Chapter 3: 最大似然估计和贝叶斯参数估计

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;

• 3.1 引 言

- □ 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：1）样本对于类条件估计太少了；2） 特征空间维数太大了，计算复杂度太高。

- □ 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。

- 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。

- 贝叶斯方法把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，使得对于每个新样本，后验概率密度函数在待估参数的真实值附近形成最大尖峰。

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

• 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；
- 比其他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）$P(x \mid \omega_j)$ 形式已知但参数未知，例如$P(x \mid \omega_j) \sim N(\mu_j, \Sigma_j)$；

3）记 $P(x \mid \omega_j) \equiv P(x \mid \omega_j, \theta_j)$，其中

$$\theta_j = (\mu_j, \Sigma_j)$$

- 使用训练样本提供的信息估计

  θ = (θ_1, θ_2, …, θ_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,…, 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”

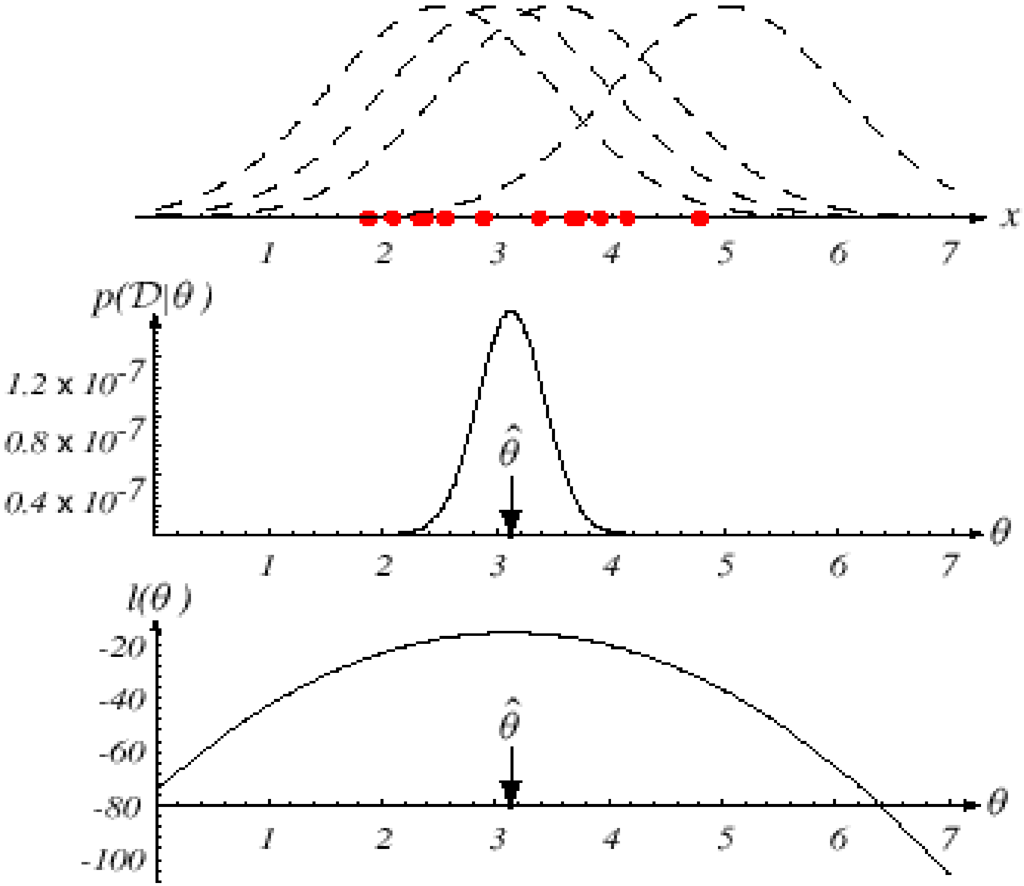


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood—that is, the log-likelihood $l(\theta)$, shown at the bottom. Note that even though they look similar, the likelihood $p(\mathcal{D}|\theta)$ is shown as a function of θ whereas the conditional density $p(x|\theta)$ is shown as a function of x. Furthermore, as a function of θ, the likelihood $p(\mathcal{D}|\theta)$ is not a probability density function and its area has no significance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- 最优估计
  - □ 令θ = (θ$_1$, θ$_2$, …, θ$_p$)^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(θ) = ln P(D | θ)
  - □ 新问题陈述:

    求解 θ 为使对数似然最大的值

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

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.

3.2.3 高斯情况： μ未知

- □ P(x_k | μ) ~ N(μ, Σ)
  (样本从一组多变量正态分布中提取)

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

和 $\nabla_\mu \ln P(x_k \mid \mu) = \sum^{-1}(x_k - \mu)$

θ = μ，因此:

- μ的最大似然估计必须满足:

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

- 乘 Σ 并且重新排序, 我们得到:

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

即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!

3.2.3 高斯情况：μ和 Σ均未知

- 未知 μ 和 σ，对于单样本x_k

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

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

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

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

对于全部样本，最后得到:

$$\begin{cases} \sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \theta_1) = 0 & (1) \\ -\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 & (2) \end{cases}$$

联合公式 (1) 和 (2), 得到如下结果:

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

3.2.4 偏差估计

- □ σ²的最大似然估计是有偏的 （渐进无偏估计）

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

- □ Σ的一个基本的无偏估计是:

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

模型错误会怎么样?

达不到最优!

• 3.3贝叶斯估计

□ 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：

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

- 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 $P(x \mid D)$

3.3.2 参数的分布

- 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；

3.3.2 参数的分布

- 注意到

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

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

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。**

3.4 贝叶斯参数估计: 高斯过程

- 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)

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

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

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

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

复制密度

$$p(\mu \mid D) \sim N(\mu_n, \sigma_n^2) \text{ [reproducing density]}$$

[称$p(\mu)$: conjugate prior]

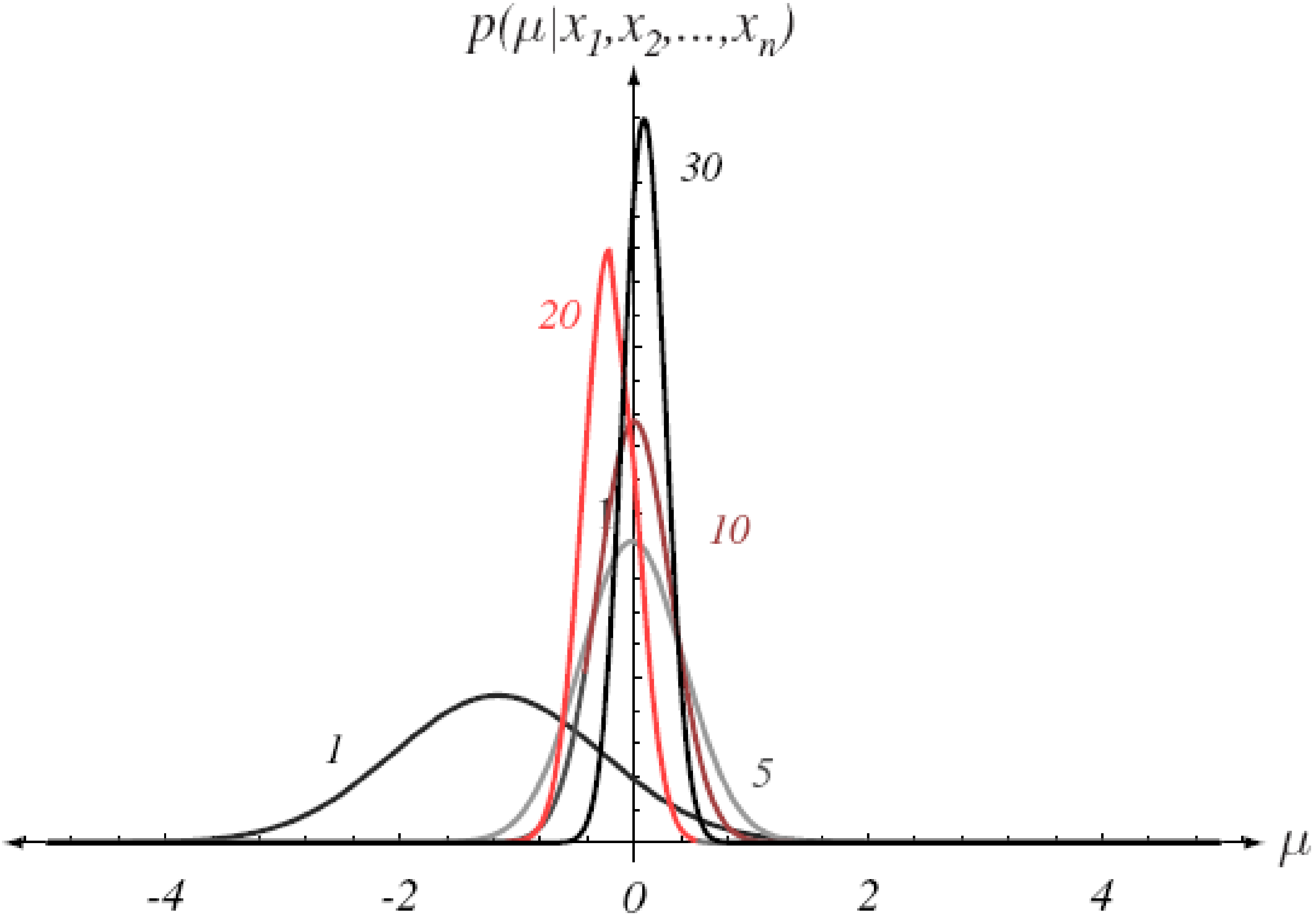
$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}, \quad \frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2}$$

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

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

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

贝叶斯学习



结论:

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 $\hat{\mu}_\mathrm{n}$,否则$\mu_\mathrm{n} = \mu_0$ 。

□ 单变量情形的 p(x|D)

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

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

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

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

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

多变量情形:

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

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

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

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

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

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

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

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

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

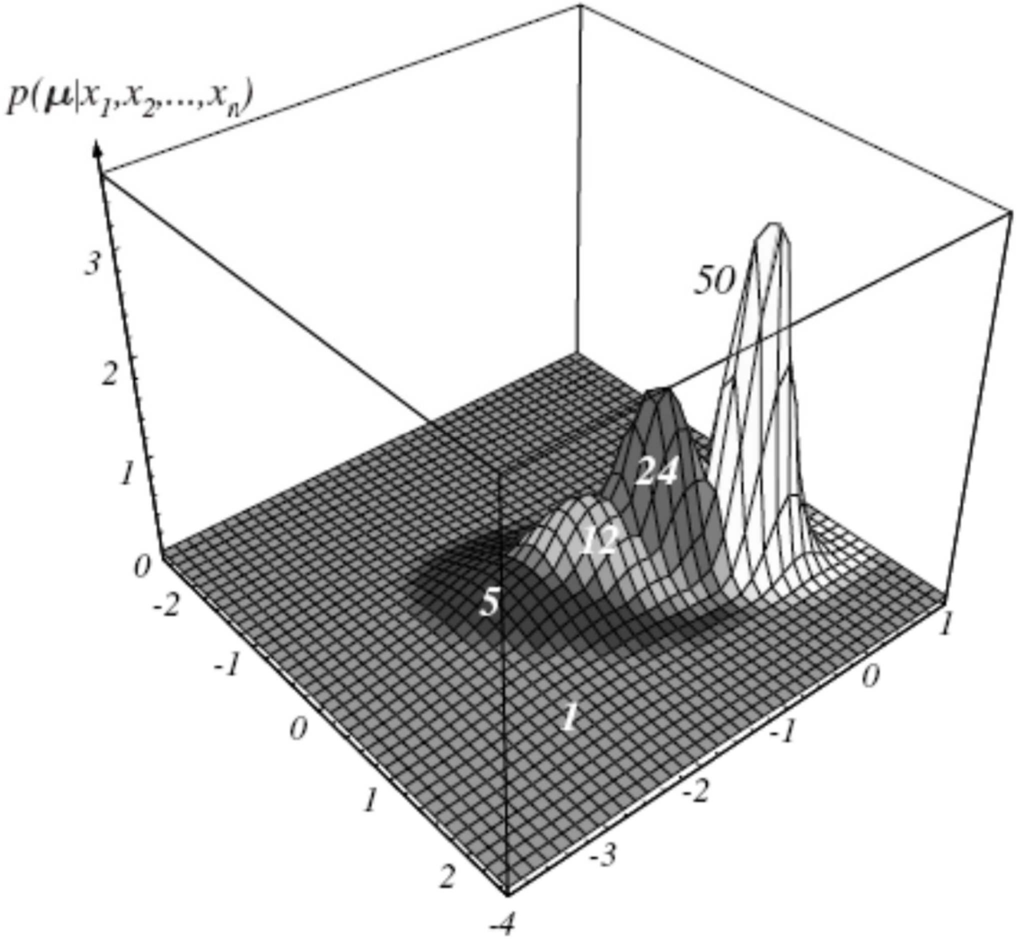
利用 $p(\mathbf{x}\mid D)=\int p(\mathbf{x}\mid\boldsymbol{\mu})p(\boldsymbol{\mu}\mid D)d\boldsymbol{\mu}$

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

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

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

多变量学习



3.5 贝叶斯参数估计：一般理论

□ 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) 的形式未知，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:

计算先验密度p(θ | D)，然后 推导出 p(x | D)。

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

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

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

问题:

p(x | D)是否能收敛到p(x)，计算复杂度如何？

递归贝叶斯学习

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

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

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

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

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

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

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。

例1：递归贝叶斯学习

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

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

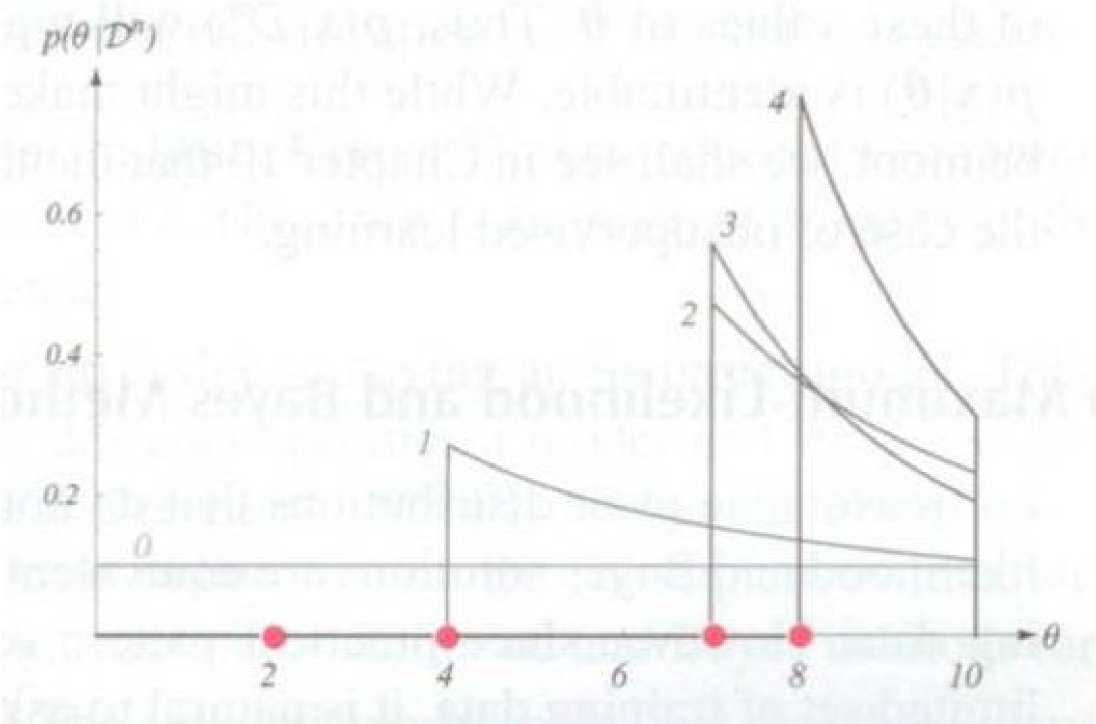
$$p(D^0\mid\theta)=p(\theta)\sim U(0,10)$$

$$p(\theta\mid D^1)\propto p(x_1\mid\theta)p(\theta\mid D^0)=\begin{cases}1/\theta & \text{对于 } 4\le\theta\le10\\ 0 & \text{其他}\end{cases}$$

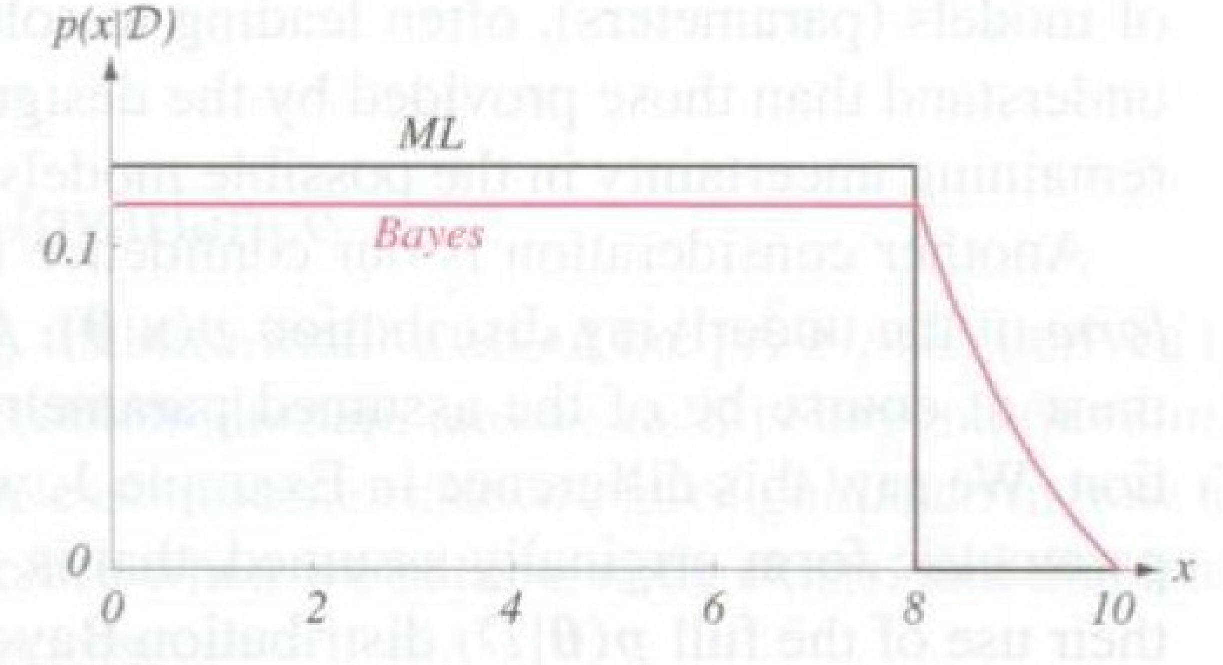
$$p(\theta\mid D^2)\propto p(x_2\mid\theta)p(\theta\mid D^1)=\begin{cases}1/\theta^2 & \text{对于 } 7\le\theta\le10\\ 0 & \text{其他}\end{cases}$$

$$p(\theta\mid D^n)\propto 1/\theta^n\quad \text{对于 } \max_x\left[D^n\right]\le\theta\le10$$

例1：递归贝叶斯学习



例1: Bayes vs. ML



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

唯一性问题

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

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

$p(\boldsymbol{\theta}|D^n)$ 将在 $\boldsymbol{\theta}$ 附近产生峰值，这时不管$p(\mathbf{x}|\boldsymbol{\theta})$ 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。

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

分类误差

- 贝叶斯错误或不可分错误，例如 $P(x \mid \omega_i)$；
- 模型错误；
- 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

- 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的所有所有信息，即 $p(D|\mathbf{s}, \theta)$ 与 θ无关；
- 如果$\boldsymbol{\theta}$是随机变量，则

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

反过来也成立。

因式分解定理：

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

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

其中 $g(.,.)$ 和$h(.)$是两个函数。

例子：多维高斯分布

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

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

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

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

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

证明：必要性

假设$\mathbf{s}$ 是关于θ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，只有一个s与之对应。

充分性：

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

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

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

上式不依赖于θ;

因此 **s** 是关于 θ的充分统计量。

核密度（Kernel density）

- 把 $P(D|\theta)$ 分解成 $g(\mathbf{s},\theta)h(D)$ 不是唯一的：
  - □ 如果$f(\mathbf{s})$ 是一个函数, $g'(\mathbf{s},\theta)=f(\mathbf{s})g(\mathbf{s},\theta)$ 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：

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

例子：多维高斯分布

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

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

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

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

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

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

核密度与参数估计

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

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

  - □ 如果我们对q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$p(\mathbf{x}|\theta)$ 可辩识时, $g(\mathbf{s},\theta)$ 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。

充分统计量与指数族函数

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

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

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

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

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

中山大學

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

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

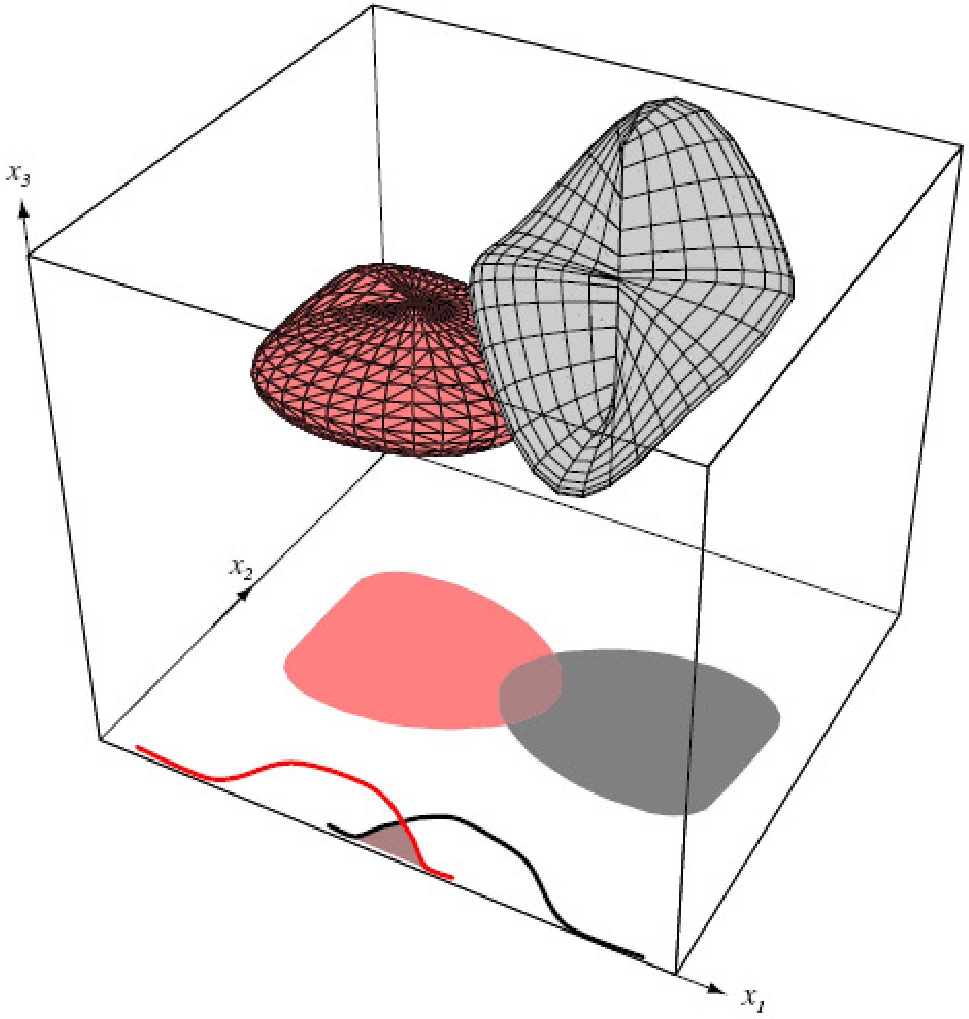
- 如果特征是独立的，则有:

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

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

- 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;
- 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。

可分性与特征维数



计算复杂度

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

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

求解$d\times d$ 矩阵的逆： $O(d^3)$

求解 $d\times d$ 行列式: $O(d^3)$

$n > d$

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

分类的计算复杂度

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 $\max_i g_i(\mathbf{x})$: $O(c)$

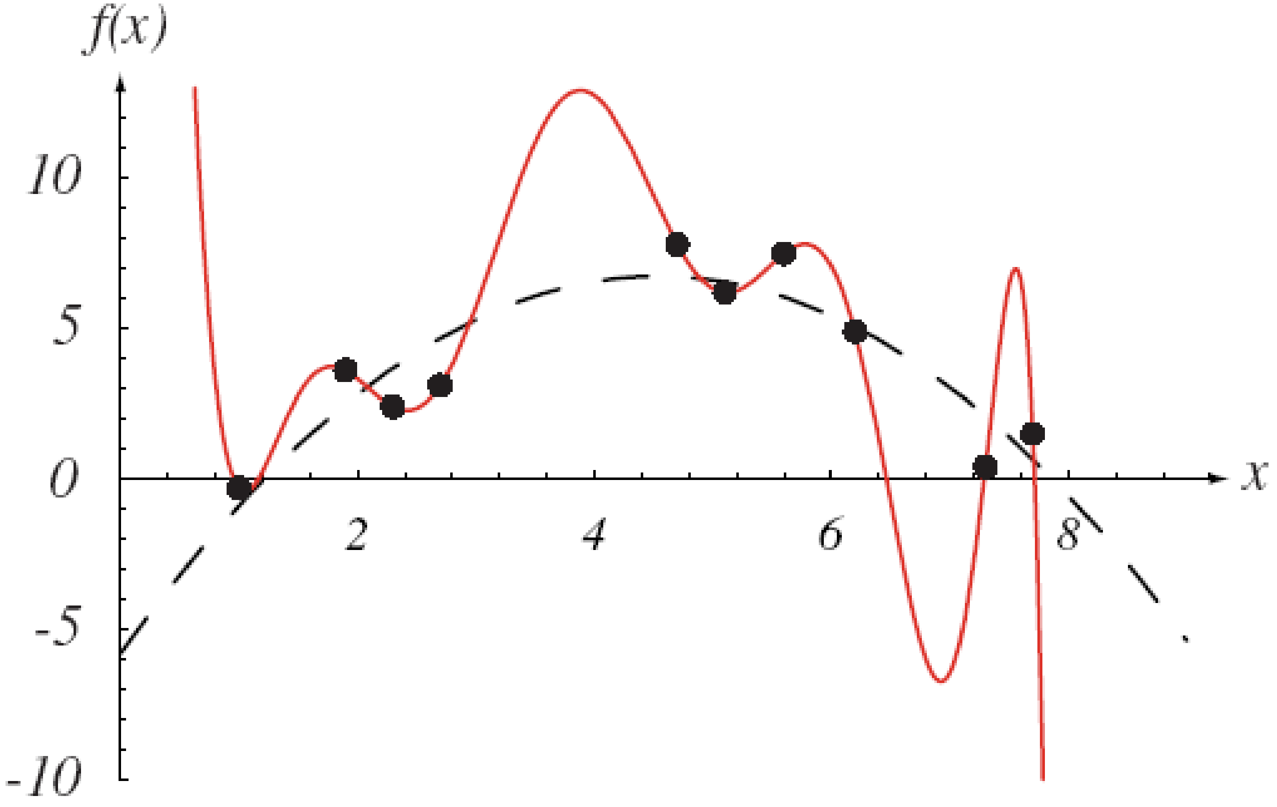
整个分类问题的复杂度: $O(d^2)$

- 分类阶段比学习阶段简单。

训练样本不足时的方法

- 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 Σ 更好的估计;
  - □ 如果有合理的先验估计 Σ_0, 则可以用如下的伪贝叶斯估计 $\lambda\Sigma_0+(1-\lambda)\widehat{\Sigma}$;
  - □ 设法将Σ_0对角化: 阈值化或假设特征之间统计独立;

过拟合的概念



正确的拟合思想是；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(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

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

Best Representative Point

Given $\mathbf{x}_1, \cdots, \mathbf{x}_n$, find $\mathbf{x}_0$ such that

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

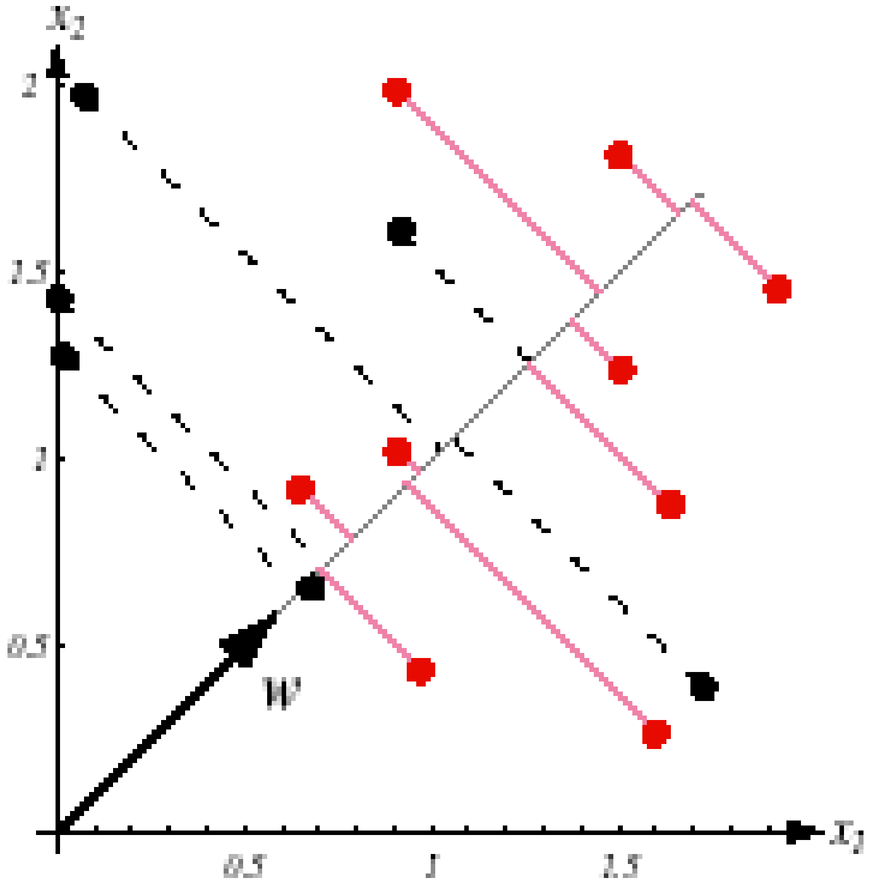
$$\mathbf{m} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k$$

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

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

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

Projection Along a Line



Best Projection to a Line Through the Sample Mean

Line $\mathbf{x} = \mathbf{m} + a\mathbf{e}$

Represent $\mathbf{x}_k$ by $\mathbf{m} + a_k\mathbf{e}$ with error

To minimize

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

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

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

Best Representative Direction

Find **e** to minimize

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

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

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

Maximize $\mathbf{e}^t\mathbf{S}\mathbf{e}$ subject to $\|\mathbf{e}\|^2 = 1$

Lagrange method : maximize $u = \mathbf{e}^t\mathbf{S}\mathbf{e} - \lambda(\mathbf{e}^t\mathbf{e} - 1)$

$\nabla_{\mathbf{e}} u = 0 \Rightarrow \mathbf{S}\mathbf{e} = \lambda\mathbf{e}$

Principal Component Analysis (PCA)

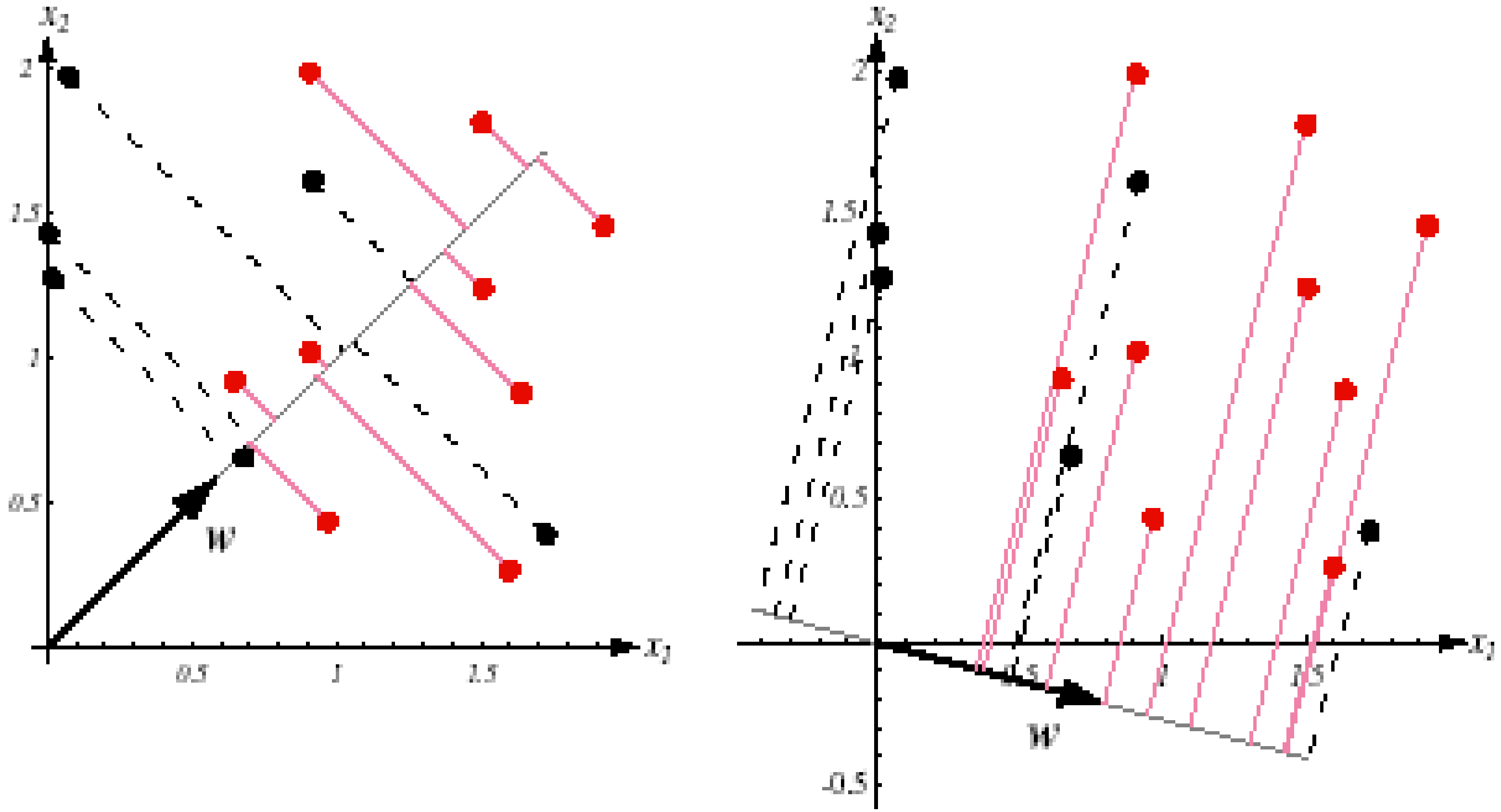
$$\text{Projection space}: \mathbf{x} = \mathbf{m} + \sum_{i=1}^{d'} a_i \mathbf{e}_i$$

Find $\mathbf{e}_i, \quad i = 1, \cdots, d'$ to minimize

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

$\Rightarrow \mathbf{e}_1, \cdots, \mathbf{e}_{d'}$ are the d' eigenvectors of $\mathbf{S}$ having the d' largest eigenvalues

Concept of Fisher Linear Discriminant



Fisher Linear Discriminant Analysis

Find $\mathbf{w}$ to get maximal separation on $\mathbf{y} = \mathbf{w}^t\mathbf{x}$

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

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

Within - class scatter : $\tilde{s}_1^2 + \tilde{s}_2^2$

To maximize $J(\mathbf{w}) = \dfrac{\left|\tilde{m}_1 - \tilde{m}_2\right|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$

Fisher Linear Discriminant Analysis

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

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

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

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

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

Fisher Linear Discriminant Analysis

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

is maximized when

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

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

$\mathbf{S}_B \mathbf{w} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t \mathbf{w}$ is always in the direction of $(\mathbf{m}_1 - \mathbf{m}_2)$

$$\therefore \mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \text{ [ignoring scales]}$$

Fisher Linear Discriminant Analysis for Multivariate Normal

Assume same covariance matrix $\mathbf{\Sigma}$

optimal decision boundary

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

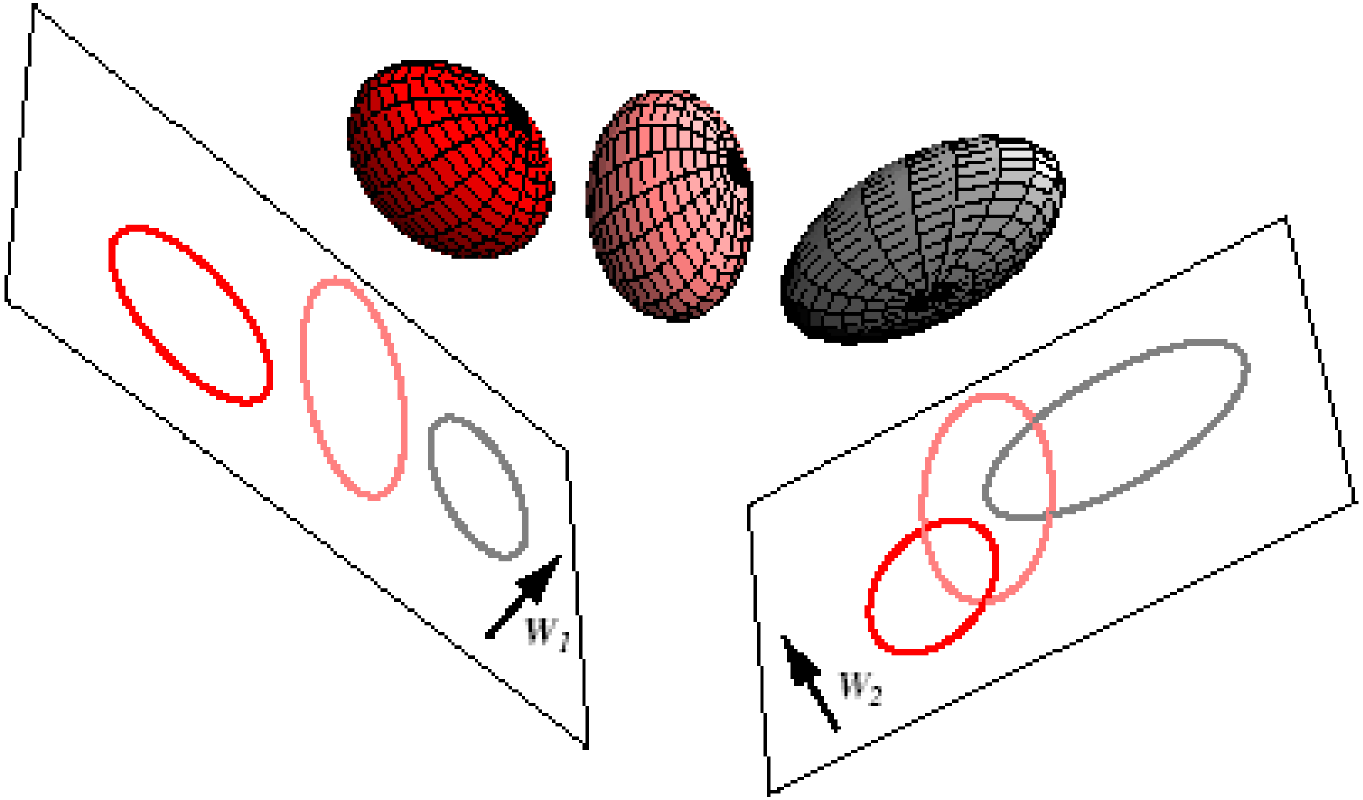
$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2)$$

With estimation for μ_1, μ_2, and $\mathbf{\Sigma}$,

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

[solution to Fisher linear discriminant analysis]

Concept of Multidimensional Discriminant Analysis



Multiple Discriminant Analysis

Consider c - class problem

Projection from d - dimensional space to (c-1) - dimensional subspace

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

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

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

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

Multiple Discriminant Analysis

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

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

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

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

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

Multiple Discriminant Analysis

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

Seek a transformation **W** to maximize the ratio of the between-class scatter to the withon-class scatter

A simple scalar measure of scatter is the determinant of the scatter matrix (equivalent to the product of variances in the principal directions)

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

Multiple Discriminant Analysis

Columns of optimal $\mathbf{W}$ satisfies

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

and is the generalized eigenvector related to the largest eigenvalue

optimal $\mathbf{W}$ is not unique, since it can be multiplied by rotation or scaling matrices, etc.

Expectation-Maximization (EM)

- Finding the maximum-likelihood estimate of the parameters of an underlying distribution
  - from a given data set when the data is incomplete or has missing values
- Two main applications
  - When the data indeed has missing values
  - When optimizing the likelihood function is analytically intractable but when the likelihood function can be simplified by assuming the existence of and values for additional but missing (or hidden) parameters

Expectation-Maximization (EM)

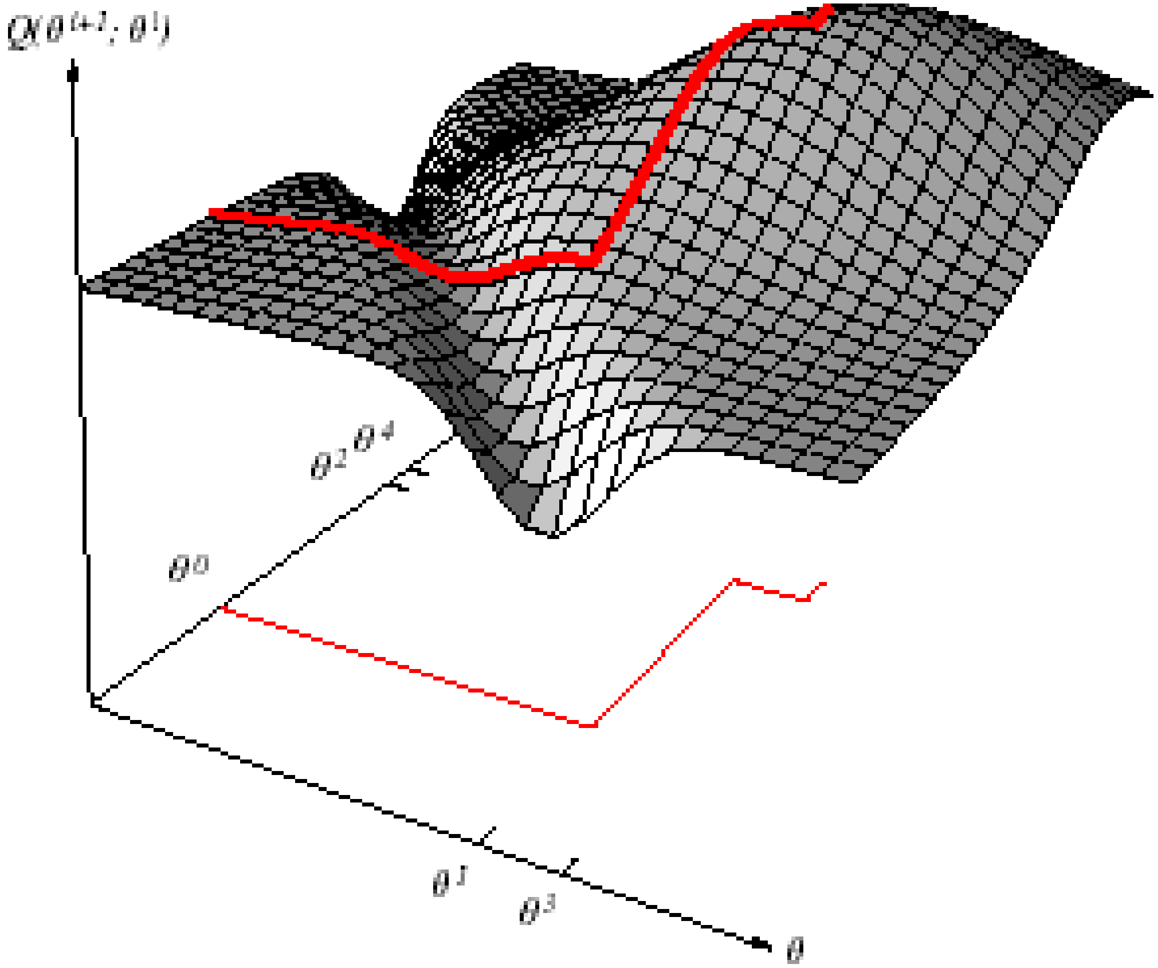
- Full sample $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_\mathrm{n}\}$
- $\mathbf{x}_\mathrm{k} = \{\, \mathbf{x}_{kg}, \mathbf{x}_{kb} \,\}$
- Separate individual features into D_g and D_b
  - D is the union of D_g and D_b
- Form the function

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

Expectation-Maximization (EM)

```
begin initialize $\theta^0$, $T$, $i \leftarrow 0$
    do $i \leftarrow i + 1$
        E step: Compute $Q(\theta; \theta^i)$
        M step: $\theta^{i+1} \leftarrow \arg\max_{\theta} \leq Q(\theta,\theta^i)$
    until $Q(\theta^{i+1};\theta^i)$-$Q(\theta^i;\theta^{i-1})$      $T$
    return $\theta \leftarrow \theta^{i+1}$
end
```

Expectation-Maximization (EM)



Example: 2D Model

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

$$D_b = x_{41}$$

Assume 2D Gaussian model with diagonal covariance matrix

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

Example: 2D Model

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

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

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

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

$$K=\int_{-\infty}^{\infty}p\left(\begin{pmatrix}x_{41}^{'}\\4\end{pmatrix}\,|\,\boldsymbol{\theta}^0\right)dx_{41}^{'}$$

Example: 2D Model

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

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

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

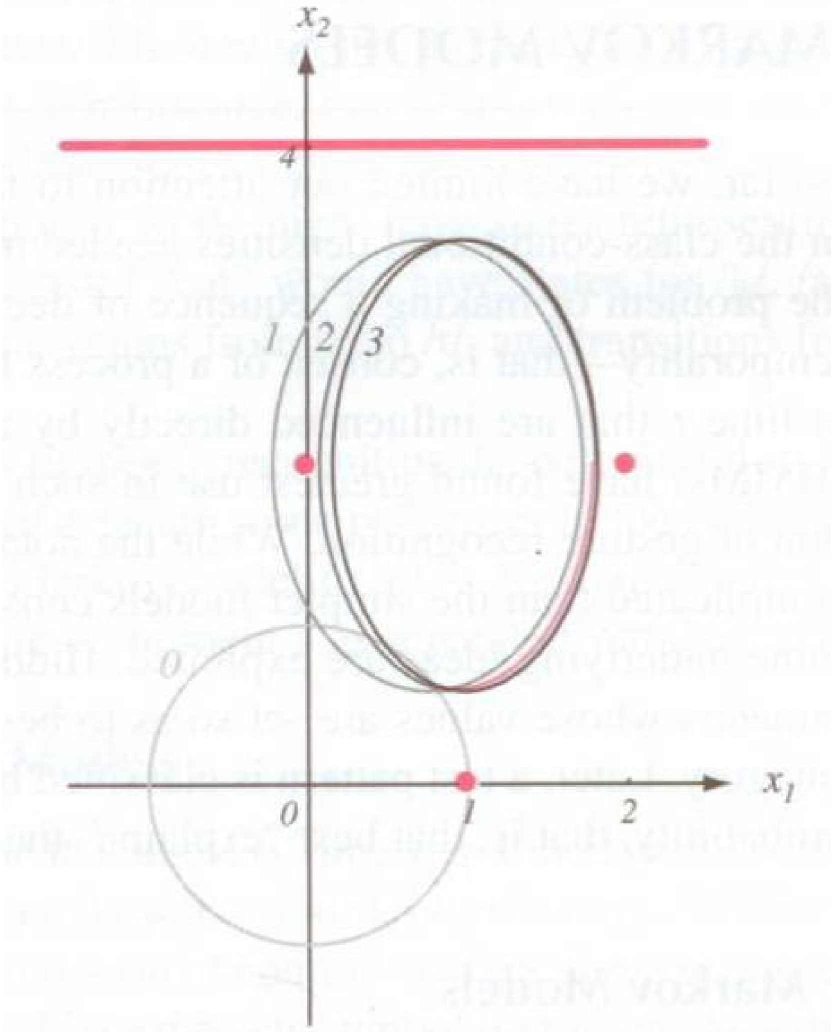
$$\boldsymbol{\theta}^1=\begin{pmatrix}0.75\\2.0\\0.938\\2.0\end{pmatrix}$$

Example: 2D Model

After 3 iterations, the algorithm converges at

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

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



Generalized Expectation-Maximization (GEM)

- Instead of maximizing $Q(\theta; \theta^i)$, we find some θ^{i+1} such that

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

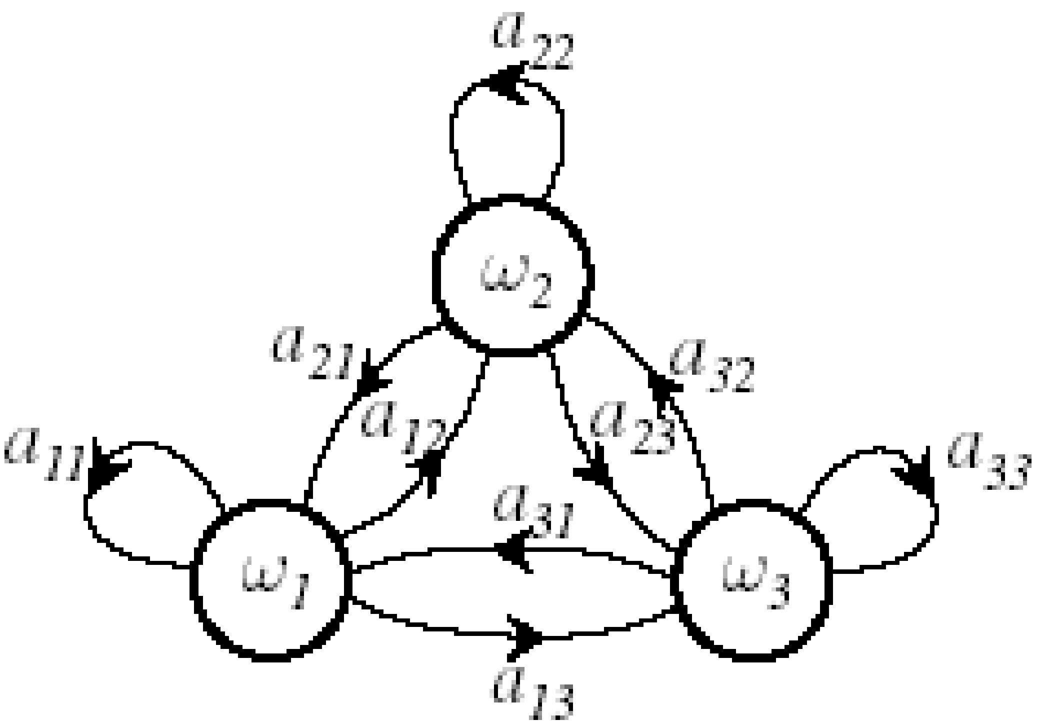
and is also guaranteed to converge

- Convergence will not as rapid
- Offers great freedom to choose computationally simpler steps
  - e.g., using maximum-likelihood value of unknown values, if they lead to a greater likelihood

Hidden Markov Model (HMM)

- Used for problems of making a series of decisions
  - *e.g.*, speech or gesture recognition
- Problem states at time t are influenced directly by a state at t-1
- More reference:
  - L. A. Rabiner and B. W. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993, Chapter 6.

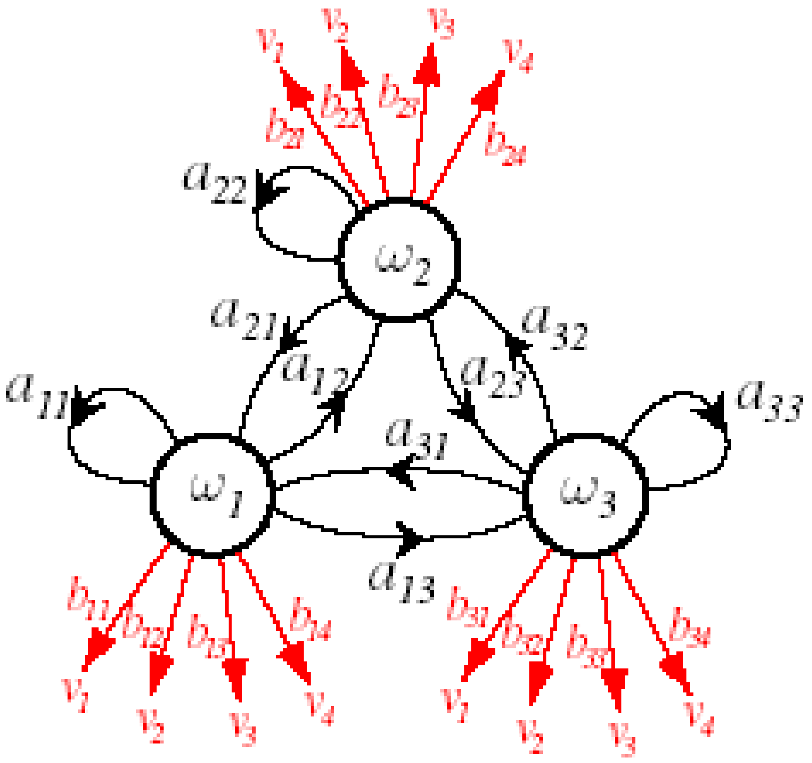
First Order Markov Models



sequence of states $\boldsymbol{\omega}^T = \{\omega(1), \omega(2), \cdots, \omega(T)\}$

$e.g., \boldsymbol{\omega}^6 = \{\omega_1, \omega_3, \omega_2, \omega_2, \omega_1, \omega_3\}, \quad P(\boldsymbol{\omega}^6 \mid \boldsymbol{\theta}) = a_{13}a_{32}a_{22}a_{21}a_{13}$

First Order Hidden Markov Models



$$\text{Sequence of visiblestates } \mathbf{V}^T = \{v(1), v(2), \cdots, v(T)\}$$

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

Hidden Markov Model Probabilities

final or absorbing state $\omega_0 : a_{00} = 1$

transition probability : $a_{ij} = P(\omega_j(t+1) \mid \omega_i(t))$

probability of emission of a visible state :

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

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

Hidden Markov Model Computation

- Evaluation problem
  - □ Given a_{ij} and b_{jk}, determine $P(\mathbf{V}^T|\theta)$
- Decoding problem
  - □ Given $\mathbf{V}^T$, determine the most likely sequence of hidden states that lead to $\mathbf{V}^T$
- Learning problem
  - □ Given training observations of visible symbols and the coarse structure but not the probabilities, determine a_{ij} and b_{jk}

Evaluation

$$P(\mathbf{V}^T) = \sum_{r=1}^{r_{\max}} P(\mathbf{V}^T \mid \omega_r^T) P(\omega_r^T)$$

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

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

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

HMM Forward

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

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

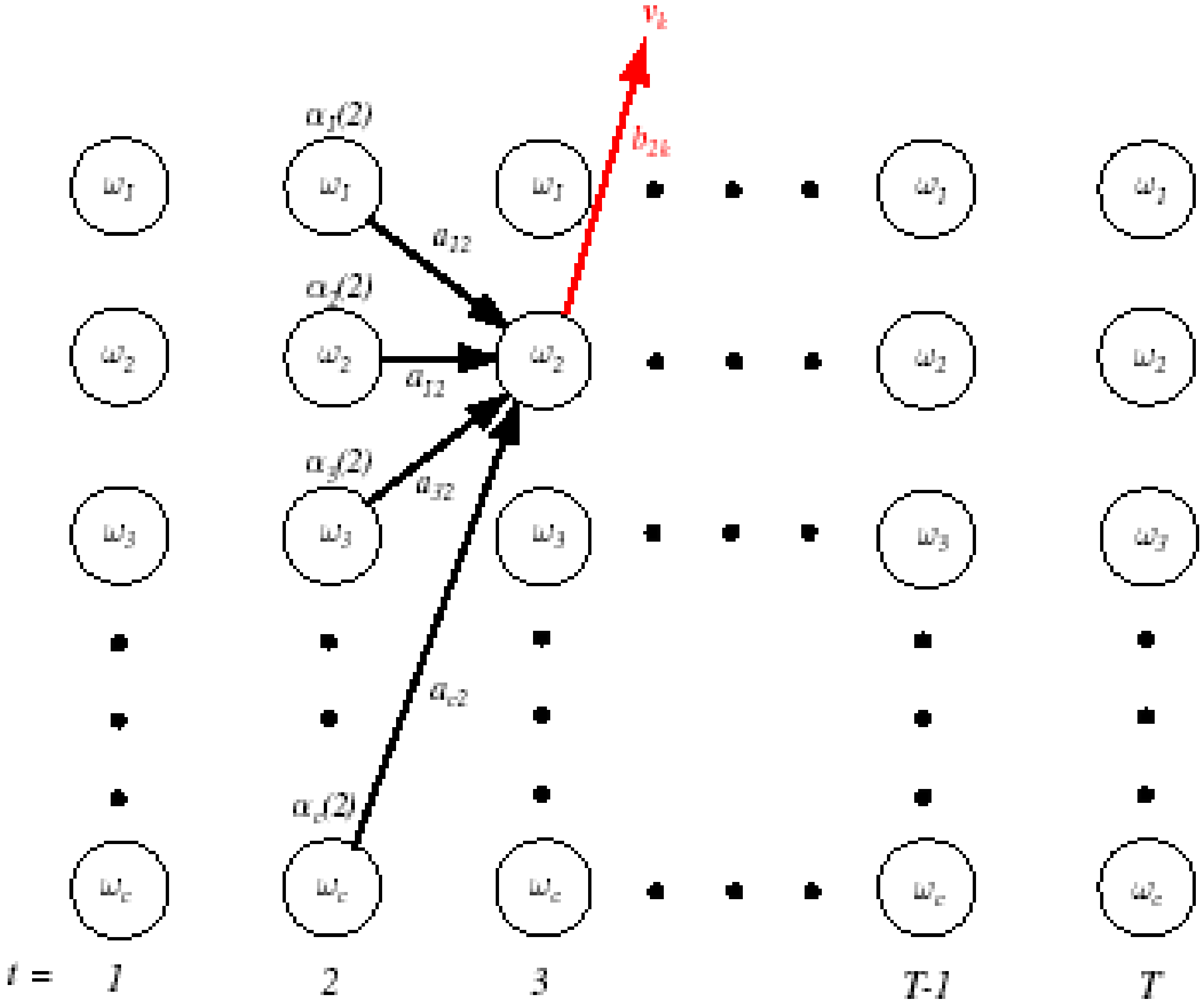
$$= P(v_k \mid \omega_j(t)) \sum_{i=1}^{c} P(\omega_j(t) \mid \omega_i(t-1)) P(\omega_i(t-1), \mathbf{V}^{t-1})$$

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

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

$$\alpha_0(T) = P(\omega_0(T), \mathbf{V}^T) = P(\omega_0(T) \mid \mathbf{V}^T) P(\mathbf{V}^T) = P(\mathbf{V}^T)$$

HMM Forward and Trellis



HMM Forward

```
initialize $t \leftarrow 0, a_{ij}, b_{jk}, \mathbf{V}^T, \alpha_j(0)$
    for $t \leftarrow t+1,\ j = 0, \cdots, c$
        $\alpha_j(t) \leftarrow b_{jkv(t)} \sum_{i=1}^{c} \alpha_i(t-1) a_{ij}$
    until $t = T$
return $P(\mathbf{V}^T) = \alpha_0(T)$ for final state
end
```

HMM Backward

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

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

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

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

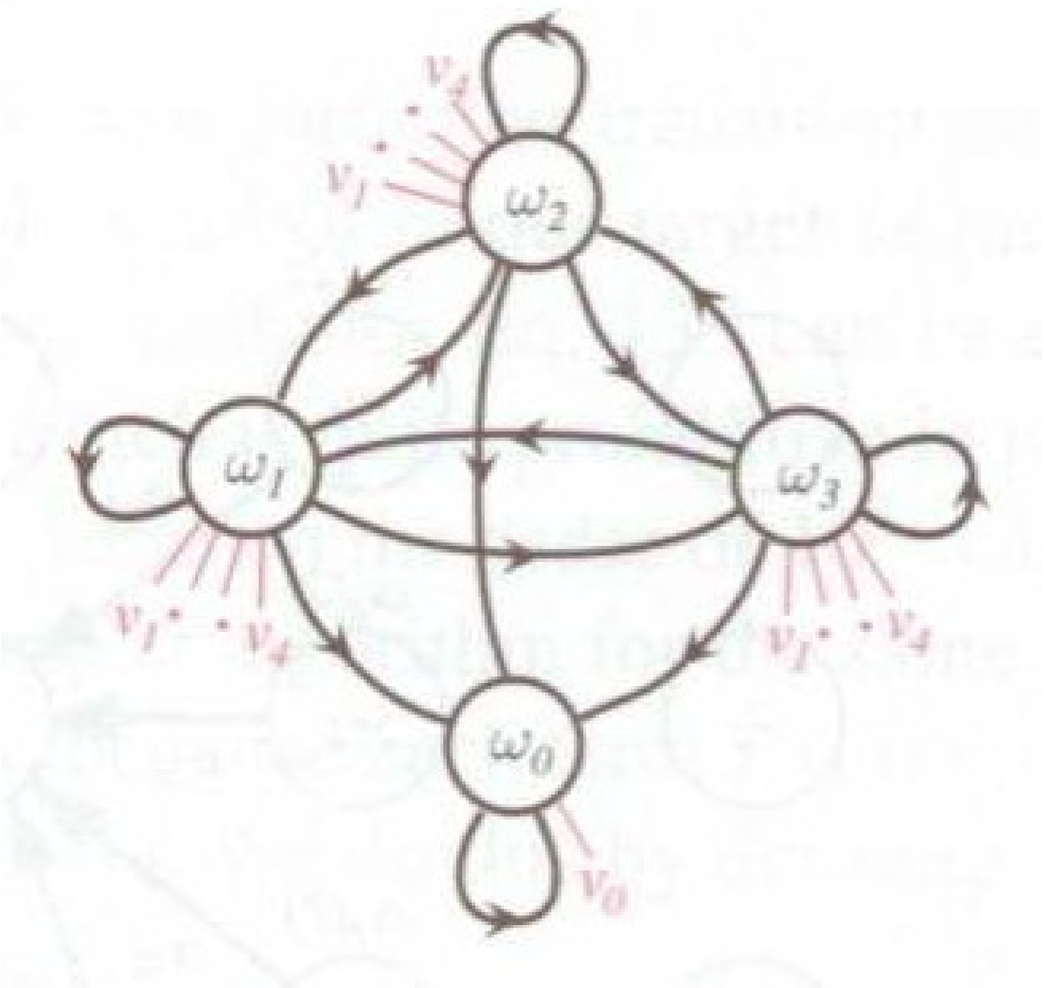
$$\beta_i(T) = \begin{cases} 0, & i \neq 0 \\ 1, & i = 0 \end{cases}$$

$$\beta_{init}(0) = P(\omega_{init}(0), \mathbf{V}^T) = P(\omega_{init}(0) \mid \mathbf{V}^T) P(\mathbf{V}^T) = P(\mathbf{V}^T)$$

HMM Backward

```
initialize $t \leftarrow T, a_{ij}, b_{jk}, \mathbf{V}^T, \beta_j(T)$
    for $t \leftarrow t-1, i = 1, \cdots, c$
        $\beta_i(t) \leftarrow \sum_{j=0}^{c} \beta_j(t+1) a_{ij} b_{jkv(t+1)}$
    until $t = 0$
return $P(\mathbf{V}^T) = \beta_0(0)$ for initial state
end
```

Example 3: Hidden Markov Model

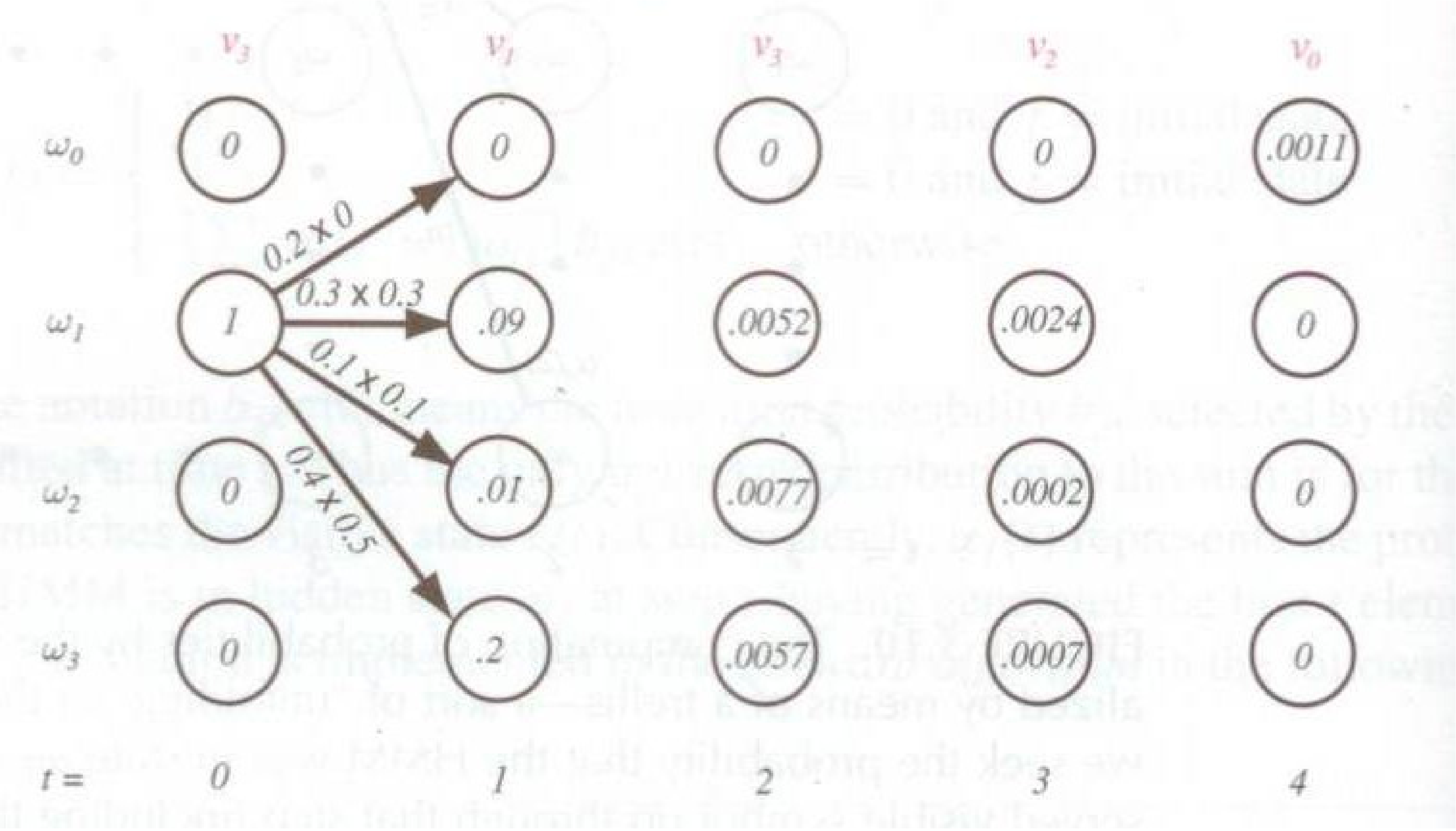


Example 3: Hidden Markov Model

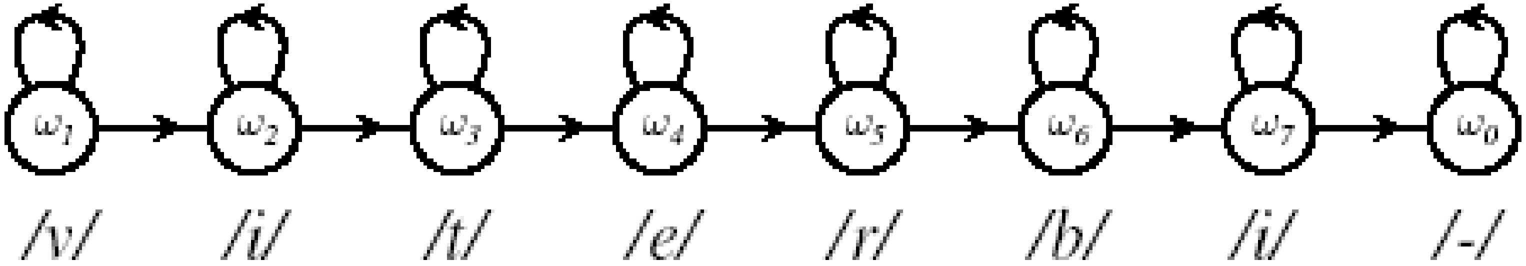
$$a_{ij} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.2 & 0.3 & 0.1 & 0.4 \\ 0.2 & 0.5 & 0.2 & 0.1 \\ 0.8 & 0.1 & 0.0 & 0.1 \end{pmatrix}$$

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

Example 3: Hidden Markov Model

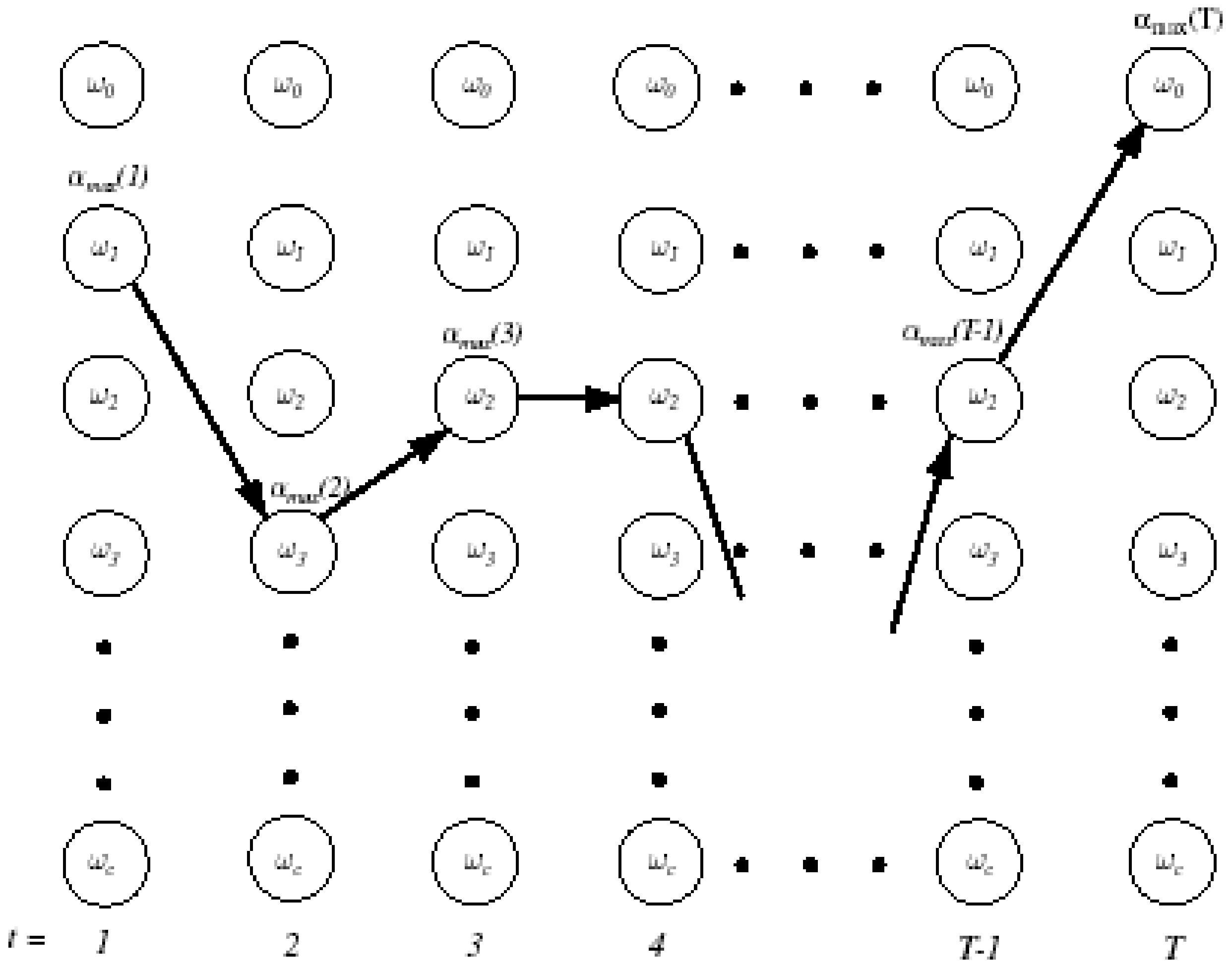


Left-to-Right Models for Speech



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

HMM Decoding



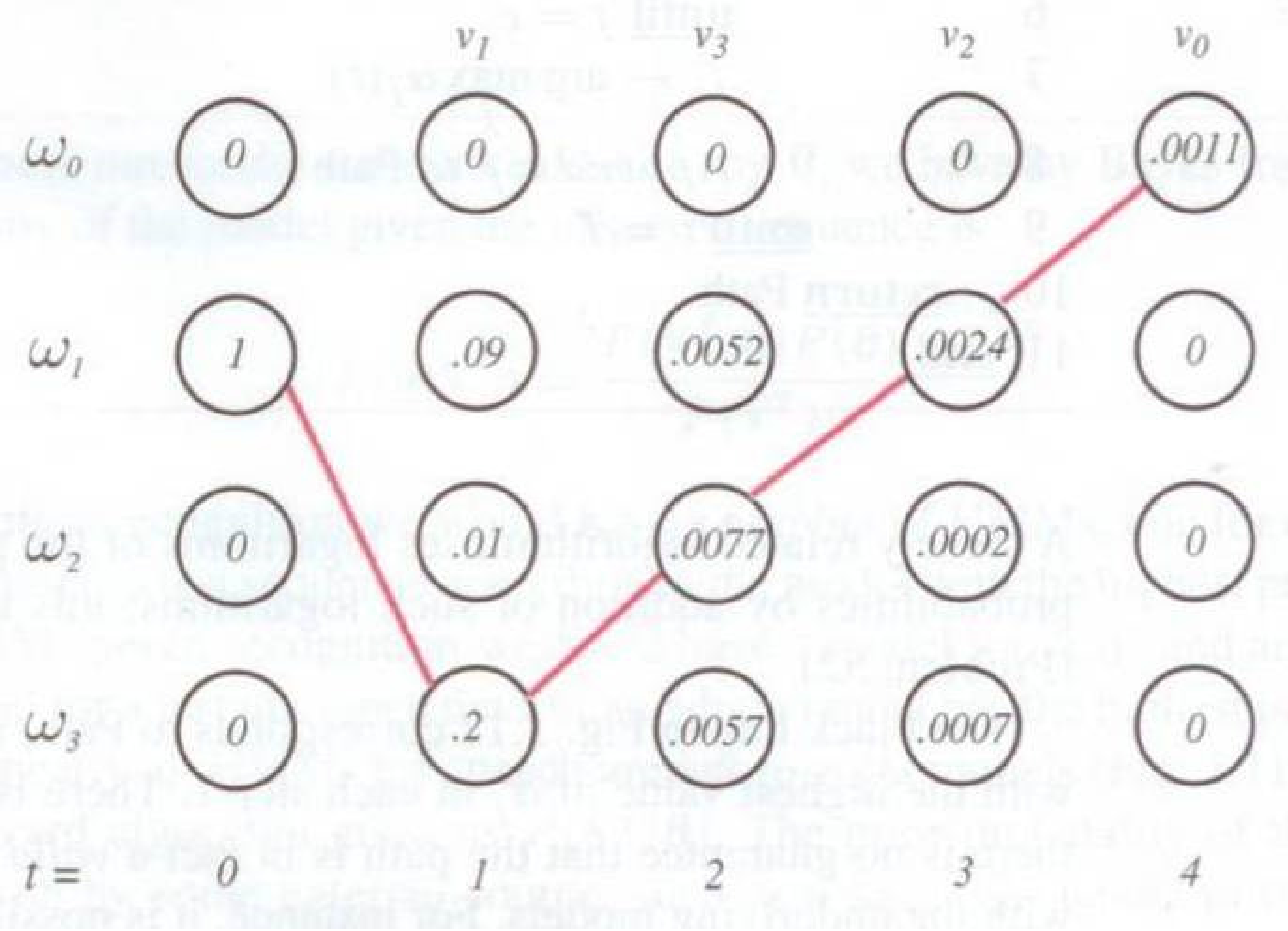
Problem of Local Optimization

- This decoding algorithm depends only on the single previous time step, not the full sequence
- Not guarantee that the path is indeed allowable

HMM Decoding

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

Example 4: HMM Decoding



Forward-Backward Algorithm

- Determines model parameters, a_{ij} and b_{jk}, from an ensemble of training samples
- An instance of a generalized expectation-maximization algorithm
- No known method for the optimal or most likely set of parameters from data

Probability of Transition

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

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

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

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

Improved Estimate for a_{ij}

Expected number of transitions between state $\omega_i(t-1)$ and $\omega_j(t)$ at any time in the sequence :

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

Total expected number of any transitions from ω_i : $\sum_{t=1}^{T} \sum_{k} \gamma_{ik}(t)$

Estimate of a_{ij} :

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

Improved Estimate for b_{jk}

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

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
```

- 成分分析与辨别式
  - □组合特征从而降低特征空间的维数
  - □线形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传递
    - PCA (主成份分析) “在最小均方误差意义下的数据的最优表示的映射”
    - MDA (多类判别分析) “在最小均方误差意义下的数据的最有分类的映射”

- 隐藏马尔可夫模型：
  - □ 马尔可夫链
  - □ 目标: 建立一系列决策
    - Processes that unfold in time, states at time t are influenced by a state at time t-1
    - 应用: 语音识别, 姿势识别,部分语音追踪和DNA 排序
    - 所有无记忆的随机过程

      $\omega^T = \{\omega(1), \omega(2), \omega(3), \ldots, \omega(T)\}$ 为状态序列

      我们可以得到 $\omega^6 = \{\omega1, \omega4, \omega2, \omega2, \omega1, \omega4\}$
    - 这个系统能够再现不同阶段的状态而且不是所有的状态都需要巡视

- 一阶马尔可夫模型
  - 所有序列的结果都由传递概率表示

$$P(\omega_j(t + 1) \mid \omega_i (t)) = a_{ij}$$

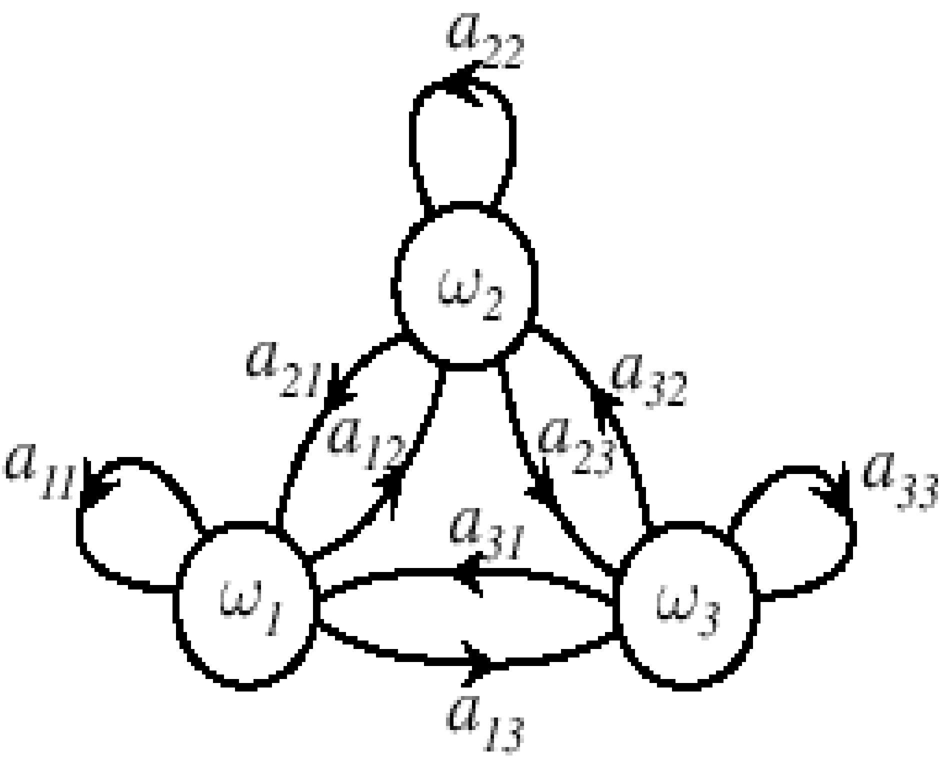


FIGURE 3.8. The discrete states, ω_i, in a basic Markov model are represented by nodes, and the transition probabilities, a_{ij}, are represented by links. In a first-order discrete-time Markov model, at any step t the full system is in a particular state $\omega(t)$. The state at step $t+1$ is a random function that depends solely on the state at step t and the transition probabilities. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

$\theta = (a_{ij}, \omega^T)$

$P(\omega^T \mid \theta) = a_{14} \cdot a_{42} \cdot a_{22} \cdot a_{21} \cdot a_{14} \cdot P(\omega(1) = \omega_i)$

例子: 语音识别

"production of spoken words"

Production of the word: "模式" 由音素表示

/p/ /a/ /tt/ /er/ /n/ // (// = silent state)

Transitions from /p/ to /a/, /a/ to /tt/, /tt/ to er/, /er/ to /n/ and /n/ to a silent state

隐性马尔可夫模型 (HMM)

- 可视状态与隐藏状态的相互影响

  $\sum b_{jk} = 1$ *对所有 j 当* $b_{jk} = P(V_k(t) \mid \omega_j(t))$.

- 这个模型存在三个问题
  - 估计问题
  - 解码问题
  - 学习问题

■估计问题

该模型生产出一列可视状态V^T是有可能的，即:

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

当每个r指示T个隐藏状态所组成的一组特殊序列

$$\omega_r^T=\{\omega(1),\omega(2),\ldots,\omega(T)\}$$

$$(1)\qquad P(V^T\,|\,\omega_r^T)=\prod_{t=1}^{t=T}P(v(t)\,|\,\omega(t))$$

$$(2)\qquad P(\omega_r^T)=\prod_{t=1}^{t=T}P(\omega(t)\,|\,\omega(t-1))$$

使用方程 (1) 和 (2), 我们能够写成:

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

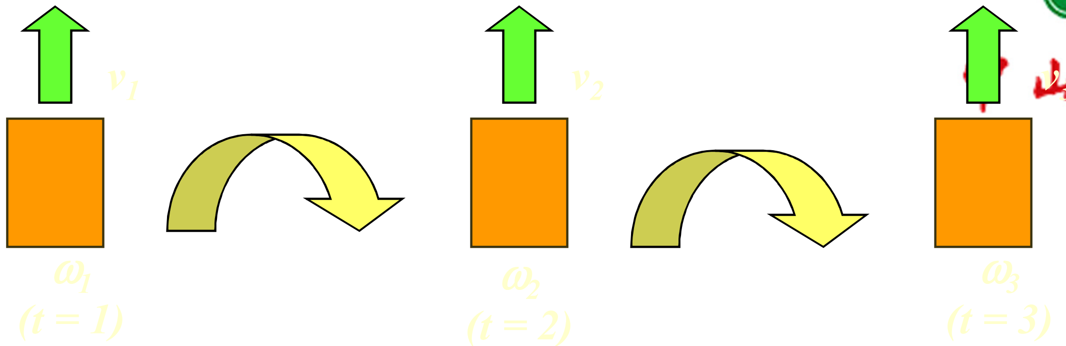
例子: 令 $\omega_1, \omega_2, \omega_3$ 为隐藏状态; v_1, v_2, v_3 为可视状态

$V^3 = \{v_1, v_2, v_3\}$ 为可视状态序列

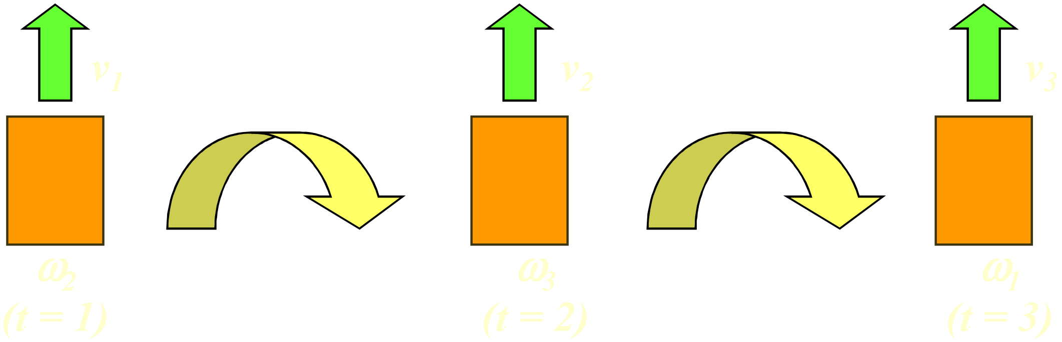
$P(\{v_1, v_2, v_3\}) = P(\omega_1).P(v_1 \mid \omega_1).P(\omega_2 \mid \omega_1).P(v_2 \mid \omega_2).P(\omega_3 \mid \omega_2).P(v_3 \mid \omega_3)$

+…+ (总的可能项数= 所有的可能性 (3^3= 27) 的情况 !)

第一概率:



第二概率:



$$P(\{v_1, v_2, v_3\}) = P(\omega_2).P(v_1 \mid \omega_2).P(\omega_3 \mid \omega_2).P(v_2 \mid \omega_3).P(\omega_1 \mid \omega_3).P(v_3 \mid \omega_1) + \ldots +$$

因此:

$$P(\{v_1, v_2, v_3\}) = \sum_{\substack{\text{possible sequence}\\ \text{of hidden states}}} \prod_{t=1}^{t=3} P(v(t) \mid \omega(t)).P(\omega(t) \mid \omega(t-1))$$

■ 解码问题 (最优状态序列)

假设V^T为可视状态序列，解码问题就是找出最有可能的隐藏状态序列 .

这个问题用数学的方式表示如下：
找出单个的“最佳”状态序列 *(隐藏状态)*

$$\hat{\omega}(1),\hat{\omega}(2),\ldots,\hat{\omega}(T)\ \ such\ \ that:$$

$$\hat{\omega}(1),\hat{\omega}(2),\ldots,\hat{\omega}(T)=\underset{\omega(1),\omega(2),\ldots,\omega(T)}{\arg\max}\ P\left[\omega(1),\omega(2),\ldots,\omega(T),v(1),v(2),\ldots,V(T)\mid\lambda\right]$$

注意最后的总和消失了，因为我们只想找到唯一的一个最佳情况

当: $\lambda = [\pi, A, B]$

$$\pi = P(\omega(1) = \omega)\ (\text{最初的状态概率})$$

$$A = a_{ij} = P(\omega(t+1) = j \mid \omega(t) = i)$$

$$B = b_{jk} = P(v(t) = k \mid \omega(t) = j)$$

在之前的例子中，这些计算与最佳路径的选择一致:

$$\{\omega_1(t=1), \omega_2(t=2), \omega_3(t=3)\},\ \{\omega_2(t=1), \omega_3(t=2), \omega_1(t=3)\}$$

$$\{\omega_3(t=1), \omega_1(t=2), \omega_2(t=3)\},\ \{\omega_3(t=1), \omega_2(t=2), \omega_1(t=3)\}$$

$$\{\omega_2(t=1), \omega_1(t=2), \omega_3(t=3)\}$$

▪ 学习问题 (参数估计)

第三个问题涉及到找出一种方法调整模型参数λ = [π,A,B]使之满足一个特定的最优标准。我们需要找出最好的模型。

$$\hat{\lambda} = [\hat{\pi}, \hat{A}, \hat{B}]$$

然后最大化可视序列的概率：

$$\underset{\lambda}{Max}\, P(V^T \mid \lambda)$$

我们使用一个迭代 Welch或者Gradient来得到一个最优解