Chapter 3: 最大似然估计和贝叶斯参数估计

1

中山大學

要点:

- 重点掌握最大似然估计和贝叶斯参数估计的原理;
- 熟练掌握主成分分析和Fisher线性分析;
- 掌握隐马尔可夫模型;
- 了解维数问题;

2

2

中山大學

本章目录

3

3

中山大學

本章目录(*Cont.*)

4

4

• 3.1 引 言

中山大學

- 贝叶斯框架下的数据收集
  - 在以下条件下我们可以设计一个可选择的分类器：
    - $P(\omega_i)$ (先验)
    - $P(x \mid \omega_i)$ (类条件密度)

  不幸的是，我们极少能够完整的得到这些信息!

- 从一个传统的样本中设计一个分类器
  - 先验估计不成问题
  - 对类条件密度的估计存在两个问题：
    - 样本对于类条件估计太少了；
    - 特征空间维数太大了，计算复杂度太高。

5

5

中山大學

- 如果可以将类条件密度参数化，则可以显著降低难度。
- 例如：$P(x \mid \omega_i)$的正态性

$$P(x \mid \omega_i) \sim N(\mu_i, \Sigma_i)$$

  - 用两个参数表示

  将概率密度估计问题转化为参数估计问题。

- 估计
  - 最大似然估计 (ML) 和贝叶斯估计；
  - 结果通常很接近, 但是方法本质是不同的。

6

6

中山大學

- 最大似然估计将参数看作是确定的量，只是其值是未知! 通过最大化所观察的样本概率得到最优的参数—用分析方法。

- 贝叶斯方法将参数当成服从某种先验概率分布的随机变量。对样本进行观测的过程，就是把先验概率密度转化成为后验概率密度，每得到一个新的观测样本，都使得后验概率密度函数变得更加尖锐，使其在待估参数的真实值附近形成最大尖峰。

- 在参数估计完后，两种方法都用后验概率**P(ω$_i$ | x)**表示分类准则!

7

7

中山大學

• 3.2 最大似然估计

□ 最大似然估计的优点：
- 当样本数目增加时，收敛性质会更好；
- 比其它可选择的技术更加简单 。

3.2.1 基本原理

假设有**c**类样本，并且

1）每个样本集的样本都是独立同分布的随机变量；

2）$p(\mathbf{x}|\omega_j)$形式已知但参数未知，

例如 $p(\mathbf{x}|\omega_j) \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ ；

3）记$p(\mathbf{x}|\omega_j) \equiv p(\mathbf{x}|\omega_j,\ \Theta_j)$ ，其中 $\theta_j = (\mu_j, \Sigma_j)$

8

8

- 使用训练样本提供的信息估计 $\theta = (\theta_1, \theta_2, \ldots, \theta_c)$, 每个 θ_i (i = 1, 2, …, c) 只和每一类相关 。

- 假定D包括n个样本, $x_1, x_2, \ldots, x_n$

$$p(D\,|\,\theta) = \prod_{k=1}^{k=n} p(x_k\,|\,\theta) = F(\theta)$$

$p(D\,|\,\theta)$ 被称为样本集D下的似然函数

- θ的最大似然估计是通过定义最大化 ***p*(D | θ)**的值 $\hat{\theta}$
“θ值与实际观察中的训练样本最相符”

[2] 9

9

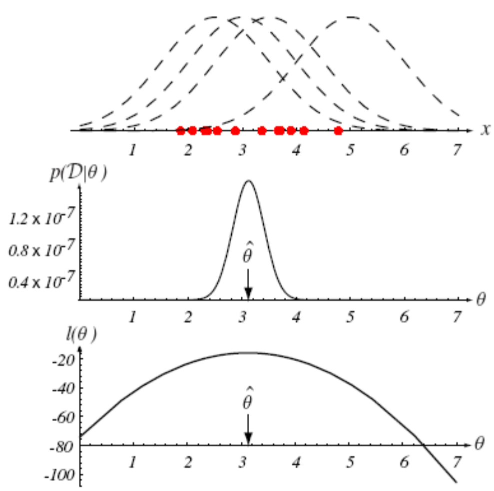


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean.

10

10

- 最优估计
  - 令θ = ($\theta_1, \theta_2, \ldots, \theta_p$)^t 并令 ∇_θ 为梯度算子 **the gradient operator**

$$\nabla_{\theta} = \left[\frac{\partial}{\partial \theta_1}, \frac{\partial}{\partial \theta_2}, \ldots, \frac{\partial}{\partial \theta_p} \right]^t$$

  - 我们定义 **L(θ)** 为对数似然函数: $l(\Theta) \equiv \ln p(D \mid \Theta)$
  - 新问题陈述:
    求解 θ 为使对数似然最大的值

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

11

对数似然函数**L(θ)**显然是依赖于样本集**D**, 有:

$$l(\theta) = \sum_{k=1}^{n} \ln p(x_k \mid \theta)$$

最优求解条件如下:

$$\nabla_\theta l(\theta) = \sum_{k=1}^{n} \nabla_\theta \ln P(x_k \mid \theta)$$

令:

$$\nabla_\theta l(\theta) = 0$$

来求解.

12

3.2.2 高斯情况： μ未知

中 山 大 學

- $P(x_k \mid \mu) \sim N(\mu, \Sigma)$
(样本从一组多变量正态分布中提取)

$$\ln p(\mathbf{x_k} \mid \boldsymbol{\mu}) = -\frac{1}{2}\ln\left[(2\pi)^d \left|\boldsymbol{\Sigma}\right|\right] - \frac{1}{2}(\mathbf{x_k} - \boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x_k} - \boldsymbol{\mu})$$

$$\text{和} \quad \nabla_\mu \ln p(\mathbf{x_k} \mid \boldsymbol{\mu}) = \Sigma^{-1}(\mathbf{x_k} - \boldsymbol{\mu})$$

θ = μ，因此:

- μ的最大似然估计必须满足:

$$\sum_{k=1}^{n} \Sigma^{-1}(\mathbf{x_k} - \hat{\boldsymbol{\mu}}) = 0$$

2 13

13

中 山 大 學

- 乘 Σ 并且重新排序, 我们得到:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x_k}$$

这是一个非常好的结果！即：对均值的最大似然估计就是对全体训练样本的算术平均值!

结论:

如果$P(x_k \mid \omega_j)$ (j = 1, 2, …, c)被假定为*d*维特征空间中的高斯分布; 然后我们能够估计向量 $\theta = (\theta_1, \theta_2, \ldots, \theta_c)^t$ 从而得到最优分类!

2 14

14

3.2.3 高斯情况：μ和 Σ均未知

中山大學

- **未知 μ 和 σ，对于单样本x_k**

  $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$

$$l(\theta) = \ln P(x_k \mid \theta) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_\theta l = \begin{pmatrix} \frac{\sigma}{\sigma\theta_1}(\ln P(x_k \mid \theta)) \\ \frac{\sigma}{\sigma\theta_2}(\ln P(x_k \mid \theta)) \end{pmatrix} = 0$$

$$\begin{cases} \frac{1}{\theta_2}(x_k - \theta_1) = 0 \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} = 0 \end{cases}$$

15

15

中山大學

对于全部样本，最后得到:

$$\begin{cases} \sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 & (1) \\ -\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 & (2) \end{cases}$$

联合公式 (1) 和 (2),并用 得到如下结果:

$$\mu = \sum_{k=1}^{n} \frac{x_k}{n} \quad ; \quad \sigma^2 = \frac{\sum_{k=1}^{n}(x_k - \mu)^2}{n}$$

2 16

16

3.2.4 偏差估计

- σ^2的最大似然估计是有偏的 （渐进无偏估计）

$$E\left[\frac{1}{n}\Sigma(x_i-\bar{x})^2\right]=\frac{n-1}{n}\sigma^2\neq\sigma^2$$

- Σ的一个基本的无偏估计是:

$$\underbrace{\mathrm{C}=\frac{1}{\text{n-1}}\sum_{k=1}^{k=n}(x_k-\hat{\mu})(x_k-\hat{\mu})^t}_{\textit{Sample}\ \text{covariance matrix}}$$

2 17

17

模型错误会怎么样?

达不到最优!

18

18

3.3贝叶斯估计

中山大學

- 在最大似然估计中 θ 被假定为固定值
- 在贝叶斯估计中 θ 是随机变量

3.3.1 类条件密度

- 目标: 计算 P(ω_i | x, D)

假设样本集为D，贝叶斯方程可以写成：

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

19

中山大學

- 先验概率通常可以事先获得，因此

$$P(\omega_i) = P(\omega_i \mid D)$$

- 每个样本只依赖于所属的类，有：

$$P(x \mid \omega_i, D) = P(x \mid \omega_i, D_i)$$

故：

$$P(\omega_i \mid x, D) = \frac{P(x \mid \omega_i, D_i).P(\omega_i)}{\sum_{j=1}^{c} P(x \mid \omega_j, D_j).P(\omega_j)}$$

即：只要在每类中独立计算 $P(x \mid \omega_i, D_i)$ 就可以确定x的类别。

因此，核心工作就是要估计 $P(x \mid D)$

20

3.3.2 参数的分布

中山大學

- 假设 $p(x)$ 的形式已知, 参数θ的值未知，因此条件概率密度 $p(x|\theta)$ 的函数形式是知道的；
- 假设参数θ是随机变量，先验概率密度函数**p(θ)**已知，利用贝叶斯公式可以计算后验概率密度函数**p(θ | D)**；
- 希望后验概率密度函数**p(θ | D)** 在θ的真实值附件有非常显著的尖峰，则可以使用后验密度**p(θ | D)** 估计 θ；

21

21

中山大學

- 注意到

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

如果*p*(θ | D) 在某个值 $\hat{\theta}$ 附件有非常显著的尖峰，
则 $p(x|D)\approx p(x|\hat{\theta})$

即: 如果条件概率密度具有一个已知的形式，则利用已有的训练样本，就能够通过**p(θ | D)** 对**p(x | D)** 进行估计。

22

22

3.4 贝叶斯参数估计: 高斯过程

中山大學

- 单变量情形的 p(μ | D)

$p(x|\mu) \sim N(\mu, \sigma^2)$, μ 是未知的。

假设 $p(\mu) \sim N(\mu_0, \sigma_0^2)$, μ_0 和 σ_0^2 已知

(μ_0是μ最好的估计; σ_0^2是该估计的不确定性)

$$D = \{x_1, \cdots, x_n\}, \quad p(\mu|D) = \frac{p(D|\mu)p(\mu)}{\int p(D|\mu)p(\mu)d\mu}$$

$$p(\mu|D) = \alpha \prod_{k=1}^{n} p(x_k|\mu)p(\mu)$$

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

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

23

23

再生密度

中山大學

$p(\mu|D) \sim N(\mu_n, \sigma_n^2)$ [reproducing density]

[称$p(\mu)$: conjugate prior]

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}, \quad \frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2}$$

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

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

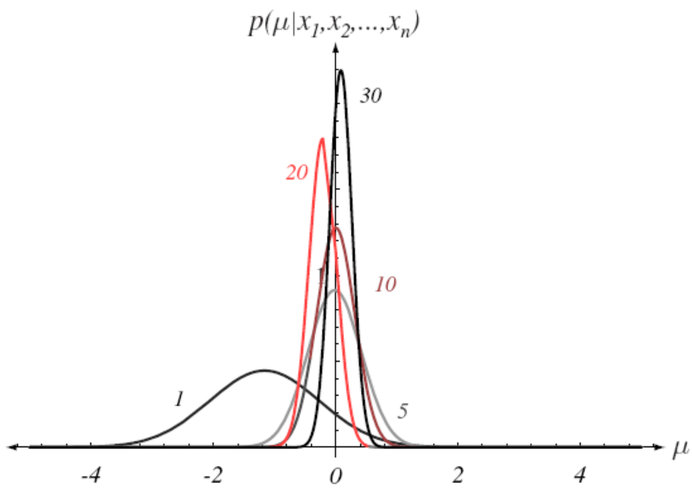
其中, $\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$

24

24

贝叶斯学习

中山大學



25

25

结论:

中山大學

μ_n 是 $\hat{\mu}_n$ 和 μ_0的线性组合，

总是位于$\hat{\mu}_n$ 和 μ_0 的连线上；

当σ_0^2 有限时，μ_n 将逼近 $\hat{\mu}_n$，否则$\mu_n = \mu_0$ 。

26

26

□ 单变量情形的 p(x|D)

中山大學

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

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

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

$$=\sqrt{2\pi\left(\frac{\sigma^2\sigma_n^2}{\sigma^2+\sigma_n^2}\right)}$$

$$p(x|D)\sim N(\mu_n,\sigma^2+\sigma_n^2)$$

27

27

多变量情形:

中山大學

$$p(\mathbf{x}|\boldsymbol{\mu})\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma}),\ p(\boldsymbol{\mu})\sim N(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$$

$$D=\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}$$

$$p(\boldsymbol{\mu}|D)=\alpha\prod_{k=1}^{n}p(x_k|\boldsymbol{\mu})p(\boldsymbol{\mu})$$

$$=\alpha'\exp\left[-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_n)^t\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_n)\right]$$ 再生密度

$$\boldsymbol{\Sigma}_n^{-1}=n\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_0^{-1},\quad \boldsymbol{\Sigma}_n^{-1}\boldsymbol{\mu}_n=n\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_n+\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0$$

其中，$\hat{\boldsymbol{\mu}}_n=\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$

28

28

利用$\left(\mathbf{A}^{-1}+\mathbf{B}^{-1}\right)^{-1}=\mathbf{A}\left(\mathbf{A}+\mathbf{B}\right)^{-1}\mathbf{B}=\mathbf{B}\left(\mathbf{A}+\mathbf{B}\right)^{-1}\mathbf{A}$，得

$$\boldsymbol{\mu}_n=\boldsymbol{\Sigma}_0\left(\boldsymbol{\Sigma}_0+\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\hat{\boldsymbol{\mu}}_n+\frac{1}{n}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}_0+\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu}_0$$

$$\boldsymbol{\Sigma}_n=\boldsymbol{\Sigma}_0\left(\boldsymbol{\Sigma}_0+\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\frac{1}{n}\boldsymbol{\Sigma}$$

利用 $p(\mathbf{x}\mid D)=\int p(\mathbf{x}\mid\boldsymbol{\mu})p(\boldsymbol{\mu}\mid D)d\boldsymbol{\mu}$

令 $\mathbf{y}=\mathbf{x}-\boldsymbol{\mu}$ $\quad p(\mathbf{y}\mid\mu)\sim N(\mathbf{0},\boldsymbol{\Sigma})$

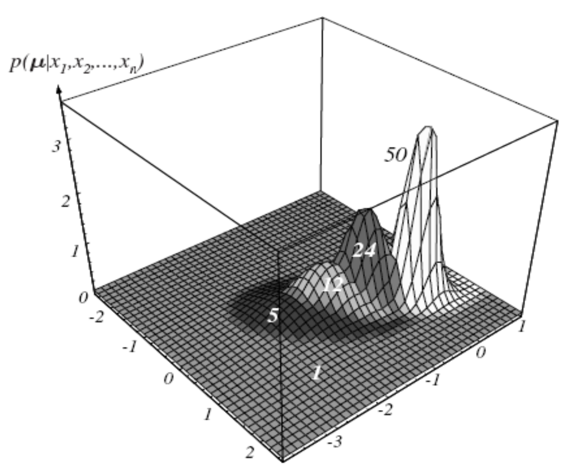
$p(\boldsymbol{\mu}\mid D)\sim N(\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n)$

$\therefore p(\mathbf{x}\mid D)=p(\mathbf{y}+\boldsymbol{\mu}\mid D)\sim N(\boldsymbol{\mu}_n,\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_n)$

29

29

多变量学习



30

30

3.5 贝叶斯参数估计：一般理论

中山大學

- p(x | D) 的计算可推广于所有能参数化未知密度的情况中，基本假设如下:
  - 假定 p(x | θ) 的形式已知，但是θ的值未知。
  - θ被假定为满足一个已知的先验密度 *p*(θ)
  - 其余的 θ的信息包含在集合D中，其中D是由n维随机变量x_1, x_2, …, x_n组成的集合，它们服从于概率密度函数p(x)。

基本的问题是:

计算后验概率密度*p*(θ | D)，然后 推导出 *p*(x | D)。

31

31

中山大學

$$p(\mathbf{x} \mid D) = \int p(\mathbf{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta} \qquad (25)$$

$$p(\boldsymbol{\theta} \mid D) = \frac{p(D \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(D \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

$$p(D \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

问题:

p(x | D)是否能收敛到p(x)，计算复杂度如何？

32

32

递归贝叶斯学习

中山大學

$$D^n = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}, \quad p(D^n \mid \boldsymbol{\theta}) = p(\mathbf{x}_n \mid \boldsymbol{\theta}) p(D^{n-1} \mid \boldsymbol{\theta})$$

$$p(\boldsymbol{\theta} \mid D^n) = \frac{p(D^n \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(D^n \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

$$= \frac{p(\mathbf{x}_n \mid \boldsymbol{\theta}) p(D^{n-1} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(\mathbf{x}_n \mid \boldsymbol{\theta}) p(D^{n-1} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

$$p(\boldsymbol{\theta} \mid D^{n-1}) = \frac{p(D^{n-1} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(D^{n-1} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

$$p(\boldsymbol{\theta} \mid D^n) = \frac{p(\mathbf{x}_n \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D^{n-1})}{\int p(\mathbf{x}_n \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D^{n-1}) d\boldsymbol{\theta}}$$

$$p(\boldsymbol{\theta} \mid D^0) = p(\boldsymbol{\theta})$$

该过程称为参数估计的递归贝叶斯方法，一种增量学习方法。

33

33

例1：递归贝叶斯学习

中山大學

假设：$p(x \mid \theta) \sim U(0,\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & \text{其他} \end{cases}$

$$p(\theta) \sim U(0,10), \quad D = \{4,7,2,8\}$$

$$p(D^0 \mid \theta) = p(\theta) \sim U(0,10)$$

$$p(\theta \mid D^1) \propto p(x_1 \mid \theta) p(\theta \mid D^0) = \begin{cases} 1/\theta & \text{对于 } 4 \le \theta \le 10 \\ 0 & \text{其他} \end{cases}$$

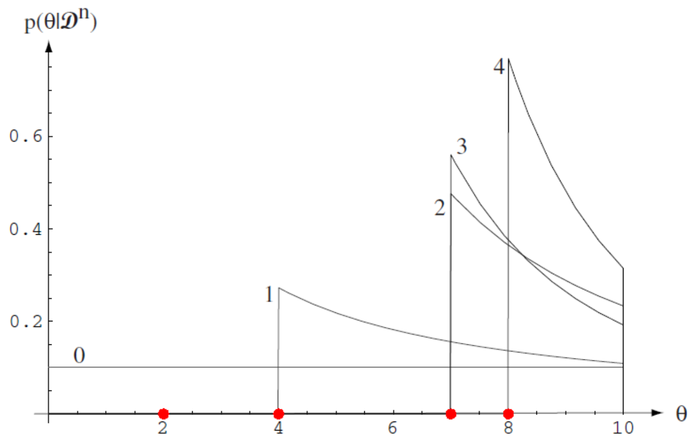
$$p(\theta \mid D^2) \propto p(x_2 \mid \theta) p(\theta \mid D^1) = \begin{cases} 1/\theta^2 & \text{对于 } 7 \le \theta \le 10 \\ 0 & \text{其他} \end{cases}$$

$$p(\theta \mid D^n) \propto 1/\theta^n \quad \text{对于 } \max_x \left[D^n \right] \le \theta \le 10$$

34

34

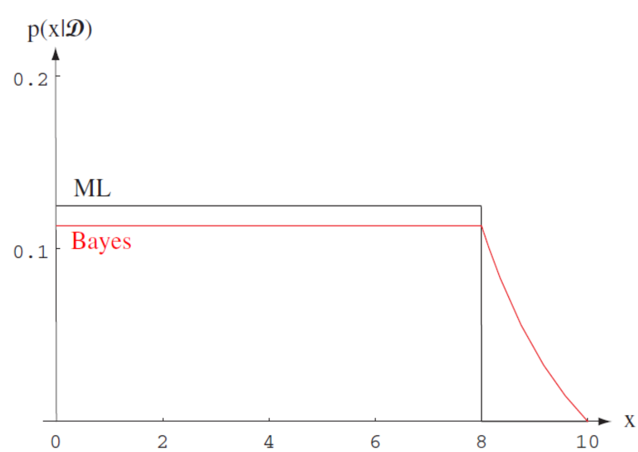
例1：递归贝叶斯学习



35

35

例1: Bayes vs. ML



贝叶斯参数估计：$p(\mathbf{x}|D)=\int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|D)d\boldsymbol{\theta}$

36

36

唯一性问题

- $p(\mathbf{x}|\theta)$ 是唯一的：
  - 后验概率序列 $p(\theta|D^n)$ 收敛到 delta 函数；
  - 只要训练样本足够多，则 $p(\mathbf{x}|\theta)$ 能唯一确定θ。

在某些情况下，不同 θ 值会产生同一个 $p(\mathbf{x}|\theta)$ 。

$p(\theta|D^n)$ 将在 θ 附近产生峰值，这时不管$p(\mathbf{x}|\theta)$ 是否唯一， $p(\mathbf{x}|D^n)$总会收敛到$p(\mathbf{x})$ 。

因此不确定性客观存在。

37

37

最大似然估计和贝叶斯参数估计的区别

	最大似然估计	贝叶斯参数估计
计算复杂度	微分	多重积分
可理解性	确定易理解	不确定不易理解
先验信息的信任程度	不准确	准确
例如 $p(\mathbf{x}\|\theta)$	与初始假设一致	与初始假设不一致

38

38

分类误差

- **贝叶斯误差或不可分误差，例如 $p(\mathbf{x}|\omega_i)$之间相互重叠引起，固有问题，无法消除；**
- **模型错误，ML与Bayes犯错一样；**
- **估计错误，训练样本个数有限产生。**

Gibbs 算法

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

依据$p(\boldsymbol{\theta}|D)$选择 $\boldsymbol{\theta}=\boldsymbol{\theta}_0$

使得 $p(\mathbf{x}|D)\approx p(\mathbf{x}|\boldsymbol{\theta}_0)$ [Gibbs算法]

在较弱的假设条件下，Gibbs算法的误差概率至多是贝叶斯最优分类器的两倍。

• 3.6 充分统计量

中山大學

- 统计量
  - 任何样本集的函数;
- 充分统计量即是一个样本集D的函数s，其中s包含了有助于估计参数θ的所有信息，即$p(D|s,\theta)$与θ无关;
- 如果θ是随机变量，则

$$p(\theta \mid s, D) = \frac{p(D \mid s, \theta)\, p(\theta \mid s)}{p(D \mid s)} = p(\theta \mid s)$$

反之亦然，即如果统计量s能使得 $p(\theta \mid s, D) = p(\theta \mid s)$

并且 $p(\theta \mid s) \neq 0$，那么$p(D|s,\theta)$与θ无关。

41

41

因式分解定理:

中山大學

- 一个关于参数θ的统计量**s是**充分统计量，当且仅当概率分布函数 $P(D|\theta)$ 能够写成乘积形式:

$$P(D|\theta) = g(\mathrm{s}, \theta)\, h(D)$$

其中 $g(.,.)$ 和$h(.)$是两个函数。

42

42

例子：多维高斯分布

中山大學

已知协方差；未知均值

$$p(\mathbf{x}\,|\,\boldsymbol{\theta}) \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$$

$$p(D\,|\,\boldsymbol{\theta}) = \prod_{k=1}^{n} \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_k-\boldsymbol{\theta})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\theta})\right]$$

$$= \exp\left[-\frac{n}{2}\boldsymbol{\theta}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}^t\boldsymbol{\Sigma}^{-1}\left(\sum_{k=1}^{n}\mathbf{x}_k\right)\right] \times$$

$$\frac{1}{(2\pi)^{nd/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left[-\frac{1}{2}\sum_{k=1}^{n}\mathbf{x}_k^t\boldsymbol{\Sigma}^{-1}\mathbf{x}_k\right]$$

$\mathbf{s} = \sum_{k=1}^{n}\mathbf{x}_k$ and thus $\hat{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$ are sufficient for $\boldsymbol{\theta}$

43

43

证明：必要性

中山大學

假设 s 是关于$\boldsymbol{\theta}$ 的充分统计量，即

$P(D\,|\,\mathbf{s},\boldsymbol{\theta})$不依赖于$\boldsymbol{\theta}$

$$P(D\,|\,\boldsymbol{\theta}) = \sum_{\mathbf{s}} P(D,\mathbf{s}\,|\,\boldsymbol{\theta}) = \sum_{\mathbf{s}} P(D\,|\,\mathbf{s},\boldsymbol{\theta})P(\mathbf{s}\,|\,\boldsymbol{\theta})$$

$$= P(D\,|\,\mathbf{s},\boldsymbol{\theta})P(\mathbf{s}\,|\,\boldsymbol{\theta})$$

$$= h(D)P(\mathbf{s}\,|\,\boldsymbol{\theta}) = h(D)g(\mathbf{s},\boldsymbol{\theta})$$

注意到 $\mathbf{s} = \varphi(D)$, 对于一个给定的样本集，只有一个**s**与之对应。

44

44

充分性：

$$\mathbf{s}=\varphi(D),\quad \bar{D}=\{D\,|\,\varphi(D)=\mathbf{s}\}$$

$$P(D\,|\,\mathbf{s},\boldsymbol{\theta})=\frac{P(D,\mathbf{s}\,|\,\boldsymbol{\theta})}{P(\mathbf{s}\,|\,\boldsymbol{\theta})}=\frac{P(D,\mathbf{s}\,|\,\boldsymbol{\theta})}{\sum_{D\in\bar{D}}P(D,\mathbf{s}\,|\,\boldsymbol{\theta})}$$

$$=\frac{P(D\,|\,\boldsymbol{\theta})}{\sum_{D\in\bar{D}}P(D\,|\,\boldsymbol{\theta})}=\frac{g(\mathbf{s},\boldsymbol{\theta})h(D)}{\sum_{D\in\bar{D}}g(\mathbf{s},\boldsymbol{\theta})h(D)}=\frac{h(D)}{\sum_{D\in\bar{D}}h(D)}$$

上式不依赖于$\boldsymbol{\theta}$；

因此 **s** 是关于 $\boldsymbol{\theta}$的充分统计量。

45

核密度（Kernel density）

- 把 $P(D|\boldsymbol{\theta})$ 分解成 $g(\mathbf{s},\boldsymbol{\theta})h(D)$ 不是唯一的：
  - 如果$f(\mathbf{s})$是一个函数, $g'(\mathbf{s},\boldsymbol{\theta})=f(\mathbf{s})g(\mathbf{s},\boldsymbol{\theta})$ 和 $h'(D)=h(D)/f(\mathbf{s})$ 也是等价的分解；
- 这种二义性可以用定义核密度函数的方法来得到消除：

$$\bar{g}(\mathbf{s},\boldsymbol{\theta})=\frac{g(\mathbf{s},\boldsymbol{\theta})}{\int g(\mathbf{s},\boldsymbol{\theta})d\boldsymbol{\theta}}$$

46

例子：多维高斯分布

$$p(\mathbf{x}\mid\boldsymbol{\theta}) \sim N(\boldsymbol{\theta},\boldsymbol{\Sigma})$$

$$p(D\mid\boldsymbol{\theta}) = \exp\left[-\frac{n}{2}\boldsymbol{\theta}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}^t\boldsymbol{\Sigma}^{-1}\left(\sum_{k=1}^{n}\mathbf{x}_k\right)\right]\times$$

$$\frac{1}{(2\pi)^{nd/2}|\boldsymbol{\Sigma}|^{n/2}}\exp\left[-\frac{1}{2}\sum_{k=1}^{n}\mathbf{x}_k^t\boldsymbol{\Sigma}^{-1}\mathbf{x}_k\right]$$

$$= g(\hat{\boldsymbol{\mu}}_n,\boldsymbol{\theta})h(D), \quad \hat{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

$$g(\hat{\boldsymbol{\mu}}_n,\boldsymbol{\theta}) = \exp\left[-\frac{n}{2}(\boldsymbol{\theta}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^t\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_n)\right]$$

$$\bar{g}(\hat{\boldsymbol{\mu}}_n,\boldsymbol{\theta}) = \frac{1}{(2\pi)^{d/2}\left|\frac{1}{n}\boldsymbol{\Sigma}\right|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\mu}}_n)^t\left(\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\mu}}_n)\right]$$

47

47

核密度与参数估计

- **对于最大似然估计情形，只需最大化 $g(s,\theta)$，因为：$P(D|\theta) = g(s,\theta)\,h(D)$**
- **对于贝叶斯估计情形：**

$$p(\boldsymbol{\theta}\mid D) = \frac{p(D\mid\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{\int p(D\mid\boldsymbol{\theta})\,p(\boldsymbol{\theta})\,d\boldsymbol{\theta}} = \frac{g(\mathrm{s},\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{\int g(\mathrm{s},\boldsymbol{\theta})\,p(\boldsymbol{\theta})\,d\boldsymbol{\theta}}$$

  - **如果我们对θ的先验概率不确定，$p(\theta)$ 通常选择均匀分布，则$p(\theta|D)$ 几乎等于核密度；**
  - **如果$p(\mathrm{x}|\theta)$ 可辨识时，$g(s,\theta)$ 通常在某个值处有明显的尖峰，并且如果$p(\theta)$ 在该值处连续并且非零，则$p(\theta|D)$ 将趋近核密度函数。**

48

48

充分统计量与指数族函数

$$p(\mathbf{x}\,|\,\boldsymbol{\theta}) = \alpha(\mathbf{x})\exp\left[\mathbf{a}(\boldsymbol{\theta}) + \mathbf{b}(\boldsymbol{\theta})^t\mathbf{c}(\mathbf{x})\right]$$

$$p(D\,|\,\boldsymbol{\theta}) = \exp\left[n\mathbf{a}(\boldsymbol{\theta}) + \mathbf{b}(\boldsymbol{\theta})^t\sum_{k=1}^{n}\mathbf{c}(\mathbf{x}_k)\right]\prod_{k=1}^{n}\alpha(\mathbf{x}_k)$$

$$= g(\mathbf{s},\boldsymbol{\theta})h(D)$$

$$\mathbf{s} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{c}(\mathbf{x}_k), \quad g(\mathbf{s},\boldsymbol{\theta}) = \exp\left[n\left\{\mathbf{a}(\boldsymbol{\theta}) + \mathbf{b}(\boldsymbol{\theta})^t\mathbf{s}\right\}\right]$$

$$h(D) = \prod_{k=1}^{n}\alpha(\mathbf{x}_k)$$

49

49

- 3.7 维数问题

□分类问题通常涉及50或100维以上的特征.

□分类精度取决于维数和训练样本的数量

- 考虑有相同协方差矩阵的两组多维向量情况:

$$p(\mathbf{x}\,|\,\omega_j) \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}), \quad j = 1,2$$

如果它们的先验概率相同，则贝叶斯误差概率为:

$$P(error) = \frac{1}{\sqrt{2\pi}}\int_{r/2}^{\infty} e^{\frac{-u^2}{2}}\,du$$

其中: $r^2 = (\mu_1 - \mu_2)^t\Sigma^{-1}(\mu_1 - \mu_2)$

$$\lim_{r\to\infty} P(error) = 0$$

50

50

中山大學

- 如果特征是独立的，则有:

$$\Sigma = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_d^2)$$

$$r^2 = \sum_{i=1}^{d} \left(\frac{\mu_{i1} - \mu_{i2}}{\sigma_i} \right)^2$$

- 最有用的特征是两类均值之间的距离大于标准方差的那些特征: $|\mu_1 - \mu_2| > \sigma$

- 在实际观察中我们发现，当特征个数增加到某个临界点后会导致更糟糕的结果而不是好的结果: 我们的模型有误，或者由于训练样本个数有限导致分布估计不精确，等等。

51

51

中山大學

可分性与特征维数

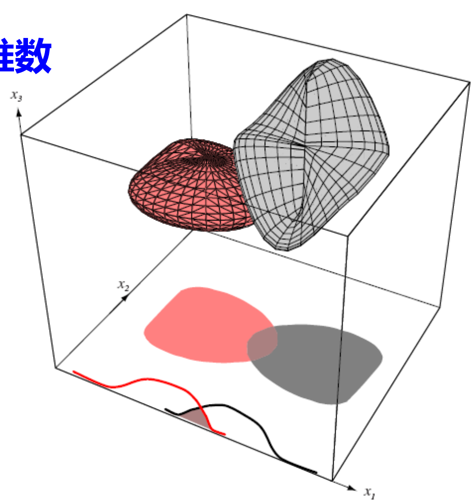


FIGURE 3.3. Two three-dimensional distributions have non-overlapping densities, and thus in three dimensions the Bayes error vanishes. When projected to a subspace here, the two-dimensional x1−x2 subspace or a one-dimensional x1 subspace—there can be greater overlap of the projected distributions, and hence greater Bayes error.

52

52

学习过程的计算复杂度

$$\hat{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k: \quad O(nd)$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_n)(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_n)^t: \quad O(nd^2)$$

求解$d \times d$ 矩阵的逆： $O(d^3)$

求解 $d \times d$ 行列式: $O(d^3)$

$n > d$

$$g(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \underbrace{\hat{\boldsymbol{\mu}}_n}_{O(nd)})^t \underbrace{\hat{\boldsymbol{\Sigma}}^{-1}}_{O(nd^2)}(\mathbf{x} - \hat{\boldsymbol{\mu}}_n) - \underbrace{\frac{1}{2}\ln\left|\hat{\boldsymbol{\Sigma}}\right|}_{O(d^3)} + \underbrace{\ln P(\omega)}_{O(n)} - \underbrace{\frac{d}{2}\ln 2\pi}_{O(1)}$$

53

Complexity of the ML Estimation

$$g(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \overbrace{\hat{\mu}}^{O(d.n)})^t \overbrace{\hat{\Sigma}^{-1}}^{O(n.d^2)} (\mathbf{x} - \hat{\mu}) - \overbrace{\frac{d}{2}\ln 2\pi}^{O(1)} - \underbrace{\frac{1}{2}\ln|\hat{\Sigma}|}_{O(d^2.n)} + \underbrace{\ln P(\omega)}_{O(n)}$$

Total = $O(nd^2)$

Total for c classes = $O(cnd^2) \cong O(nd^2)$

Cost increase when *d* and *n* are large!

54

分类的计算复杂度

中山大學

给定 $\mathbf{x}$

计算 $(\mathbf{x}-\hat{\mu}_n)$: $O(d)$

将协方差矩阵的逆矩阵与差向量相乘: $O(d^2)$

判别 $\max_i g_i(\mathbf{x})$: $O(c)$

整个分类问题的复杂度: $O(d^2)$

➢ **分类阶段比学习阶段简单。**

55

55

训练样本不足时的方法

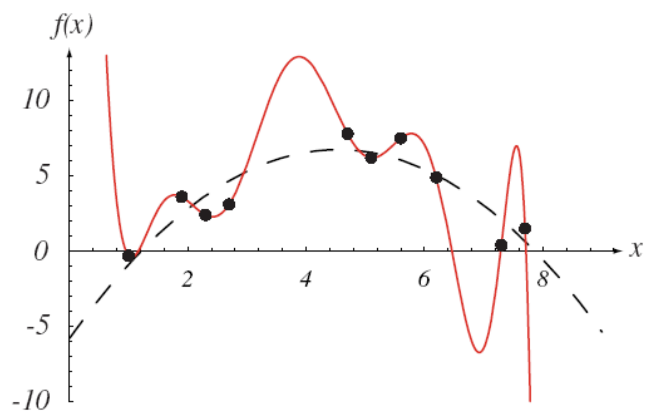
中山大學

- **降维**
  - □ **重新设计特征提取模块;**
  - □ **选择现有特征的子集;**
  - □ **将几个特征组合在一起;**
  - □ **假设各个类的协方差矩阵都相同,将全部数据都归到一起;**
- **寻找协方差矩阵 *S* 更好的估计;**
  - □ **如果有合理的先验估计S_0, 则可以用如下的伪贝叶斯估计** $\lambda\Sigma_0+(1-\lambda)\hat{\Sigma}$ **;**
  - □ **设法将S_0对角化: 阈值化或假设特征之间统计独立;**

56

56

过拟合的概念



正确的拟合思想：一开始用高阶的多项式曲线来拟合，然后依次去掉高阶项来逐渐简化模型，获得更光滑的结果。

缩并 Shrinkage (Regularized Discriminant Analysis)

假设两类分布分别为$N(\mu_1,\mathbf{\Sigma}_1)$和$N(\mu_2,\mathbf{\Sigma}_2)$

i 为 c 个类中的任何一个下标，$\mathbf{\Sigma}$ 是缩并后的协方差，我们有：

$$\mathbf{\Sigma}_i(\alpha)=\frac{(1-\alpha)n_i\mathbf{\Sigma}_i+\alpha n\mathbf{\Sigma}}{(1-\alpha)n_i+\alpha n},\quad 0<\alpha<1$$

或将共同的协方差向单位矩阵缩并为

$$\mathbf{\Sigma}(\beta)=(1\text{-}\beta)\mathbf{\Sigma}+\beta\mathbf{I},\quad 0<\beta<1$$

• 3.8 成分分析与判别函数

中山大學

➢ 采用组合特征降低特征空间的维数
- 线性组合通常比较容易计算和处理
- 将高维数据投影到一个低维空间里去
- 使用两种分类方法寻找理想一点的线性变换:
  - PCA (主成份分析) "在最小均方意义下的原始数据最优表示的映射"
  - MDA (多类判别分析) "在最小均方意义下的数据的最优分类的映射"

59

59

PCA

中山大學

给定 $\mathbf{x}_1,\cdots,\mathbf{x}_n$,找 $\mathbf{x}_0$ 使得

$$J_0(\mathbf{x}_0)=\sum_{k=1}^{n}\|\mathbf{x}_k-\mathbf{x}_0\|^2 \text{ 最小}$$

容易证明，$\mathbf{x}_0=\mathbf{m}$

其中，$\mathbf{m}=\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$

$$J_0(\mathbf{x}_0)=\sum_{k=1}^{n}\|(\mathbf{x}_k-\mathbf{m})-(\mathbf{x}_0-\mathbf{m})\|^2$$

$$=\sum_{k=1}^{n}\|\mathbf{x}_0-\mathbf{m}\|^2+\sum_{k=1}^{n}\|\mathbf{x}_k-\mathbf{m}\|^2$$

该项不依赖$\mathbf{x}_0$

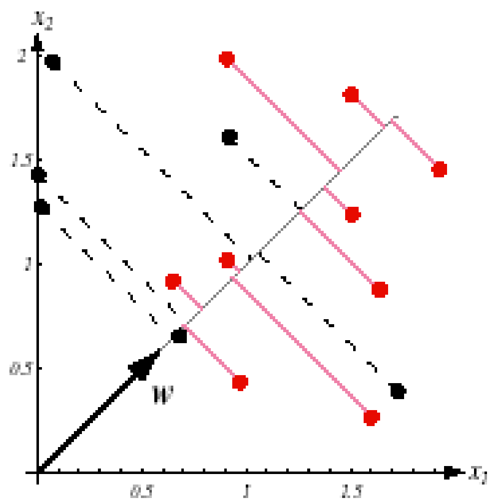
当$\mathbf{x}_0=\mathbf{m}$ 使得minimizes $J_0(\mathbf{x}_0)$

60

60

中山大學

沿直线投影：



中山大學

对于通过样本均值直线的最佳投影

直线：$\mathbf{x} = \mathbf{m} + a\mathbf{e}$

如果用$\mathbf{m} + a_k\mathbf{e}$ 表达 $\mathbf{x}_k$ ，那么通过最小化平方误差准则函数，可以得到一组最优a_k的集合，过程如下：

$$J_1(a_1, \cdots, a_n; \mathbf{e}) = \sum_{k=1}^{n} \left\| (\mathbf{m} + a_k\mathbf{e}) - \mathbf{x}_k \right\|^2$$

$$= \sum_{k=1}^{n} a_k^2 \left\|\mathbf{e}\right\|^2 - 2\sum_{k=1}^{n} a_k \mathbf{e}^t(\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^{n} \left\|\mathbf{x}_k - \mathbf{m}\right\|^2$$

由于$\left\|\mathbf{e}\right\| = 1$, 通过对$a_k$求偏导数，并且令结果为0, 得到：

$$a_k = \mathbf{e}^t(\mathbf{x}_k - \mathbf{m})$$

寻找最佳表达方向

寻找 $\mathbf{e}$， 最小化下面过程:

$$J_1(\mathbf{e}) = \sum_{k=1}^{n} a_k^2 - 2\sum_{k=1}^{n} a_k^2 + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2$$

$$= -\sum_{k=1}^{n} \mathbf{e}^t \underline{(\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t} \mathbf{e} + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2$$

$$= -\mathbf{e}^t \mathbf{S} \mathbf{e} + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2$$

其中,**S**为散布矩阵（Scatter matrix）

即是最大化 $\mathbf{e}^t\mathbf{S}\mathbf{e}$, 满足 $\|\mathbf{e}\|^2 = 1$

Lagrange 方法: 最大化 $u = \mathbf{e}^t\mathbf{S}\mathbf{e} - \lambda(\mathbf{e}^t\mathbf{e} - 1)$

$$\nabla_{\mathbf{e}} u = 0 \Rightarrow \frac{\partial u}{\partial e} = 2\mathbf{S}\mathbf{e} - 2\lambda\mathbf{e} = 0 \Rightarrow \mathbf{S}\mathbf{e} = \lambda\mathbf{e}$$

63

63

$$S = \sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t$$

Scatter Matrix -- 散布矩阵 或 散度矩阵

64

64

主成分分析 (PCA)

—Principal component analysis

由1维推广到d′维：

投影空间：$\mathbf{x} = \mathbf{m} + \sum_{i=1}^{d'} a_i \mathbf{e}_i$

寻找 $\mathbf{e}_i$， $i = 1, \cdots, d'$ 来最小化下式：

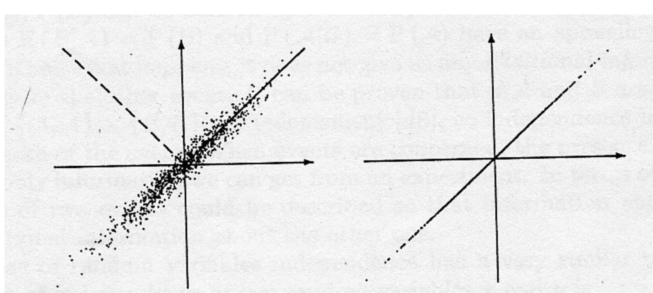
$$J_{d'}(\mathbf{e}_1, \cdots, \mathbf{e}_{d'}) = \sum_{k=1}^{n} \left\| \left(\mathbf{m} + \sum_{i=1}^{d'} a_{ki} \mathbf{e}_i \right) - \mathbf{x}_k \right\|^2$$

$\Rightarrow \mathbf{e}_1, \cdots, \mathbf{e}_{d'}$ 是S的 d' 个具有最大特征值 的特征向量

65

65

- L个N维空间的向量，构成N维空间的L个点。如果大多数点落在一个M维超平面上，只要能找到M维空间的坐标系，则可以将*L*个向量投影到*M*维空间，获得低维的表达。



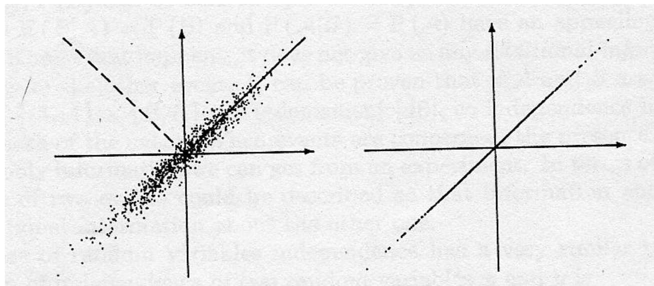
K-L变换　　　PCA

- K-L变换是压缩与特征提取的有效方法。

66

66

中山大學



PCA的变换矩阵是协方差矩阵，**K-L**变换的变换矩阵可以有很多种（二阶矩阵、协方差矩阵、总类内离散度矩阵等等）。

当**K-L**变换矩阵为协方差矩阵时，等同于**PCA**。

67

67

中山大學

3.8.2 Fisher 线性判别分析

—Fisher Linear Discriminant Analysis

考虑两类样本的分类问题：

有一组n个d维样本：$\mathbf{x}_1$、$\mathbf{x}_1$、. . . . 、$\mathbf{x}_n$

其中n_1个样本属于ω_1，n_2个样本属于ω_2。

找 $\mathbf{w}$，最大化分离 $\mathbf{y}=\mathbf{w}^t\mathbf{x}$

$$\mathbf{m}_i=\frac{1}{n_i}\sum_{\mathbf{x}\in D_i}\mathbf{x},\ i=1,2$$

$$\tilde{m}_i=\frac{1}{n_i}\sum_{\mathbf{x}\in D_i}\mathbf{w}^t\mathbf{x}=\mathbf{w}^t\mathbf{m}_i,\quad \tilde{s}_i^2=\sum_{\mathbf{x}\in D_i}(\mathbf{w}^t\mathbf{x}-\tilde{m}_i)^2$$

内类散度矩阵: $\tilde{s}_1^2+\tilde{s}_2^2$

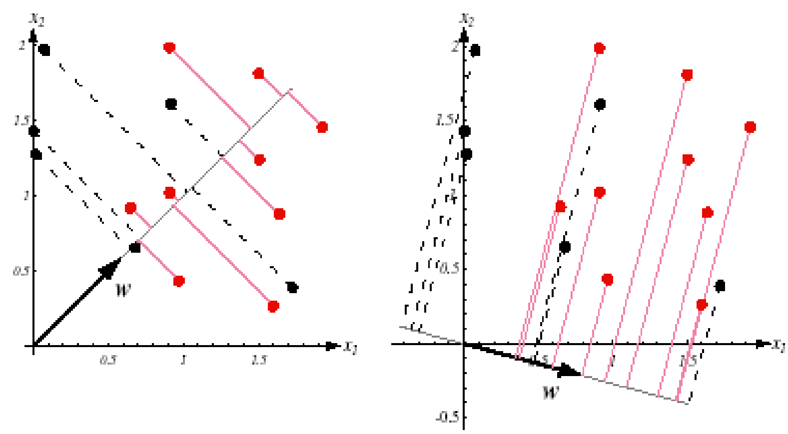
最大化 $J(\mathbf{w})=\dfrac{|\tilde{m}_1-\tilde{m}_2|^2}{\tilde{s}_1^2+\tilde{s}_2^2}$

68

68

Fisher 线性分类的概念

中山大學



- 以“O”、 “Q”为例，比较PCA与LDA的差别。

69

Fisher 线性判别分析

中山大學

$$\tilde{s}_i^2 = \sum_{x \in D_i} \left(\mathbf{w}^t \mathbf{x} - \mathbf{w}^t \mathbf{m}_i \right)^2 = \mathbf{w}^t \mathbf{S}_i \mathbf{w}$$

$$\mathbf{S}_i = \sum_{x \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t$$

$$\tilde{s}_1^2 + \tilde{s}_2^2 = \mathbf{w}^t \mathbf{S}_W \mathbf{w}, \ \mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

$$\left| \tilde{m}_1 - \tilde{m}_2 \right|^2 = \left(\mathbf{w}^t \mathbf{m}_1 - \mathbf{w}^t \mathbf{m}_2 \right)^2 = \mathbf{w}^t \mathbf{S}_B \mathbf{w}$$

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$$

70

中山大學

$J(\mathbf{w}) = \dfrac{\mathbf{w}^t \mathbf{S}_B \mathbf{w}}{\mathbf{w}^t \mathbf{S}_W \mathbf{w}}$，称为generalized Rayleigh quotient

（广义瑞利商），

使$J(\mathbf{w})$达到最大的 $\mathbf{w}$必须满足：

广义本征值问题

$\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}$ (generalized eigenvalue problem)

$\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w} = \lambda \mathbf{w}$

$\mathbf{S}_B \mathbf{w} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t \mathbf{w}$ 总是位于 $(\mathbf{m}_1 - \mathbf{m}_2)$方向上

$\therefore \mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$ [忽略常数]

71

71

中山大學

对于正态分布的LDA

假设各类的协方差矩阵 $\mathbf{\Sigma}$相同

最佳判决边界的方程为：

$\mathbf{w}^t \mathbf{x} + w_0 = 0 \quad P31$

其中，$\mathbf{w} = \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2)$

如果用样本估计 μ_1, μ_2,和 $\mathbf{\Sigma}$,则

$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$

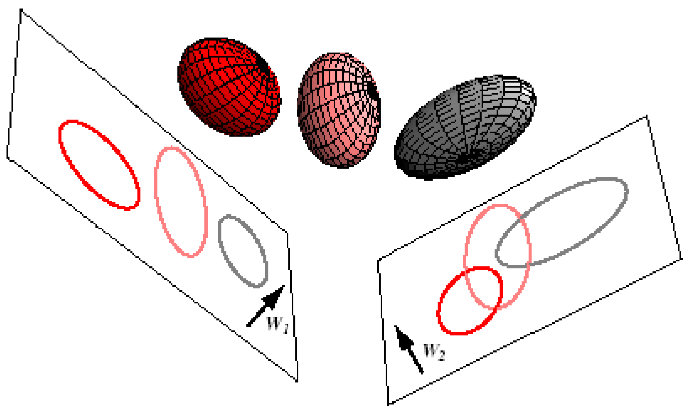
[这就是Fisher linear discriminant analysis与贝叶斯准则的关系]

72

72

3.8.3 多重判别分析—MDA

中山大學



73

73

Multiple Discriminant Analysis

中山大學

考虑c类的分类问题

将 d维空间投影到 $(c-1)$维子空间（d>c）

$$y_i = \mathbf{w}_i^t \mathbf{x}, \quad i = 1, \cdots, c-1 \Rightarrow \mathbf{y} = \mathbf{W}^t \mathbf{x}$$

$$\tilde{\mathbf{m}}_i = \frac{1}{n} \sum_{\mathbf{x} \in D_i} \mathbf{W}^t \mathbf{x}, \quad \tilde{\mathbf{m}} = \frac{1}{n} \sum_{i=1}^{c} n_i \tilde{\mathbf{m}}_i$$

$$\tilde{\mathbf{S}}_W = \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} \left(\mathbf{W}^t \mathbf{x} - \tilde{\mathbf{m}}_i \right) \left(\mathbf{W}^t \mathbf{x} - \tilde{\mathbf{m}}_i \right)^t = \mathbf{W}^t \mathbf{S}_W \mathbf{W}$$

$$\mathbf{S}_W = \sum_{i=1}^{c} \mathbf{S}_i, \quad \mathbf{S}_i = \sum_{\mathbf{x} \in D_i} \left(\mathbf{x} - \mathbf{m}_i \right) \left(\mathbf{x} - \mathbf{m}_i \right)^t, \quad \mathbf{m}_i = \frac{1}{n} \sum_{\mathbf{x} \in D_i} \mathbf{x}$$

74

74

$$\mathbf{m} = \frac{1}{n}\sum_{\mathbf{x}} \mathbf{x} = \frac{1}{n}\sum_{i=1}^{c} n_i \mathbf{m}_i$$

$$\mathbf{S}_T = \sum_{\mathbf{x}} (\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^t$$

$$= \sum_{i=1}^{c}\sum_{\mathbf{x}\in D_i} (\mathbf{x}-\mathbf{m}_i+\mathbf{m}_i-\mathbf{m})(\mathbf{x}-\mathbf{m}_i+\mathbf{m}_i-\mathbf{m})^t$$

$$= \sum_{i=1}^{c}\sum_{\mathbf{x}\in D_i} (\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t + \sum_{i=1}^{c}\sum_{\mathbf{x}\in D_i} (\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^t$$

$$= \mathbf{S}_W + \sum_{i=1}^{c} n_i (\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^t = \mathbf{S}_W + \mathbf{S}_B$$

$$\tilde{\mathbf{S}}_B = \sum_{i=1}^{c} n_i (\tilde{\mathbf{m}}_i-\tilde{\mathbf{m}})(\tilde{\mathbf{m}}_i-\tilde{\mathbf{m}})^t = \mathbf{W}^t\mathbf{S}_B\mathbf{W}$$

75

75

我们希望寻找 **W** 使得between-class scatter与 within-class scatter的比最大。

用散度矩阵行列式度量的情形：

$$\therefore \text{let } J(\mathbf{W}) = \frac{|\tilde{\mathbf{S}}_B|}{|\tilde{\mathbf{S}}_W|} = \frac{|\mathbf{W}^t\mathbf{S}_B\mathbf{W}|}{|\mathbf{W}^t\mathbf{S}_W\mathbf{W}|}$$

W 的列满足：

$$\mathbf{S}_B\mathbf{w}_i = \lambda_i \mathbf{S}_W \mathbf{w}_i$$

这是相对大特征值的广义特征向量。最优的 **W** 是不唯一的, 对坐标轴进行适当地旋转和伸缩，对准则函数和最后的分类器没有影响。

76

76

期望最大化 (EM)

中山大學

- 将最大似然估计推广到允许包含丢失特征样本来学习特定分布的参数问题;
- 完整的样本集 $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$
- $\mathbf{x}_k = \{ \mathbf{x}_{kg}, \mathbf{x}_{kb} \}$
- 把不同的特征分成两部分 D_g 和 D_b
  - D 是 D_g 和 D_b的并集
- 组成函数

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^i) = E_{D_b}\left[\ln p(D_g, D_b;\boldsymbol{\theta}) \mid D_g;\boldsymbol{\theta}^i\right]$$

左边是一个关于$\boldsymbol{\theta}$的函数,$\boldsymbol{\theta}^i$描述整个分布的真实参数。

也可以看作参数向量$\boldsymbol{\theta}^i$是当前对整个分布的最好估计,

而$\boldsymbol{\theta}$是在当前估计基础上进一步改善的估计。

右边表示关于丢失的特征求对数似然函数的期望。

77

77

中山大學

begin initialize $\boldsymbol{\theta}^0, T, i \leftarrow 0$

  do $i \leftarrow i + 1$

    E step: Compute $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^i)$

    M step: $\boldsymbol{\theta}^{i+1} \leftarrow \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\boldsymbol{\theta}^i)$

  until $Q(\boldsymbol{\theta}^{i+1};\boldsymbol{\theta}^i)-Q(\boldsymbol{\theta}^i;\boldsymbol{\theta}^{i-1}) \leq T$

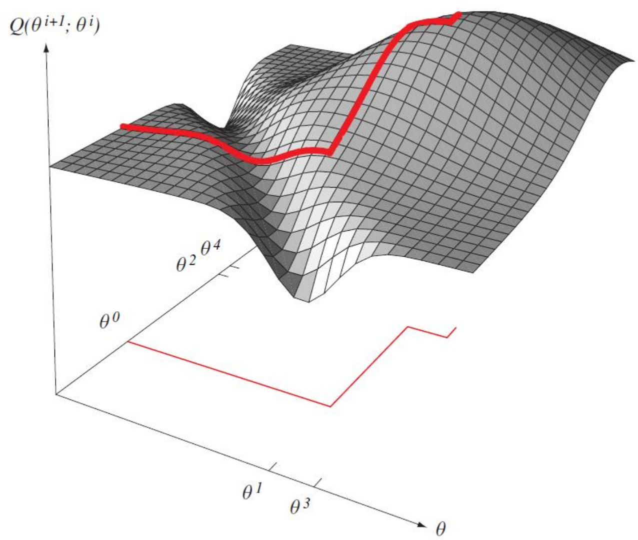
  return $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{i+1}$

end

78

78

Expectation-Maximization (EM)



79

79

Example: 2D 模型

$$D = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\} = \left\{ \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} * \\ 4 \end{pmatrix} \right\}$$

$$D_b = x_{41}$$

假设 2D 高斯模型具有对角斜方差阵

$$\boldsymbol{\theta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \sigma_1^2 \\ \sigma_2^2 \end{pmatrix}, \quad \boldsymbol{\theta}^0 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

80

80

中山大學

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^0) = E_{x_{41}}\left[\ln\ p(D_g, x_{41};\boldsymbol{\theta}) \mid D_g, \boldsymbol{\theta}^0\right]$$

$$= \int_{-\infty}^{\infty}\left[\sum_{k=1}^{3}\ln\ p(\mathbf{x}_k \mid \boldsymbol{\theta}) + \ln\ p(\mathbf{x}_4 \mid \boldsymbol{\theta})\right]\times$$

$$p(x_{41} \mid \boldsymbol{\theta}^0; x_{42} = 4)dx_{41}$$

$$= \sum_{k=1}^{3}\ln p(\mathbf{x}_k \mid \boldsymbol{\theta}) + \int_{-\infty}^{\infty}\ln p\left(\begin{pmatrix} x_{41} \\ 4 \end{pmatrix} \mid \boldsymbol{\theta}\right)\frac{p\left(\begin{pmatrix} x_{41} \\ 4 \end{pmatrix} \mid \boldsymbol{\theta}^0\right)}{K}dx_{41}$$

$$K = \int_{-\infty}^{\infty} p\left(\begin{pmatrix} x'_{41} \\ 4 \end{pmatrix} \mid \boldsymbol{\theta}^0\right)dx'_{41}$$

81

81

中山大學

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^0) = \sum_{k=1}^{3}\ln\ p(\mathbf{x}_k \mid \boldsymbol{\theta}) +$$

$$\frac{1}{K}\int_{-\infty}^{\infty}\ln\ p\left(\begin{pmatrix} x_{41} \\ 4 \end{pmatrix} \mid \boldsymbol{\theta}\right)\frac{1}{2\pi}\exp\left[-\frac{1}{2}(x_{41}^2 + 4^2)\right]dx_{41}$$

$$= \sum_{k=1}^{3}\ln\ p(\mathbf{x}_k \mid \boldsymbol{\theta}) - \frac{1+\mu_1^2}{2\sigma_1^2} - \frac{(4-\mu_2)^2}{2\sigma_2^2} - \ln\left(2\pi\sigma_1\sigma_2\right)$$

$$\boldsymbol{\theta}^1 = \begin{pmatrix} 0.75 \\ 2.0 \\ 0.938 \\ 2.0 \end{pmatrix}$$

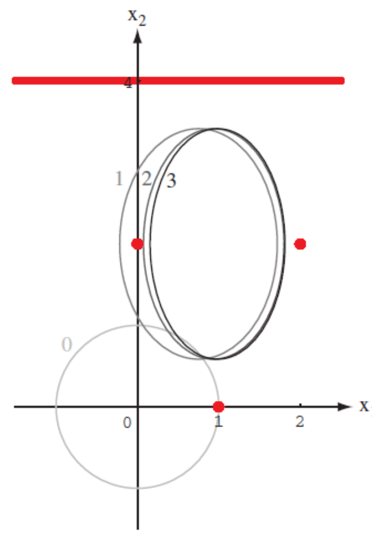
82

82

3 迭代后, 该算法在下列值收敛:

$$\boldsymbol{\mu} = \begin{pmatrix} 1.0 \\ 2.0 \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.667 & 0 \\ 0 & 2.0 \end{pmatrix}$$



83

83

广义期望最大化 (GEM)

- 代替最大化 $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^i)$, 我们在M步只需要找 $\boldsymbol{\theta}^{i+1}$ 使得

  $$Q(\boldsymbol{\theta}^{i+1};\boldsymbol{\theta}^i)>Q(\boldsymbol{\theta};\boldsymbol{\theta}^i)$$

  也能确保收敛。
- 收敛将没有那么快。
- 让用户自由选取计算更加简单的途径。
  - 有一种版本的GEM算法, 每次叠代时，都计算未知特征的最大似然函数，然后依此重新计算$\boldsymbol{\theta}$。

84

84

隐马尔可夫模型—
Hidden Markov Model (HMM)

中山大學

- 前面各章节,用一个n维特征矢量确定一个对象的状态,并基于这个状态进行统计判决;
- 本节,用一个时间的(矢量)序列或空间的(矢量)阵列来描述对象的整体状态,并基于这个整体状态进行统计判决;

- 用于处理序列判决问题
  - 应用, 在语音和手势识别方面有用。
- 在 t 时刻发生的事件要受到t-1时刻发生事件的直接影响。

85

85

First Order Markov Models
一阶马尔可夫模型

中山大學

假 设 在 t时 刻 的 状 态 记 为 ω(t) .

有一个时间长度为T的状态序列:

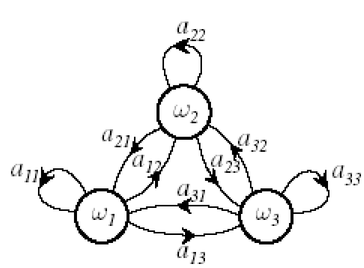
状态序列 $\boldsymbol{\omega}^T = \{\omega(1), \omega(2), \cdots, \omega(T)\}$

$e.g., \boldsymbol{\omega}^6 = \{\omega_1, \omega_3, \omega_2, \omega_2, \omega_1, \omega_3\},$

$P(\boldsymbol{\omega}^6 \mid \boldsymbol{\theta}) = a_{13}a_{32}a_{22}a_{21}a_{13}$

其中记$P(\omega_j(t+1) \mid \omega_i(t)) = a_{ij}$为转移概率.

特定序列的概率为连续的转移概率相乘。



86

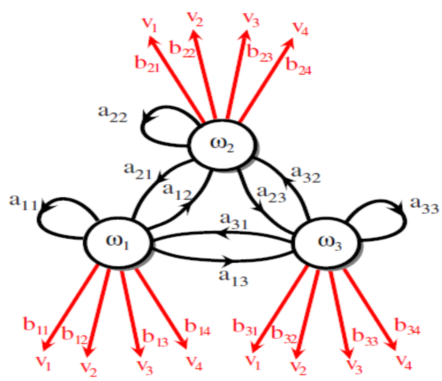
86

First Order Hidden Markov Models
一阶隐马尔可夫模型

假定在任一时刻状态$\omega(t)$都以某种概率产生

可见状态序列 $\mathbf{V}^T=\{v(1),v(2),\cdots,v(T)\}$中的一个。

$e.g., \mathbf{V}^6=\{v_4,v_1,v_1,v_4,v_2,v_3\},\quad P(v_k(t)\,|\,\omega_j(t))=b_{jk}$



Hidden Markov Model 概率

最终或吸收状态 $\omega_0 : a_{00}=1$

转移概率: $a_{ij}=P(\omega_j(t+1)\,|\,\omega_i(t))$

激发可见状态的概率:

$$b_{jk}=P(v_k(t)\,|\,\omega_j(t))$$

$$\sum_j a_{ij}=1,\quad \sum_k b_{jk}=1$$

中山大學

一阶隐马尔可夫模型的例子：

下面给出一个天气的例子。设一天的天气可能是晴天、多云、雨天，其只随机（但有统计性）地依赖于已过去的一天的天气，天气状态的转移概率 a_{ij} 如下所示：

$$\boldsymbol{A}=(a_{ij})=\begin{array}{c|ccc} & \text{晴天} & \text{多云} & \text{雨天} \\ \hline \text{晴天} & 0.50 & 0.375 & 0.125 \\ \text{多云} & 0.25 & 0.125 & 0.625 \\ \text{雨天} & 0.25 & 0.375 & 0.375 \end{array}$$

可以看出，晴天后是雨天的概率是0.125，雨天后是多云的概率是0.375，等等。一片海藻的湿润情况是天气的反映，它们间的概率如下所示：

$$\boldsymbol{B}=(b_{jk})=\begin{array}{c|cccc} & \text{干燥} & \text{稍干} & \text{潮湿} & \text{浸水} \\ \hline \text{晴天} & 0.60 & 0.20 & 0.15 & 0.05 \\ \text{多云} & 0.25 & 0.25 & 0.25 & 0.25 \\ \text{雨天} & 0.05 & 0.10 & 0.35 & 0.50 \end{array}$$

晴天时海藻干燥的概率是0.6，多云天气时海藻潮湿的概率是0.25，等等。若以每天作为观测时刻，假设只能观测海藻的湿润情况，这种两层统计系统就是一个隐马尔可夫系统。

89

89

Hidden Markov Model 的计算

中山大學

- 估值问题
  - 利用给定的 a_{ij} 和 b_{jk}, 计算某个特定观察序列 $\mathbf{V}^T$的概率 $P(\mathbf{V}^T|\boldsymbol{\theta})$。
- 解码问题
  - 给定特定观察序列 $\mathbf{V}^T$, 决定最有可能产生 $\mathbf{V}^T$的隐状态序列ω^T。
- 学习问题
  - 已知HMM的大致结构（如隐状态和可见状态的数目），但 a_{ij} 和 b_{jk}未知，如何从一组可见符号的训练集中，决定这些参数。
- 运用HMM模型识别
  - 利用各类的可见序列样本进行学习,产生代表每类的HMM参考模型; 对待识别可见序列,通过估值方法进行识别。

90

90

Evaluation(估值问题)

$$P(\mathbf{V}^T)=\sum_{r=1}^{r_{\max}}P(\mathbf{V}^T\mid\boldsymbol{\omega}_r^T)P(\boldsymbol{\omega}_r^T),\quad 其中 r_{\max}=c^T$$

$$P(\boldsymbol{\omega}_r^T)=\prod_{t=1}^{T}P(\omega(t)\mid\omega(t-1))$$

$$P(\mathbf{V}^T\mid\boldsymbol{\omega}_r^T)=\prod_{t=1}^{T}P(v(t)\mid\omega(t))$$

$$P(\mathbf{V}^T)=\sum_{r=1}^{r_{\max}}\prod_{t=1}^{T}P(v(t)\mid\omega(t))P(\omega(t)\mid\omega(t-1))$$

计算复杂度O(c^TT)

91

91

HMM Forward(前向算法)

$$P(\mathbf{V}^T)=\sum_{r=1}^{r_{\max}}\prod_{t=1}^{T}P(v(t)\mid\omega(t))P(\omega(t)\mid\omega(t-1))$$

$$\alpha_j(t)=P(\omega_j(t),\mathbf{V}^t)$$

$$=P(v_k\mid\omega_j(t))\sum_{i=1}^{c}P(\omega_j(t)\mid\omega_i(t-1))P(\omega_i(t-1),\mathbf{V}^{t-1})$$

$$=b_{jkv(t)}\sum_{i=1}^{c}a_{ij}\alpha_i(t-1)$$

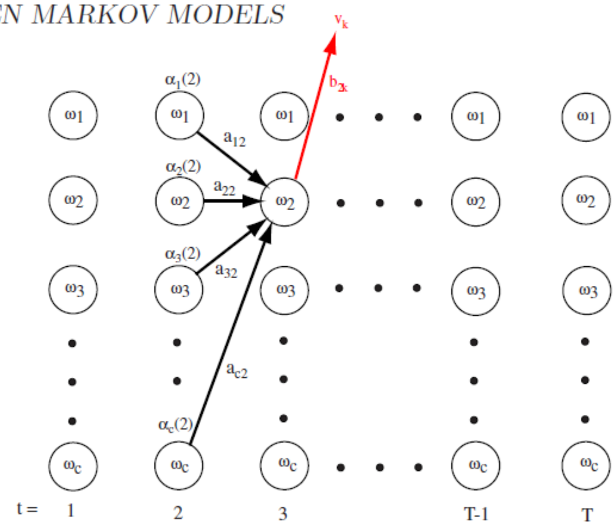
$$\alpha_j(0)=\begin{cases}0, & j\neq \text{initial state}\\ 1, & j=\text{initial state}\end{cases}$$

92

92

HMM Forward

*HIDDEN MARKOV MODELS



HMM Forward

initialize $t \leftarrow 0, a_{ij}, b_{jk}, \mathbf{V}^T, \alpha_j(0) = 1$

  for $t \leftarrow t+1,\ j = 0, 1, \cdots, c$

    $\alpha_j(t) \leftarrow b_{jkv(t)} \sum_{i=1}^{c} \alpha_i(t-1) a_{ij}$

  until $t = T$

return $P(\mathbf{V}^T) = \alpha_0(T)$ for final state

end

HMM Backward

中山大學

$$P(\mathbf{V}^T) = \sum_{r=1}^{r_{\max}} \prod_{t=1}^{T} P(v(t) \mid \omega(t)) P(\omega(t) \mid \omega(t-1))$$

$$\beta_i(t) = P(\omega_i(t), \mathbf{V}^{T-t})$$

$$= \sum_{j=1}^{c} P(v_k \mid \omega_j(t+1)) P(\omega_j(t+1) \mid \omega_i(t)) P(\omega_j(t+1), \mathbf{V}^{T-(t+1)})$$

$$= \sum_{j=1}^{c} b_{jkv(t+1)} a_{ij} \beta_j(t+1)$$

$$\beta_i(T) = \begin{cases} 0, & i \neq 0 \\ 1, & i = 0 \end{cases}$$

95

95

HMM Backward

中山大學

$$\text{initialize } t \leftarrow T, a_{ij}, b_{jk}, \mathbf{V}^T, \beta_j(T)$$

$$\quad \text{for } t \leftarrow t-1,\ i = 0, 1, \cdots, c$$

$$\qquad \beta_i(t) \leftarrow \sum\nolimits_{j=0}^{c} \beta_j(t+1) a_{ij} b_{jkv(t+1)}$$

$$\quad \text{until } t = 0$$

$$\text{return } P(\mathbf{V}^T) = \beta_0(0) \text{ for initial state}$$

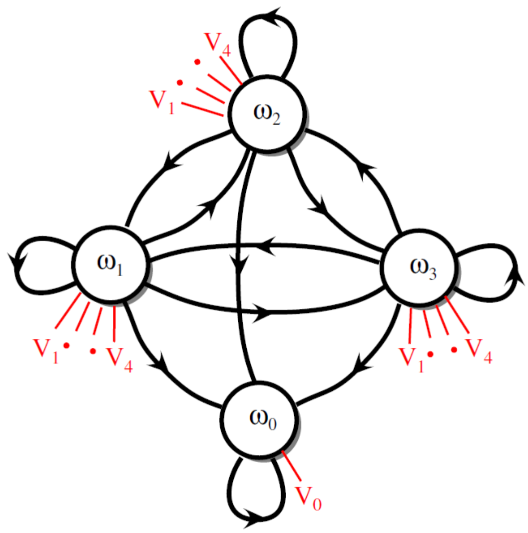
$$\text{end}$$

96

96

中山大學

Example 3: Hidden Markov Model



97

97

中山大學

Example 3: Hidden Markov Model

$$a_{ij} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.2 & 0.3 & 0.1 & 0.4 \\ 0.2 & 0.5 & 0.2 & 0.1 \\ 0.8 & 0.1 & 0.0 & 0.1 \end{pmatrix}$$

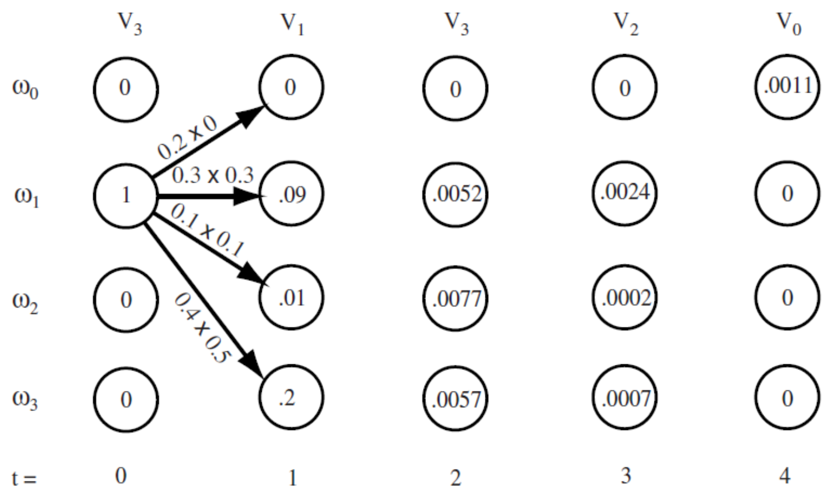
$$b_{jk} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0.3 & 0.4 & 0.1 & 0.2 \\ 0 & 0.1 & 0.1 & 0.7 & 0.1 \\ 0 & 0.5 & 0.2 & 0.1 & 0.2 \end{pmatrix}$$

98

98

中山大學

Example 3: Hidden Markov Model

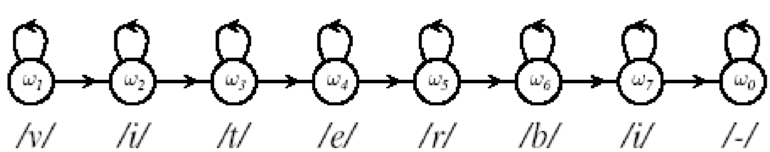


99

99

中山大學

Left-to-Right Models for Speech

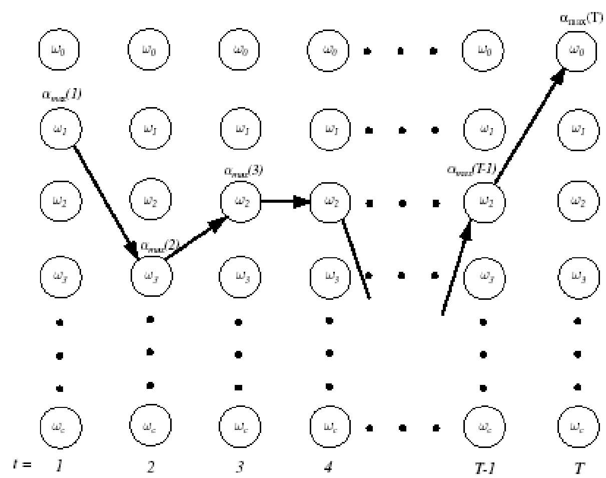


$$P(\boldsymbol{\theta} \mid \mathbf{V}^T) = \frac{P(\mathbf{V}^T \mid \boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathbf{V}^T)}$$

100

100

HMM Decoding（解码）



101

101

Problem of Local Optimization

- 局部最优：对于每个t时刻，都寻找从前状态转移过来，并且产生可见状态v_k的概率最大的概率。
- 算法本身不保证整个路径是合法的。

102

102

中山大學

HMM Decoding

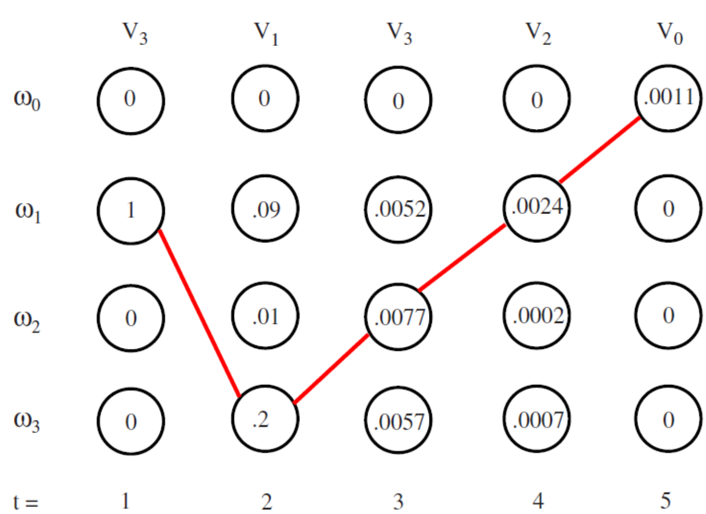
$$
\begin{aligned}
&\text{initialize } t \leftarrow 0, Path = \{\} \\
&\quad \text{for } t \leftarrow t+1, \\
&\quad\quad j \leftarrow 0 \\
&\quad\quad \text{for } j \leftarrow j+1 \\
&\quad\quad\quad \alpha_j(t) \leftarrow b_{jkv(t)} \sum_{i=1}^{c} \alpha_i(t-1) a_{ij} \\
&\quad\quad \text{until } j = c \\
&\quad\quad j' \leftarrow \arg\max_j \alpha_j(t) \\
&\quad\quad \text{Append } \omega_{j'} \text{ to } Path \\
&\quad \text{until } t = T \\
&\text{return } Path \\
&\text{end}
\end{aligned}
$$

103

103

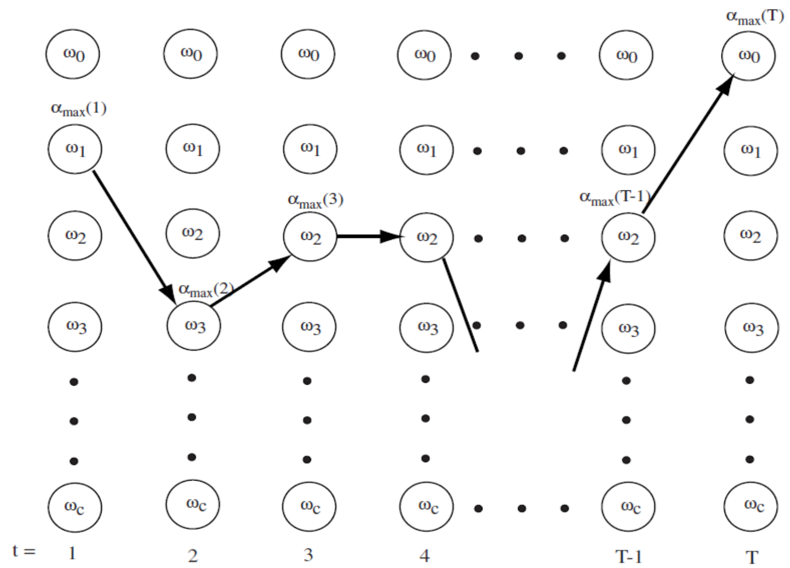
中山大學

Example 4: HMM Decoding



104

104



105

105

Forward-Backward Algorithm

- 从一组训练样本中,确定模型参数 a_{ij} 和 b_{jk}。
- “前向-后向算法”是“广义期望最大化算法”的具体实现
- 通过递归方法更新权重，以得到能够更好地解释训练样本序列的模型参数。

106

106

Probability of Transition

$$\gamma_{ij}(t) = P(\omega_i(t-1), \omega_j(t) \mid \mathbf{V}^T, \boldsymbol{\theta})$$

$$= \frac{P(\omega_i(t-1), \omega_j(t), \mathbf{V}^T \mid \boldsymbol{\theta})}{P(\mathbf{V}^T \mid \boldsymbol{\theta})}$$

$$= \frac{\alpha_i(t-1) a_{ij} b_{jk} \beta_j(t)}{P(\mathbf{V}^T \mid \boldsymbol{\theta})}$$

107

107

Improved Estimate for a_{ij}

在任何时刻，序列中从状态$\omega_i(t-1)$ 到 $\omega_j(t)$的转换估计值：

$$\sum_{t=1}^{T} \gamma_{ij}(t)$$

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

a_{ij}的估计值为:

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T} \gamma_{ij}(t)}{\sum_{t=1}^{T} \sum_k \gamma_{ik}(t)}$$

108

108

同理，可以获得 b_{jk}的估计值

$$\hat{b}_{jk} = \frac{\sum_{t=1,\ v(t)=v_k}^{T} \sum_i \gamma_{ij}(t)}{\sum_{t=1}^{T} \sum_i \gamma_{ij}(t)}$$

109

109

Forward-Backward Algorithm (Baum-Welch Algorithm)

initialize a_{ij}, b_{jk}, training sequence $\mathbf{V}^T$, threshold $\theta, z \leftarrow 0$

do $z \leftarrow z+1$

compute all $\hat{a}_{ij}(z)$ from all $a_{ij}(z-1)$ and $b_{jk}(z-1)$

compute all $\hat{b}_{jk}(z)$ from all $a_{ij}(z-1)$ and $b_{jk}(z-1)$

$a_{ij}(z) \leftarrow \hat{a}_{ij}(z)$

$b_{jk}(z) \leftarrow \hat{b}_{jk}(z)$

until $\max_{i,j,k}\left[a_{ij}(z) - a_{ij}(z-1), b_{jk}(z) - b_{jk}(z-1)\right] < \theta$

return $a_{ij} \leftarrow a_{ij}(z); b_{jk} \leftarrow b_{jk}(z)$

end

110

110