(y_{1:k-1}\right) dx_k$$

$$=p\left(y_{1:k-1}\right)\int p\left(y_k \mid x_k\right) p\left(x_k \mid y_{1:k-1}\right) dx_k$$

归一化常数

$$\mathrm{Z}_{\mathrm{k}}=\int p\left(y_k \mid x_k\right) p\left(x_k \mid y_{1:k-1}\right) dx_k$$

贝叶斯滤波——递归贝叶斯估计

已知K-1时刻状态

$$p\left(x_{k-1} \mid y_{1:k-1}\right)$$

预测K时刻状态(1步转移概率，prior)

$$p\left(x_k \mid y_{1:k-1}\right) = \int p\left(x_k, x_{k-1} \mid y_{1:k-1}\right) dx_{k-1}$$

获得K时刻测量likelihood

$$p\left(y_k \mid x_k\right)$$

更新k时刻状态估计，posterior

$$p\left(x_k \mid y_{1:k}\right) = \frac{p\left(x_k, y_{1:k}\right)}{p\left(y_{1:k}\right)} = \frac{1}{Z_k} p\left(y_k \mid x_k\right) p\left(x_k \mid y_{1:k-1}\right)$$

HMM模型，递归贝叶斯，求解posterior

看起来很美，如何确定分布形式？如何积分？

——根据不同的应用环境做不同的假设

卡尔曼滤波、扩展卡尔曼滤波、无迹卡尔曼滤波、粒子滤波…

贝叶斯滤波——递归贝叶斯估计

已知K-1时刻状态

$$p\left(x_{k-1} \mid y_{1:k-1}\right)$$

预测K时刻状态(1步转移概率，prior）

$$p\left(x_k \mid y_{1:k-1}\right) = \int p\left(x_k, x_{k-1} \mid y_{1:k-1}\right) dx_{k-1}$$

获得K时刻测量likelihood

$$p\left(y_k \mid x_k\right)$$

更新k时刻状态估计，posterior

$$p\left(x_k \mid y_{1:k}\right) = \frac{1}{Z_k} p\left(y_k \mid x_k\right) p\left(x_k \mid y_{1:k-1}\right)$$

HMM模型，递归贝叶斯，求解posterior

看起来很美，如何确定分布形式？如何积分？

——根据不同的应用环境做不同的假设

卡尔曼滤波、扩展卡尔曼滤波、无迹卡尔曼滤波、粒子滤波…

贝叶斯滤波

- 卡尔曼滤波 Kalmanlfilter：高斯噪声+线性函数

$$\begin{cases} x_k = f\left(x_{k-1}, u_k\right) + w_k \\ y_k = h\left(x_k\right) + v_k \end{cases} \Rightarrow \begin{cases} x_k = Ax_{k-1} + q_{k-1} \\ y_k = Hx_k + r_k \end{cases}, \quad \begin{matrix} q_{k-1} \sim N\left(0, Q_{k-1}\right) \\ r_{k-1} \sim N\left(0, R_{k-1}\right) \end{matrix}$$

- 扩展卡尔曼滤波：非线性近似滤波，泰勒展开，转化为线性高斯模型
- 无迹卡尔曼滤波Unscented Kalman Filter，后验概率解析计算困难，转而采用一定规律的采样和权重，可以近似获得均值和方差，确定性地选取一定数量的采样点，这些点的分布和原分布式一致的

- 粒子滤波：引入蒙特卡洛随机采样来计算后验概率，噪声不再假设为高斯形式

高斯变量的联合分布（条件->联合）

对于两个高斯分布随机变量$\boldsymbol{x} \sim \mathbb{R}^n$和$\boldsymbol{y} \sim \mathbb{R}^m$，已知边缘分布$p(\boldsymbol{x})$和条件分布$p(\boldsymbol{y}|\boldsymbol{x})$

$$p(\boldsymbol{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{P}) \qquad p(\boldsymbol{y}|\boldsymbol{x}) \sim N(\boldsymbol{H}\boldsymbol{x} + \boldsymbol{u}, \boldsymbol{R})$$

则$\boldsymbol{x}$和$\boldsymbol{y}$的联合分布以及$\boldsymbol{y}$的边缘分布如下：

$$\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{H}\boldsymbol{\mu} + \boldsymbol{u} \end{bmatrix}, \begin{bmatrix} \boldsymbol{P} & \boldsymbol{P}\boldsymbol{H}^T \\ \boldsymbol{H}\boldsymbol{P} & \boldsymbol{H}\boldsymbol{P}\boldsymbol{H}^T + \boldsymbol{R} \end{bmatrix}\right)$$

$$\boldsymbol{y} \sim N(\boldsymbol{H}\boldsymbol{\mu} + \boldsymbol{u},\ \boldsymbol{H}\boldsymbol{P}\boldsymbol{H}^T + \boldsymbol{R})$$

高斯变量的条件分布（联合->条件）

已知高斯随机变量$\boldsymbol{x}$和$\boldsymbol{y}$的联合分布为：

$$\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix}, \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C} \\ \boldsymbol{C}^T & \boldsymbol{B} \end{bmatrix}\right)$$

写出对应的概率密度函数并进行整理简化即得

则$\boldsymbol{x}$和$\boldsymbol{y}$各自的边缘分布以及条件分布如下：

$$p(\boldsymbol{x}) \sim N(\boldsymbol{a}, \boldsymbol{A}) \qquad p(\boldsymbol{y}) \sim M(\boldsymbol{b}, \boldsymbol{B})$$

$$p(\boldsymbol{x}|\boldsymbol{y}) \sim N(\boldsymbol{a} + \boldsymbol{C}\boldsymbol{B}^{-1}(\boldsymbol{y} - \boldsymbol{b}), \boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{C}^T)$$

$$p(\boldsymbol{y}|\boldsymbol{x}) \sim N(\boldsymbol{b} + \boldsymbol{C}^T\boldsymbol{A}^{-1}(\boldsymbol{x} - \boldsymbol{a}), \boldsymbol{B} - \boldsymbol{C}^T\boldsymbol{A}^{-1}\boldsymbol{C})$$

卡尔曼滤波 Kalmanlfilter

- 模型：高斯噪声+线性函数

$$
\begin{cases}
\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{q}_{k-1} \\
\mathbf{z}_k = \mathbf{H}\mathbf{x}_k + \mathbf{r}_k \\
\mathbf{q}_{k-1} \sim \mathrm{N}\left(0, \mathbf{Q}_{k-1}\right) \\
\mathbf{r}_k \sim \mathrm{N}\left(0, \mathbf{R}_k\right) \\
\mathbf{x}_0 \sim \mathrm{N}\left(\boldsymbol{\mu}_0, \mathbf{P}_0\right)
\end{cases}
\Rightarrow
\begin{aligned}
p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) &\sim N(\boldsymbol{A}_{k-1}\boldsymbol{x}_{k-1}, \boldsymbol{Q}_{k-1}) \\
p(\boldsymbol{z}_k|\boldsymbol{x}_k) &\sim N(\boldsymbol{H}_k\boldsymbol{x}_k, \boldsymbol{R}_k)
\end{aligned}
$$

- 状态预测：根据k-1时刻状态，对k时刻状态进行预测

- 测量更新：根据k时刻测量，计算后验概率

卡尔曼滤波 Kalmanlfilter

状态预测：根据k-1时刻状态，对k时刻状态进行预测

- 求解联合分布：

$$
\begin{aligned}
\begin{bmatrix} \boldsymbol{x}_k \\ \boldsymbol{x}_{k-1} \end{bmatrix} &\sim p(\boldsymbol{x}_k, \boldsymbol{x}_{k-1} | \boldsymbol{z}_{1:k-1}) \\
&= p(\boldsymbol{x}_k | \boldsymbol{x}_{k-1}) p(\boldsymbol{x}_{k-1}) \\
&= N(\boldsymbol{A}_{k-1}\boldsymbol{x}_{k-1}, \boldsymbol{Q}_{k-1}) N(\boldsymbol{\mu}_{k-1}, \boldsymbol{P}_{k-1}) \\
&= N\left(\begin{bmatrix} \boldsymbol{\mu}_{k-1} \\ \boldsymbol{A}_{k-1}\boldsymbol{\mu}_{k-1} \end{bmatrix}, \begin{bmatrix} \boldsymbol{P}_{k-1} & \boldsymbol{P}_{k-1}\boldsymbol{A}_{k-1}^T \\ \boldsymbol{A}_{k-1}\boldsymbol{P}_{k-1} & \boldsymbol{A}_{k-1}\boldsymbol{P}_{k-1}\boldsymbol{A}_{k-1}^T + \boldsymbol{Q}_k \end{bmatrix} \right)
\end{aligned}
$$

- 计算边缘分布

$$
\begin{aligned}
p(\boldsymbol{x}_k | \boldsymbol{z}_{\mathbf{1:k-1}}) = p(\boldsymbol{x}_k) &\sim N(\boldsymbol{A}_{k-1}\boldsymbol{\mu}_{k-1}, \boldsymbol{A}_{k-1}\boldsymbol{P}_{k-1}\boldsymbol{A}_{k-1}^T + \boldsymbol{Q}_k) \\
&\sim N(\hat{\boldsymbol{\mu}}_k, \boldsymbol{P}_k)
\end{aligned}
$$

卡尔曼滤波 Kalmanlfilter

- 测量更新

$$
\begin{aligned}
p(\boldsymbol{z}_k,\boldsymbol{x}_k|\boldsymbol{z}_{1:k-1}) &= p(\boldsymbol{z}_k|\boldsymbol{x}_k)p(\boldsymbol{x}_k|\boldsymbol{z}_{k-1}) \\
&= N(\boldsymbol{H}_k\boldsymbol{x}_k,\boldsymbol{R}_k)N(\hat{\boldsymbol{\mu}}_k,\hat{\boldsymbol{P}}_k) \\
\Rightarrow \begin{bmatrix} \boldsymbol{x}_k \\ \boldsymbol{z}_k \end{bmatrix} &\sim N(\begin{bmatrix} \hat{\boldsymbol{\mu}}_k \\ \boldsymbol{H}_k\hat{\boldsymbol{\mu}}_k \end{bmatrix}, \begin{bmatrix} \hat{\boldsymbol{P}}_k & \hat{\boldsymbol{P}}_k\boldsymbol{H}_k^T \\ \boldsymbol{H}_k\hat{\boldsymbol{P}}_k & \boldsymbol{H}_k\hat{\boldsymbol{P}}_k\boldsymbol{H}^T+\boldsymbol{R}_k \end{bmatrix})
\end{aligned}
$$

$$
p(\boldsymbol{x}_k|\boldsymbol{z}_k,\boldsymbol{z}_{1:k-1}) = p(\boldsymbol{x}_k|\boldsymbol{z}_k) \sim N(\tilde{\boldsymbol{\mu}}_k,\tilde{\boldsymbol{P}}_k)
$$

其中：

增益 残差

$$
\tilde{\boldsymbol{\mu}}_k = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{P}}_k\boldsymbol{H}_k^T\left(\boldsymbol{H}_k\hat{\boldsymbol{P}}_k\boldsymbol{H}^T+\boldsymbol{R}_k\right)^{-1}(\boldsymbol{z}_k-\boldsymbol{H}_k\hat{\boldsymbol{\mu}}_k) = \hat{\boldsymbol{\mu}} + \boldsymbol{K}\hat{\boldsymbol{z}}_k
$$

$$
\tilde{\boldsymbol{P}}_k = \hat{\boldsymbol{P}}_k - \hat{\boldsymbol{P}}_k\boldsymbol{H}_k^T\left(\boldsymbol{H}_k\hat{\boldsymbol{P}}_k\boldsymbol{H}^T+\boldsymbol{R}_k\right)^{-1}\boldsymbol{H}_k\hat{\boldsymbol{P}}_k = \hat{\boldsymbol{P}}_k - \boldsymbol{K}\boldsymbol{H}_k\hat{\boldsymbol{P}}_k
$$

卡尔曼滤波 Kalmanlfilter

- 模型：高斯噪声+线性函数

$$p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) \sim N(\boldsymbol{A}_{k-1}\boldsymbol{x}_{k-1}, \boldsymbol{Q}_{k-1})$$
$$p(\boldsymbol{z}_k|\boldsymbol{x}_k) \sim N(\boldsymbol{H}_k\boldsymbol{x}_k, \boldsymbol{R}_k)$$

$$\mathbf{x}_0 \sim \mathrm{N}\left(\boldsymbol{\mu}_0, \mathbf{P}_0\right)$$

- 状态预测：

$$p(\boldsymbol{x_k}|\boldsymbol{z_{1:k-1}}) \sim N(\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{P}}_k) \begin{cases} \hat{\boldsymbol{\mu}}_k = \boldsymbol{A}_{k-1}\boldsymbol{\mu}_{k-1} \\ \hat{\boldsymbol{P}}_k = \boldsymbol{A}_{k-1}\boldsymbol{P}_{k-1}\boldsymbol{A}_{k-1}^T + \boldsymbol{Q}_k \end{cases}$$

- 测量更新：

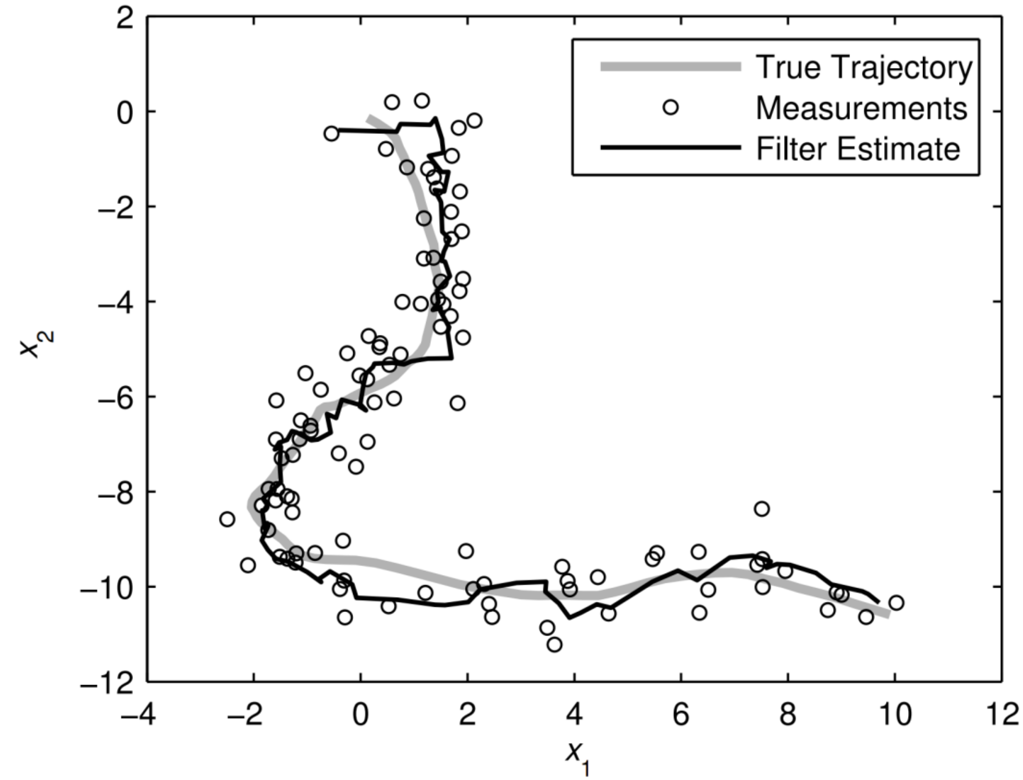
$$p(\boldsymbol{x}_k|\boldsymbol{z}_k) \sim N(\tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{P}}_k) \begin{cases} \boldsymbol{K} = \hat{\boldsymbol{P}}_k\boldsymbol{H}_k^T\left(\boldsymbol{H}_k\hat{\boldsymbol{P}}_k\boldsymbol{H}^T + \boldsymbol{R}_k\right)^{-1} \\ \tilde{\boldsymbol{\mu}}_k = \hat{\boldsymbol{\mu}} + \boldsymbol{K}\left(\boldsymbol{z}_k - \boldsymbol{H}_k\hat{\boldsymbol{\mu}}_k\right) \\ \tilde{\boldsymbol{P}}_k = \hat{\boldsymbol{P}}_k - \boldsymbol{K}\boldsymbol{H}_k\hat{\boldsymbol{P}}_k \end{cases}$$

举例

- 二维平面目标跟踪，

$$\mathbf{x}=(\underbrace{x_1,x_2}_{\text{位移}},\underbrace{x_3,x_4}_{\text{速度}})$$

$$\begin{cases} x_k = Ax_{k-1} + q_{k-1} \\ y_k = Hx_k + r_k \end{cases},$$



$$\mathbf{A} = \begin{pmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} \frac{q_1^c\,\Delta t^3}{3} & 0 & \frac{q_1^c\,\Delta t^2}{2} & 0 \\ 0 & \frac{q_2^c\,\Delta t^3}{3} & 0 & \frac{q_2^c\,\Delta t^2}{2} \\ \frac{q_1^c\,\Delta t^2}{2} & 0 & q_1^c\,\Delta t & 0 \\ 0 & \frac{q_2^c\,\Delta t^2}{2} & 0 & q_2^c\,\Delta t \end{pmatrix}$$

$$\mathbf{H} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$