Generative Models: Fundamentals and Applications

Lecture 4:
Generative models for continuous data

Shuigeng Zhou, Wenhui Liu
College of CSAI

October 13, 2026

目录

- 生成分类器
- 单高斯模型
- 高斯混合模型
- 隐马尔可夫模型

生成分类器

- 关键：类条件密度函数$p(x|y=c,\theta)$

$$p(y=c|\mathbf{x},\theta)=\frac{p(\mathbf{x},y=c|\theta)}{p(\mathbf{x}|\theta)}$$

$$=\frac{p(\mathbf{x},y=c|\theta)}{\sum_{c'\in C}p(\mathbf{x},y=c'|\theta)}$$

$$=\frac{p(y=c\mid\theta)p(\mathbf{x}|y=c,\theta)}{\sum_{c'\in C}p(y=c'\mid\theta)p(\mathbf{x}|y=c',\theta)}$$

$$\propto p(y=c\mid\theta)p(\mathbf{x}|y=c,\theta)$$

生成分类器

- 分类器

$$\hat{y}(\mathbf{x}) = \underset{c}{\operatorname{argmax}} \left[\log p(y = c|\boldsymbol{\pi}) + \log p(\mathbf{x}|\boldsymbol{\theta}_c)\right]$$

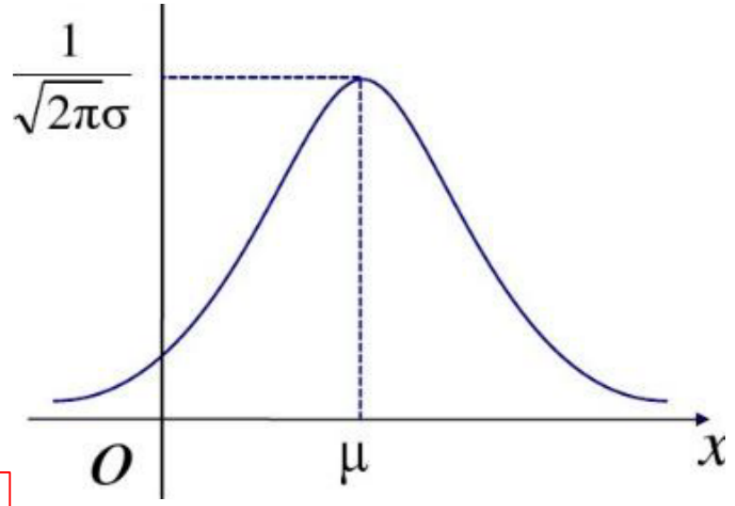
  - $\boldsymbol{\pi}$：类先验的参数
    - $\boldsymbol{\pi}_c = \frac{N_c}{N}$
  - $\boldsymbol{\theta}_c$：第c个类的模型参数，例如高斯模型的参数

目录

- 生成分类器
- 单高斯模型
- 高斯混合模型
- 隐马尔可夫模型

单高斯模型

- 一元高斯分布
  - 一维随机变量x
  - 均值μ
  - 方差σ^2
  - 概率密度函数



$$\mathcal{N}(x|\mu,\sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

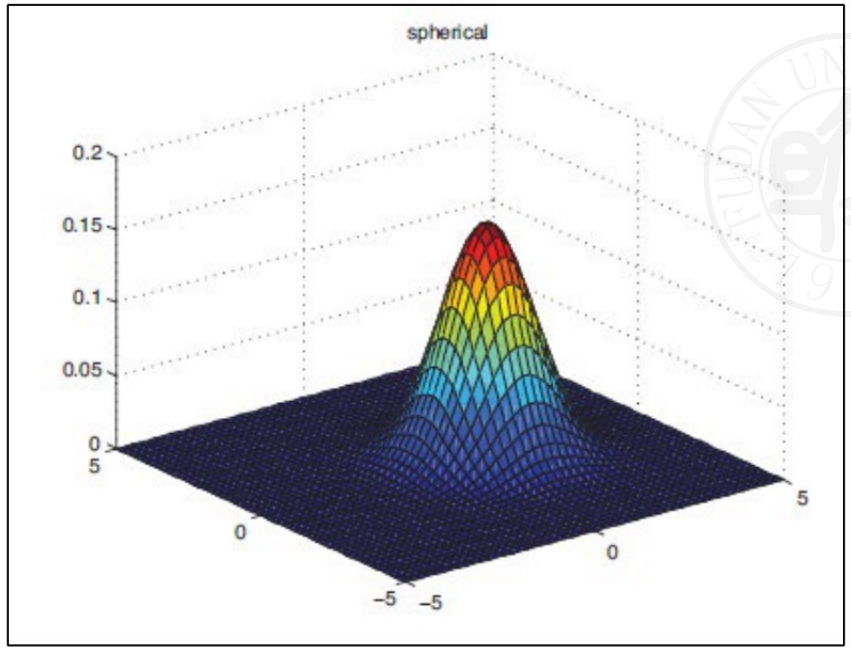
单高斯模型

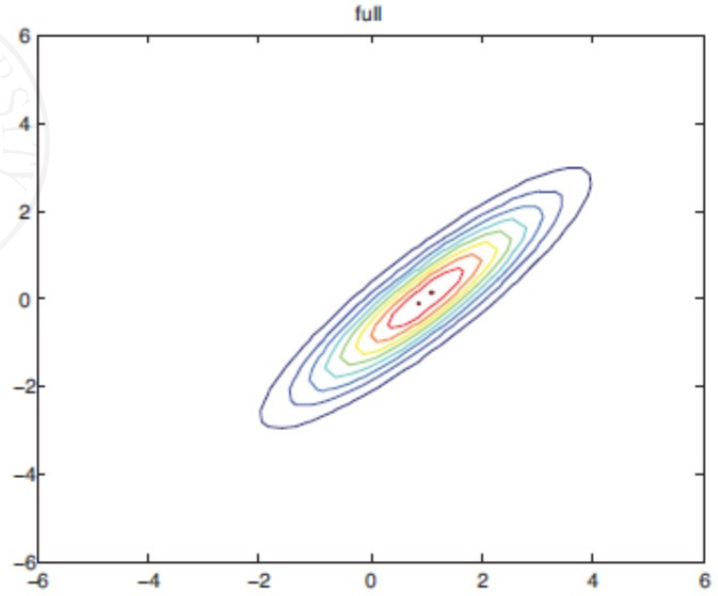
- 多元高斯分布（multivariate Gaussian）
  - D维随机变量x
  - 均值$\boldsymbol{\mu}$
  - 协方差矩阵$\boldsymbol{\Sigma}$
  - 联合概率密度函数

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

单高斯模型

- 二元高斯分布示意图





与马氏距离的联系

- 马氏距离（Mahalanobis distance）

$$D_M = (x-\mu)^T \Sigma^{-1} (x-\mu)$$

  - 等价于数据x变换之后的欧氏距离

  - 变换：平移μ，旋转U 缩放Λ
    - 对协方差矩阵Σ进行特征分解： $\Sigma = U\Lambda U^T$
      - U：特征向量组成的正交矩阵，且$U^T U = I$
      - Λ：特征值构成的对角矩阵，进行缩放

与马氏距离的联系

- 马氏距离（Mahalanobis distance）

$$D_M = (x-\mu)^T \Sigma^{-1} (x-\mu)$$

  - 变换：平移μ，旋转U 缩放Λ
    - 注意到$\Sigma = U\Lambda U^{\top}$，并记$z_i = u_i^T(x-\mu)$

$$\begin{aligned} D_M &= (x-\mu)^T U\Lambda^{-1}U^T(x-\mu) \\ &= (x-\mu)^T \left(\sum_{i=1}^{D} \frac{1}{\lambda_i} u_i u_i^T\right)(x-\mu) \\ &= \sum_{i=1}^{D} \frac{1}{\lambda_i} (x-\mu)^T u_i u_i^T (x-\mu) \\ &= \sum_{i=1}^{D} \frac{1}{\lambda_i} z_i^T z_i \end{aligned}$$

单高斯模型

- 给定N个独立同分布的样本$\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，估计模型参数$\boldsymbol{\mu}, \boldsymbol{\Sigma}$?
  - 极大似然估计 MLE
  - 最大后验概率估计 MAP

极大似然估计

联合似然函数

$$p(\mathcal{D} \mid \boldsymbol{\mu}, \Sigma) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right\}$$

对数似然函数

$$\begin{aligned}\ell(\boldsymbol{\mu}, \Sigma) = \ln p(\mathcal{D} \mid \boldsymbol{\mu}, \Sigma) &= \sum_{i=1}^{N}\left[-\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right] \\ &= -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\end{aligned}$$

平方和分解式

$$\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) = \mathrm{tr}\left(\Sigma^{-1}\boldsymbol{S}\right) + N(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T\Sigma^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})$$

$$\boldsymbol{S} = \sum_{i=1}^{N}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$$

代入可得

$$\begin{aligned}\ell(\boldsymbol{\mu}, \Sigma) = &-\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| \\ &-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\boldsymbol{S}\right) - \frac{N}{2}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T\Sigma^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})\end{aligned}$$

极大似然估计

矩阵微分必要知识

1. 迹微分：$d\,\mathrm{tr}(A) = \mathrm{tr}(dA)$

2. 行列式微分（A 对称正定）：

$$d\ln|A| = \mathrm{tr}\left(A^{-1}dA\right), \quad \frac{\partial \ln|A|}{\partial A} = A^{-1}$$

3. 逆矩阵微分：

$$d(A^{-1}) = -A^{-1}(dA)A^{-1}$$

4. 迹复合：

$$d\,\mathrm{tr}\left(BA^{-1}\right) = \mathrm{tr}\left(-A^{-1}BA^{-1}\,dA\right), \quad \frac{\partial\,\mathrm{tr}(BA^{-1})}{\partial A} = -A^{-1}BA^{-1}$$

5. 向量二次型（A 对称）：

$$\frac{\partial(\boldsymbol{a}-\boldsymbol{\theta})^{T}A(\boldsymbol{a}-\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = -2A(\boldsymbol{a}-\boldsymbol{\theta})$$

极大似然估计

对μ求梯度：

对数似然里只有最后一项含 $\boldsymbol{\mu}$：

$$\ell \supset -\frac{N}{2}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\Sigma^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})$$

Σ^{-1} 对称，套用二次型求导公式：

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{\mu}} &= -\frac{N}{2}\cdot\Big[-2\Sigma^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})\Big] \\ &= N\Sigma^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})\end{aligned}$$

令梯度等于零向量：

$$N\Sigma^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})=\mathbf{0}$$

Σ^{-1} 可逆，两边左乘 Σ/N：

$$\bar{\boldsymbol{x}}-\boldsymbol{\mu}=\mathbf{0}\quad\boxed{\hat{\boldsymbol{\mu}}_{\mathrm{MLE}}=\bar{\boldsymbol{x}}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i}$$

对Σ求梯度：

$$\ell(\Sigma)=C-\frac{N}{2}\ln|\Sigma|-\frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}\boldsymbol{S}\right)$$

C 是和 Σ 无关常数。

对 Σ 求导，分项：

$$\boxed{\frac{\partial}{\partial\Sigma}\left(-\frac{N}{2}\ln|\Sigma|\right)=-\frac{N}{2}\Sigma^{-1}}\quad\boxed{\frac{\partial}{\partial\Sigma}\left(-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}\boldsymbol{S})\right)=-\frac{1}{2}\left(-\Sigma^{-1}\boldsymbol{S}\Sigma^{-1}\right)=\frac{1}{2}\Sigma^{-1}\boldsymbol{S}\Sigma^{-1}}$$

合并梯度：

$$\frac{\partial\ell}{\partial\Sigma}=-\frac{N}{2}\Sigma^{-1}+\frac{1}{2}\Sigma^{-1}\boldsymbol{S}\Sigma^{-1}$$

令梯度为零矩阵：

$$-\frac{N}{2}\Sigma^{-1}+\frac{1}{2}\Sigma^{-1}\boldsymbol{S}\Sigma^{-1}=\mathbf{0}$$

两边同时左乘 Σ，再右乘 Σ：

$$-\frac{N}{2}\Sigma+\frac{1}{2}\boldsymbol{S}=\mathbf{0}\quad -N\Sigma+\boldsymbol{S}=\mathbf{0}\quad\boxed{\hat{\Sigma}_{\mathrm{MLE}}=\frac{1}{N}\boldsymbol{S}=\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^T}$$

最大后验概率估计

- 目标
  - 求后验概率 $p(\mu, \Sigma|\mathcal{D}) \propto p(\mathcal{D}|\mu, \Sigma)p(\mu, \Sigma)$ 最大化参数 μ, Σ

贝叶斯公式：

$$p(\boldsymbol{\mu}, \Sigma \mid \mathcal{D}) \propto p(\mathcal{D} \mid \boldsymbol{\mu}, \Sigma) p(\boldsymbol{\mu}, \Sigma)$$

关键是使用什么先验分布？

取对数：

$$\ln p(\boldsymbol{\mu}, \Sigma \mid \mathcal{D}) \propto \ell(\boldsymbol{\mu}, \Sigma) + \ln p(\boldsymbol{\mu}, \Sigma)$$

其中对数似然

$$\ell(\boldsymbol{\mu}, \Sigma) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}\boldsymbol{S}\right) - \frac{N}{2}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T\Sigma^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})$$

常数项 $-\frac{ND}{2}\ln(2\pi)$ 在比例式中可以丢弃。

最大后验概率估计

- NIW先验
  - NIW 先验：$p(\boldsymbol{\mu}, \Sigma) = \mathrm{NIW}(\boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{S}_0)$

    $p(\boldsymbol{\mu}, \Sigma) = p(\boldsymbol{\mu} \mid \Sigma)\, p(\Sigma)$

    条件高斯部分：

    $$\ln p(\boldsymbol{\mu} \mid \Sigma) \propto -\tfrac{\kappa_0}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)$$

    逆威沙特部分：

    $$\ln p(\Sigma) \propto -\tfrac{\nu_0 + D + 1}{2} \ln |\Sigma| - \tfrac{1}{2} \operatorname{tr}\left(\boldsymbol{S}_0 \Sigma^{-1}\right)$$

    合并先验对数（核）：

    $$\ln p(\boldsymbol{\mu}, \Sigma) \propto -\tfrac{\nu_0 + D + 1}{2} \ln |\Sigma| - \tfrac{\kappa_0}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) - \tfrac{1}{2} \operatorname{tr}\left(\boldsymbol{S}_0 \Sigma^{-1}\right)$$

最大后验概率估计

- NIW先验参数意义
  - NIW 先验：$p(\boldsymbol{\mu}, \Sigma) = \mathrm{NIW}(\boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{S}_0)$

1. $\boldsymbol{\mu}_0$：先验均值中心（D 维向量）
2. κ_0：均值侧的先验置信强度（伪样本量，正标量）
3. $\boldsymbol{S}_0$：协方差的先验尺度矩阵（$D \times D$ 对称正定矩阵）
4. ν_0：协方差侧的先验置信强度（IW 自由度，标量，满足 $\nu_0 > D - 1$）

最大后验概率估计

合并对数似然核，可得到后验形式

$$\ln p(\boldsymbol{\mu}, \Sigma \mid \mathcal{D}) \propto -\frac{N}{2}\ln|\Sigma| - \frac{1}{2}\mathrm{tr}(\Sigma^{-1}\boldsymbol{S}) - \frac{N}{2}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\Sigma^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})$$
$$-\frac{\nu_0+D+1}{2}\ln|\Sigma| - \frac{\kappa_0}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\Sigma^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) - \frac{1}{2}\mathrm{tr}(\boldsymbol{S}_0\Sigma^{-1})$$

$-\frac{N}{2}\ln|\Sigma| - \frac{\nu_0+D+1}{2}\ln|\Sigma| = -\frac{(N+\nu_0)+D+1}{2}\ln|\Sigma|$

定义后验自由度：

$\nu_N = \nu_0 + N$

$-\frac{1}{2}\left[\kappa_0(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\Sigma^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) + N(\boldsymbol{\mu}-\bar{\boldsymbol{x}})^T\Sigma^{-1}(\boldsymbol{\mu}-\bar{\boldsymbol{x}})\right]$

定义

$\kappa_N = \kappa_0 + N, \qquad \boldsymbol{\mu}_N = \frac{\kappa_0\boldsymbol{\mu}_0 + N\bar{\boldsymbol{x}}}{\kappa_0+N}$

配方恒等式：

$$\kappa_0(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\Sigma^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) + N(\boldsymbol{\mu}-\bar{\boldsymbol{x}})^T\Sigma^{-1}(\boldsymbol{\mu}-\bar{\boldsymbol{x}})$$
$$=\kappa_N(\boldsymbol{\mu}-\boldsymbol{\mu}_N)^T\Sigma^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_N) + \frac{\kappa_0 N}{\kappa_0+N}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)^T\Sigma^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)$$

代入后二次型拆成两块：

$$-\frac{\kappa_N}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_N)^T\Sigma^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_N) - \frac{1}{2}\cdot\frac{\kappa_0 N}{\kappa_0+N}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)^T\Sigma^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)$$

第二项不含 $\boldsymbol{\mu}$，利用 $\boldsymbol{a}^TA\boldsymbol{a} = \mathrm{tr}(A\boldsymbol{a}\boldsymbol{a}^T)$ 并入迹：

$$-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\cdot\frac{\kappa_0 N}{\kappa_0+N}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)^T\right)$$

1. $-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}\boldsymbol{S})$
2. $-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}\boldsymbol{S}_0)$
3. $-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\cdot\frac{\kappa_0 N}{\kappa_0+N}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)^T\right)$

合并在一起：

$$-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\left[\boldsymbol{S}_0 + \boldsymbol{S} + \frac{\kappa_0 N}{\kappa_0+N}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)^T\right]\right)$$

定义后验尺度矩阵：

$$\boldsymbol{S}_N = \boldsymbol{S}_0 + \boldsymbol{S} + \frac{\kappa_0 N}{\kappa_0+N}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)^T$$

最大后验概率估计

$$\ln p(\boldsymbol{\mu}, \Sigma \mid \mathcal{D}) \propto -\frac{\nu_N + D + 1}{2} \ln |\Sigma| - \frac{\kappa_N}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_N)^T \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_N) - \frac{1}{2} \operatorname{tr}\left(\Sigma^{-1} \boldsymbol{S}_N\right)$$

换回密度形式（指数）：

$$p(\boldsymbol{\mu}, \Sigma \mid \mathcal{D}) \propto |\Sigma|^{-(\nu_N + D + 1)/2} \exp\left\{-\tfrac{\kappa_N}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_N)^T \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_N) - \tfrac{1}{2} \operatorname{tr}\left(\Sigma^{-1} \boldsymbol{S}_N\right)\right\}$$

这个结构正是 NIW 的核：

$$\boxed{p(\boldsymbol{\mu}, \Sigma \mid \mathcal{D}) = \mathrm{NIW}\left(\boldsymbol{\mu}_N, \kappa_N, \nu_N, \boldsymbol{S}_N\right)}$$

最大后验概率估计

- 最大化后验分布$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{D})$
  - 后验众数（mode）

$$\hat{\boldsymbol{\mu}}_{\text{mode}} = \boldsymbol{\mu}_N = \frac{\kappa_0 \boldsymbol{\mu}_0 + N\bar{\boldsymbol{x}}}{\kappa_0 + N}$$

$$\hat{\Sigma}_{\text{mode}} = \frac{\boldsymbol{S}_N}{\nu_N + D + 1} = \frac{\boldsymbol{S}_0 + \boldsymbol{S} + \frac{\kappa_0 N}{\kappa_0 + N}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T}{(\nu_0 + N) + D + 1}$$

  - 如果设置$\kappa_0 = 0$

$$\hat{\boldsymbol{\mu}}_{\text{mode}} = \bar{\boldsymbol{x}}$$

$$\hat{\Sigma}_{\text{mode}} = \frac{\boldsymbol{S}_0 + \boldsymbol{S}}{(\nu_0 + N) + D + 1}$$

高斯判别分析

- 高斯判别分析（GDA）
  - 假设每个类 c 的样本服从多元高斯分布

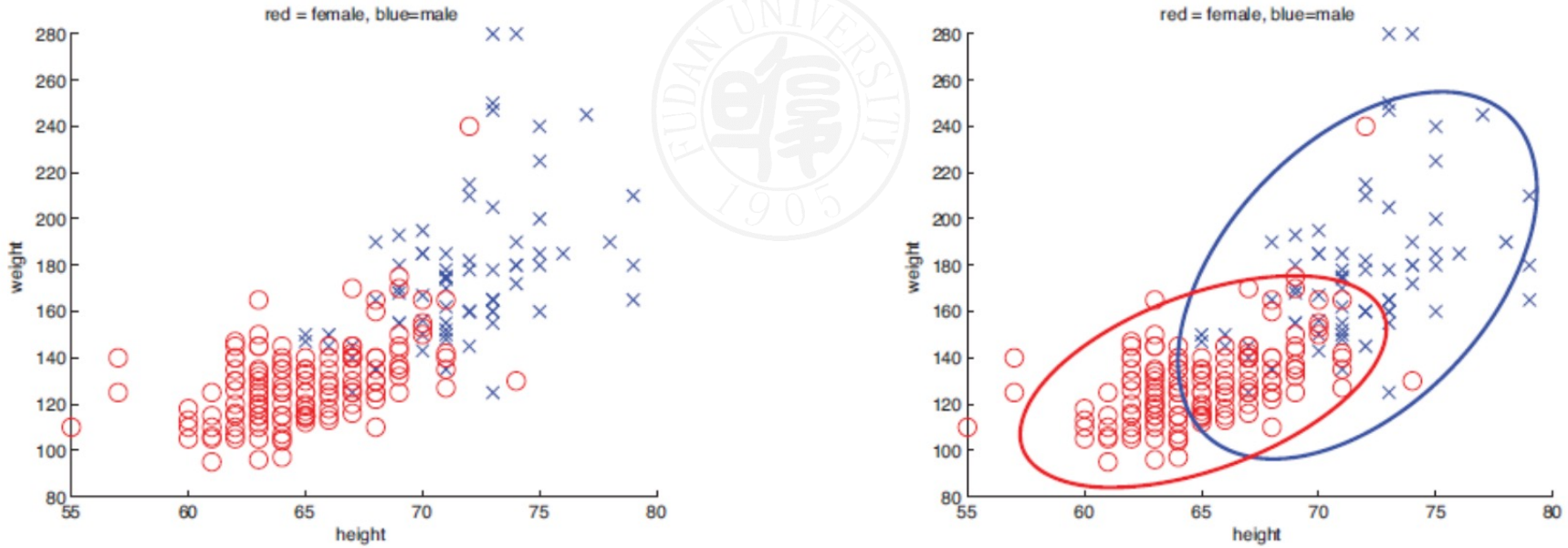
$$p(\mathbf{x}|y=c,\boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

  - 给定样本 $\boldsymbol{x}$，生成分类器预测最可能的类别

$$\begin{aligned}\hat{y}(\mathbf{x}) &= \underset{c}{\operatorname{argmax}} \left[\log p(y=c|\boldsymbol{\pi}) + \log p(\mathbf{x}|\boldsymbol{\theta}_c)\right] \\ &= \underset{c}{\operatorname{argmax}} \left[\log p(y=c|\boldsymbol{\pi}) + \log \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\right.\end{aligned}$$

高斯判别分析

- 左图：身高/体重数据
- 右图：每个类的可视化



(a) 高斯判别分析就是给每个类各拟合一个椭球，然后看新点落在谁的椭球里更合理 (b)

高斯判别分析

- 生成分类器
  - 如果类先验$\boldsymbol{\pi}_c$是均匀的

$$\hat{y}(\mathbf{x}) = \text{argmax}_c \left[\log p(y = c|\boldsymbol{\pi}_c) + \log \mathcal{N}(x|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\right]$$

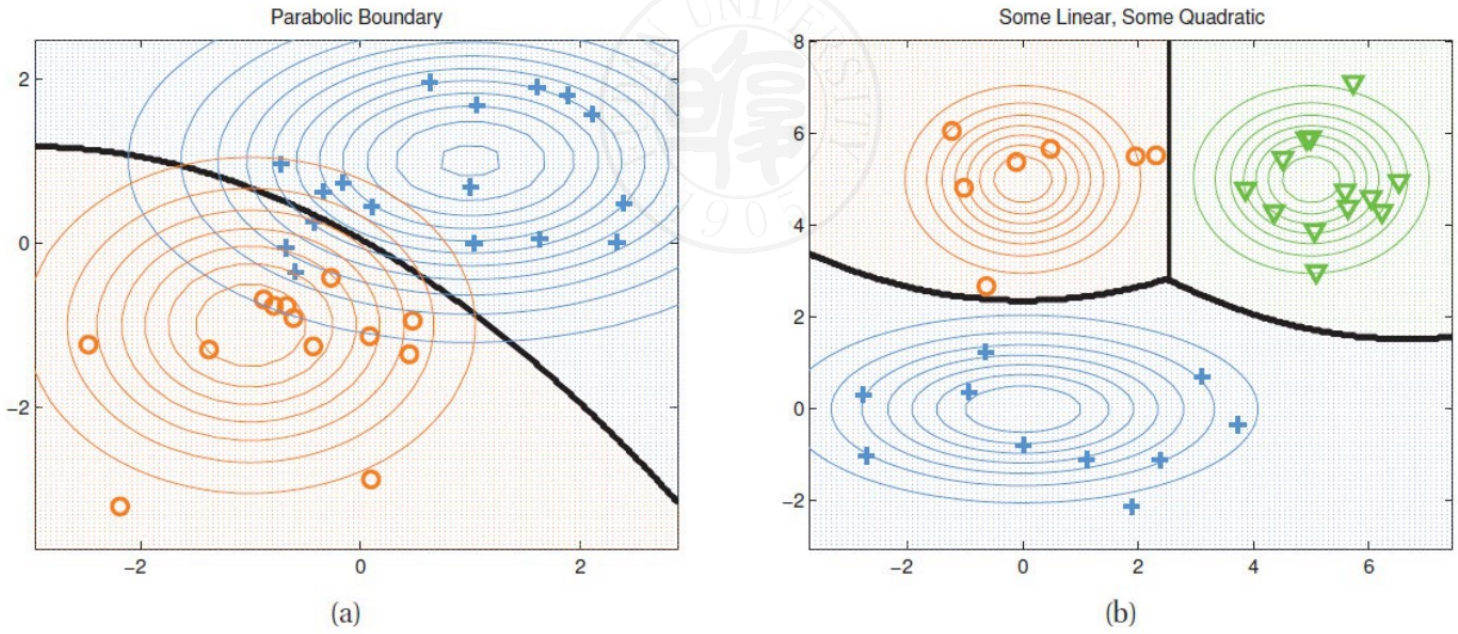
$$= \text{argmax}_c \log \left\{ \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}_c|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c{}^{-1}(\mathbf{x} - \boldsymbol{\mu}_c)\right] \right\}$$

$$= \text{argmax}_c -\frac{1}{2}\log|\boldsymbol{\Sigma}_c| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c{}^{-1}(\mathbf{x} - \boldsymbol{\mu}_c)$$

$$= \text{argmin}_c (\mathbf{x} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c{}^{-1}(\mathbf{x} - \boldsymbol{\mu}_c) + \log|\boldsymbol{\Sigma}_c|$$

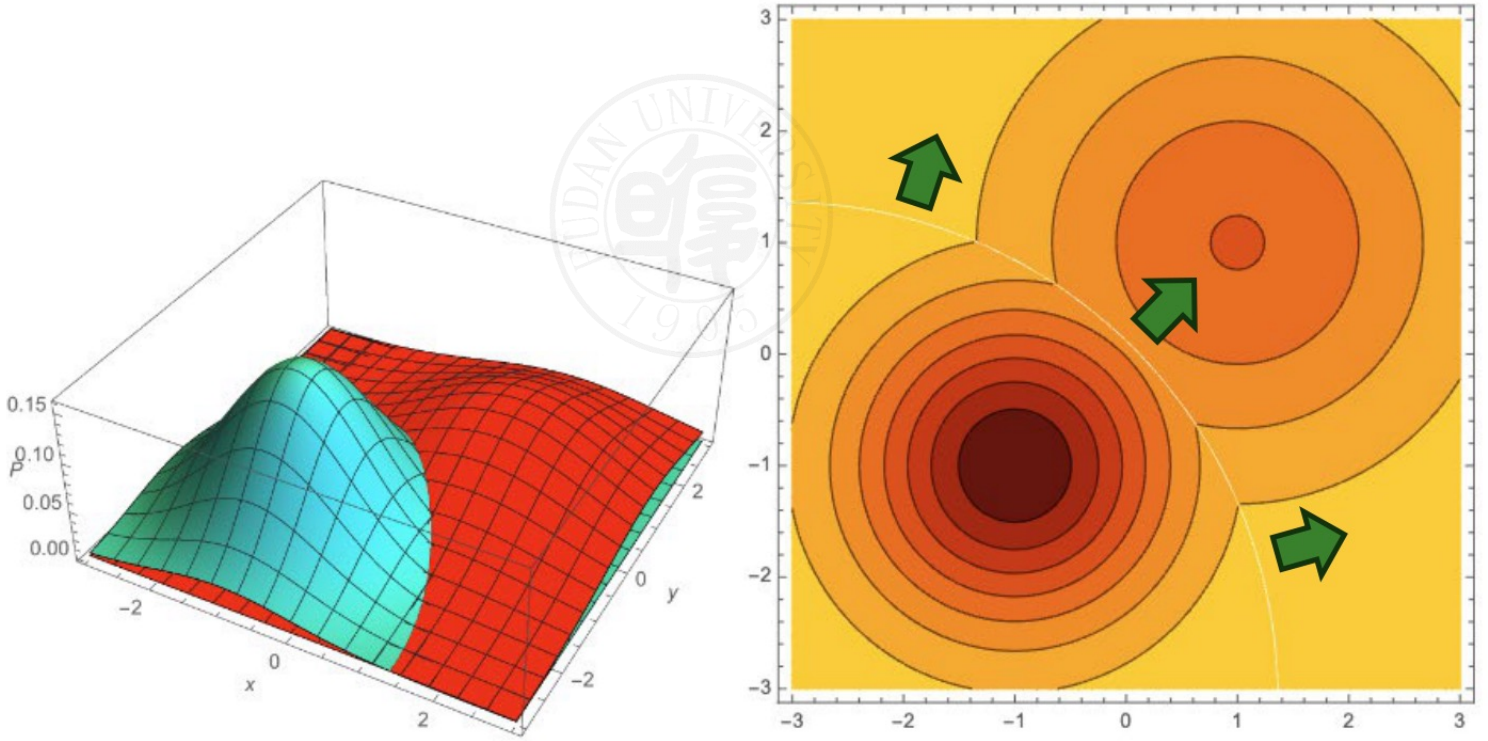
高斯判别分析

- 高斯判别分析又称为二次判别分析（QDA）
  - 判别边界 $(\mathbf{x}-\boldsymbol{\mu}_c)^T\boldsymbol{\Sigma}_c{}^{-1}(\mathbf{x}-\boldsymbol{\mu}_c)+\log|\boldsymbol{\Sigma}_c|$ 是二次形式



高斯判别分析

- 判别边界：两个类密度分布的“等势面”



高斯判别分析

- 特例：$\Sigma_c = \Sigma$

$$
\begin{aligned}
p(y=c|\mathbf{x},\boldsymbol{\theta}) &\propto \pi_c \exp\left[\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c\right] \\
&= \exp\left[\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c + \log\pi_c\right]\exp\left[-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right]
\end{aligned}
$$

与类别c无关的二次项，判别时可以忽略

  - 定义

$$
\begin{aligned}
\gamma_c &= -\frac{1}{2}\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c + \log\pi_c \\
\boldsymbol{\beta}_c &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c
\end{aligned}
$$

高斯判别分析

- 特例：$\Sigma_c = \Sigma$
  - 生成分类器

$$p(y = c|\mathbf{x}, \boldsymbol{\theta}) \propto \exp[\beta_c^T \mathbf{x} + \gamma_c]$$

- 任意两个类c和c'的决策边界恰好是一条直线

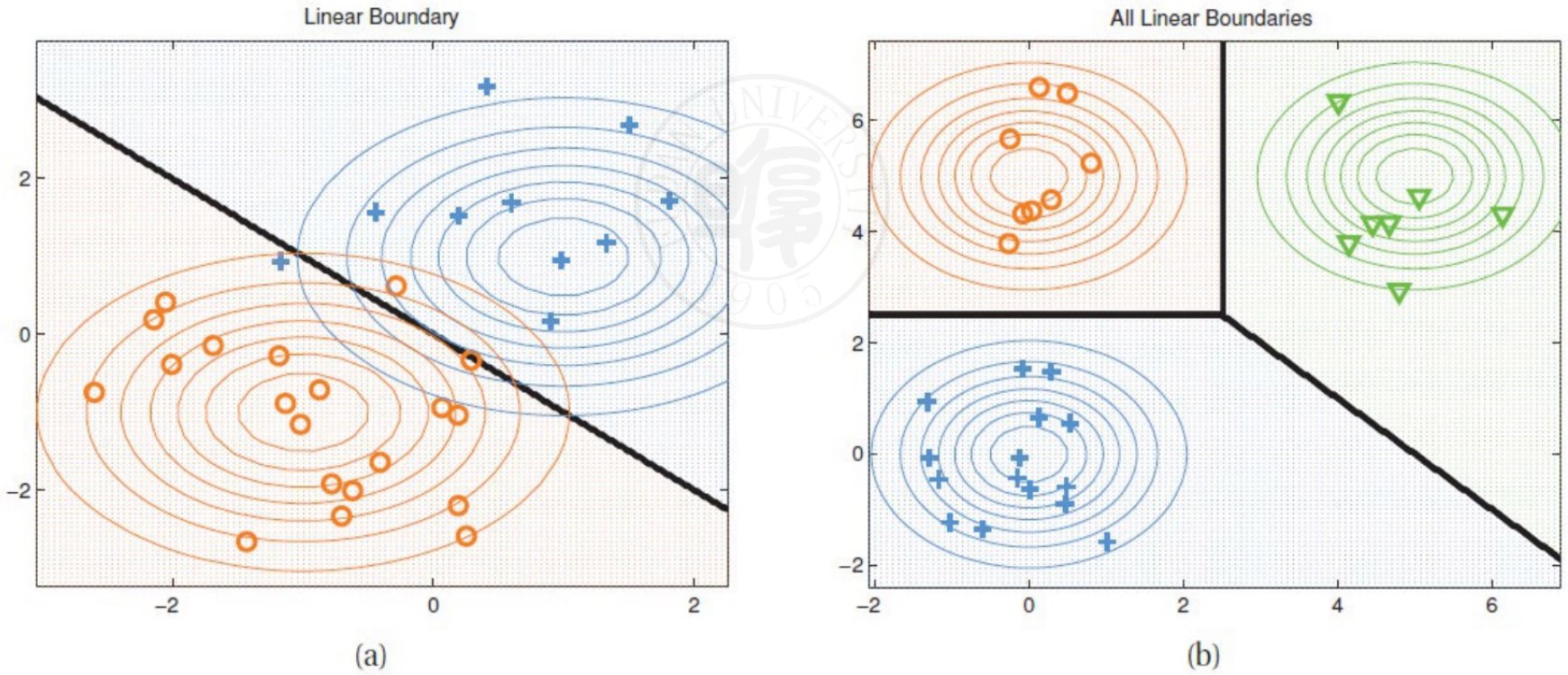
$$p(y = c|x, \theta) = p(y = c'|x, \theta) \Longrightarrow \beta_c^T x + \gamma_c = \beta_{c'}^T x + \gamma_{c'}$$
$$\Longrightarrow (\beta_c^T - \beta_{c'}^T)x = \gamma_{c'} - \gamma_c$$

- 亦称为线性判别分析(linear discriminant analysis, LDA)

$$\log p(y = c|\mathbf{x}, \boldsymbol{\theta}) \propto \beta_c^T \mathbf{x} + \gamma_c$$

高斯判别分析

- 特例：$\Sigma_c = \Sigma$

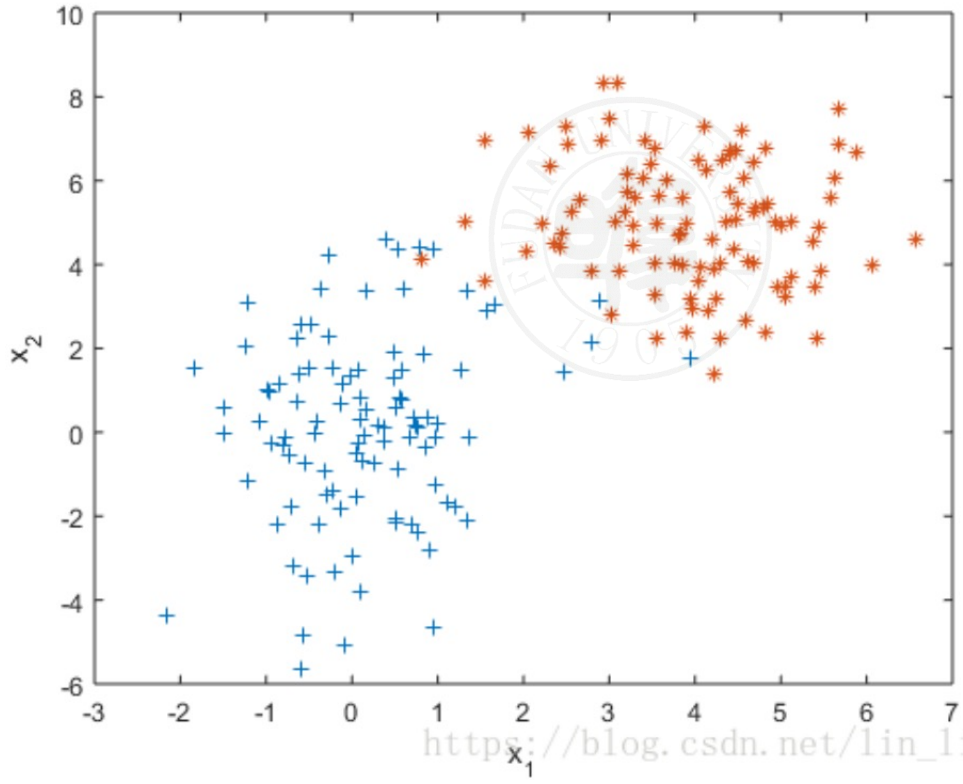


目录

- 生成分类器
- 单高斯模型
- 高斯混合模型
- 隐马尔可夫模型

为什么有高斯混合模型

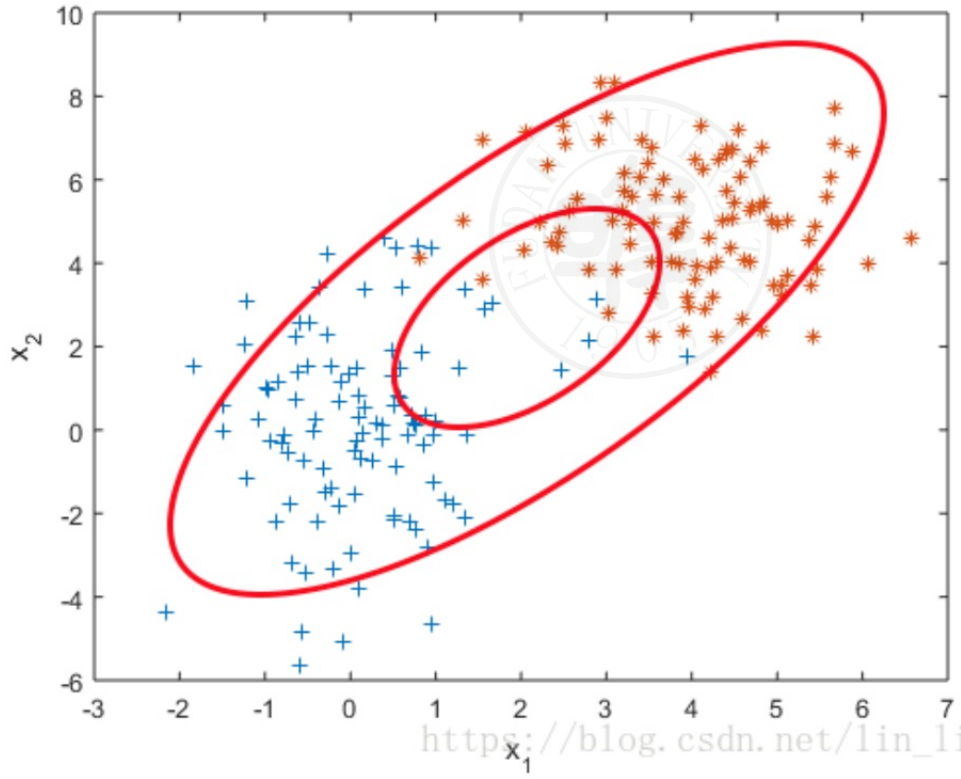
- 假设这组数据是由某个高斯分布产生



https://blog.csdn.net/lin_lim

为什么有高斯混合模型

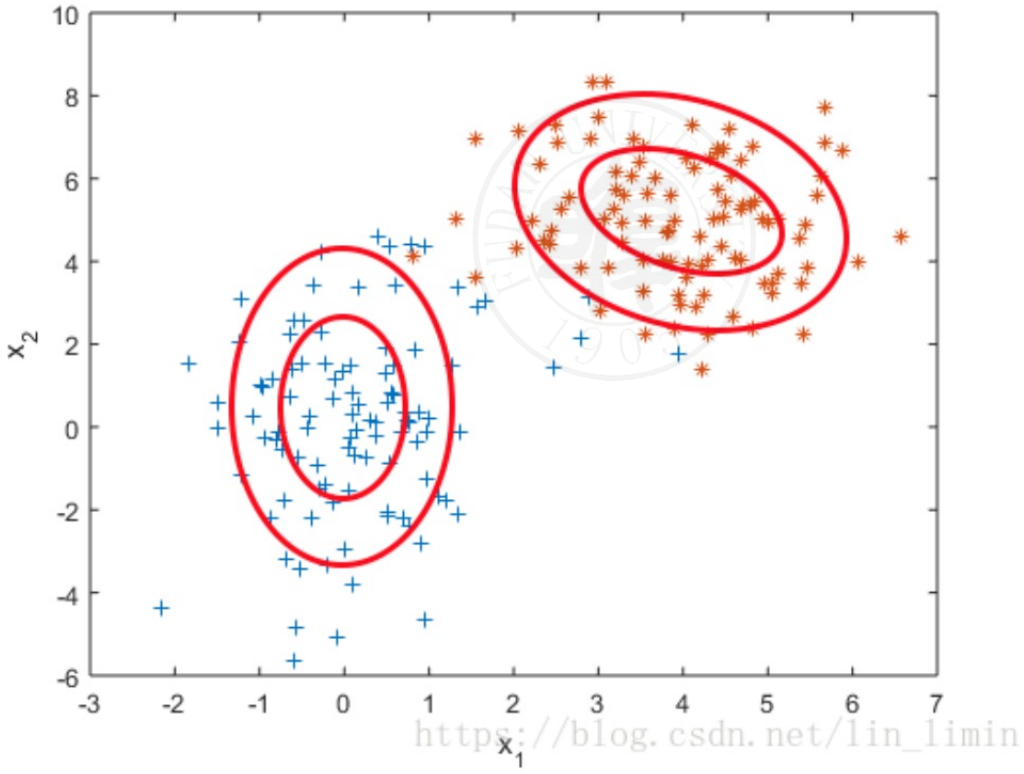
- 样本服从单高斯分布的假设并不合理



https://blog.csdn.net/lin_lim

为什么有高斯混合模型

- 这是用两个不同的高斯分布模型产生的数据



https://blog.csdn.net/lin_limin

高斯混合模型

- 假设混合高斯模型由K个高斯模型组成，其概率密度为

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_z p(z|\boldsymbol{\theta})p(\mathbf{x}|z,\boldsymbol{\theta})$$

$$= \sum_{k=1}^{K} p(z=k|\boldsymbol{\theta})p(\mathbf{x}|z=k,\boldsymbol{\theta})$$

$$= \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}|\theta_k)$$

  - 隐含类别 $z \in \{1,\dots,K\}$：确定数据$\mathbf{x}$是由哪个高斯模型产生的

高斯混合模型

- 高斯混合模型参数为$\boldsymbol{\theta}$

$$\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K)$$

  - π_k：第k个类的混合权重

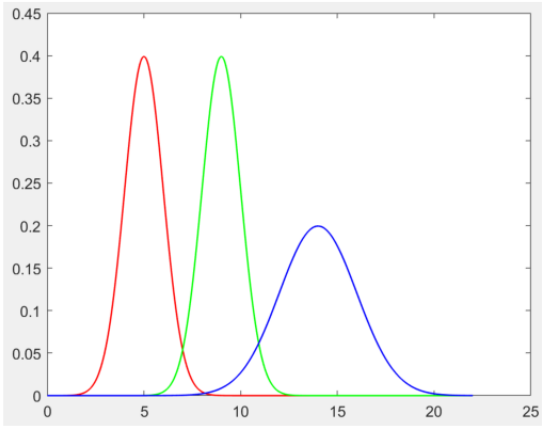
$$0 \leq \pi_k \leq 1 \qquad \sum_{k=1}^{K} \pi_k = 1$$

  - θ_k：第k个单高斯模型参数（均值、协方差），概率密度函数为p_k

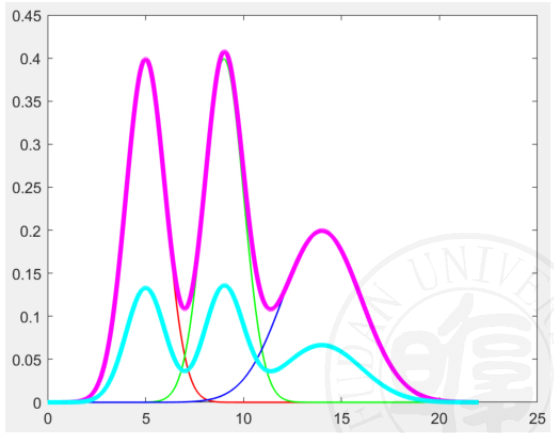
$$p_k(x|\theta_k) \sim \mathcal{N}(x|\mu_k, \Sigma_k)$$

高斯混合模型

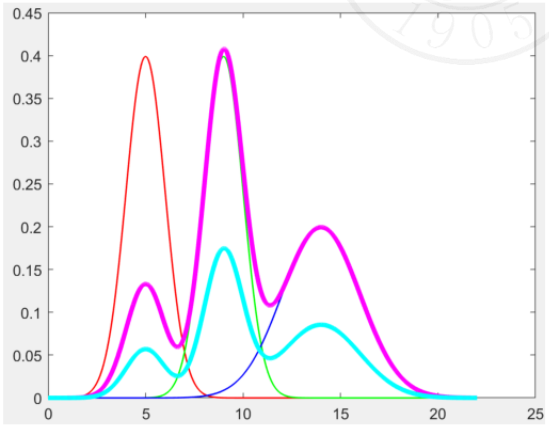
- 一维的例子



红线：N(5,1)
绿线：N(9,1)
蓝线：N(14,2)



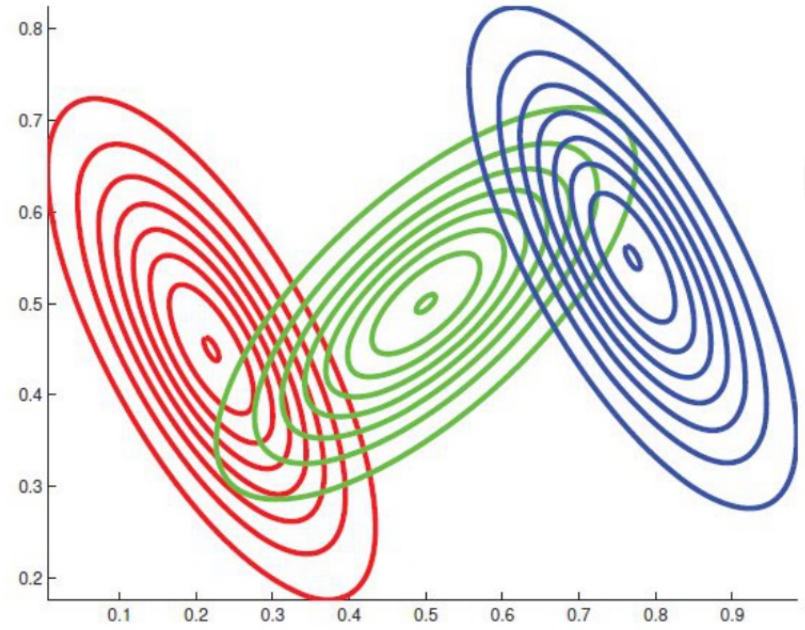
品红：N(5,1)+N(9,1)+N(14,2)
淡蓝：(1/3)*N(5,1)+ (1/3)*N(9,1)+ (1/3)*N(14,2)



品红：(1/3)*N(5,1)+ N(9,1)+ N(14,2)
淡蓝：(1/7)*N(5,1)+ (3/7)*N(9,1) + (3/7)*N(14,2)

二维高斯混合模型

- 直观上，高斯混合模型可以模拟任意分布函数



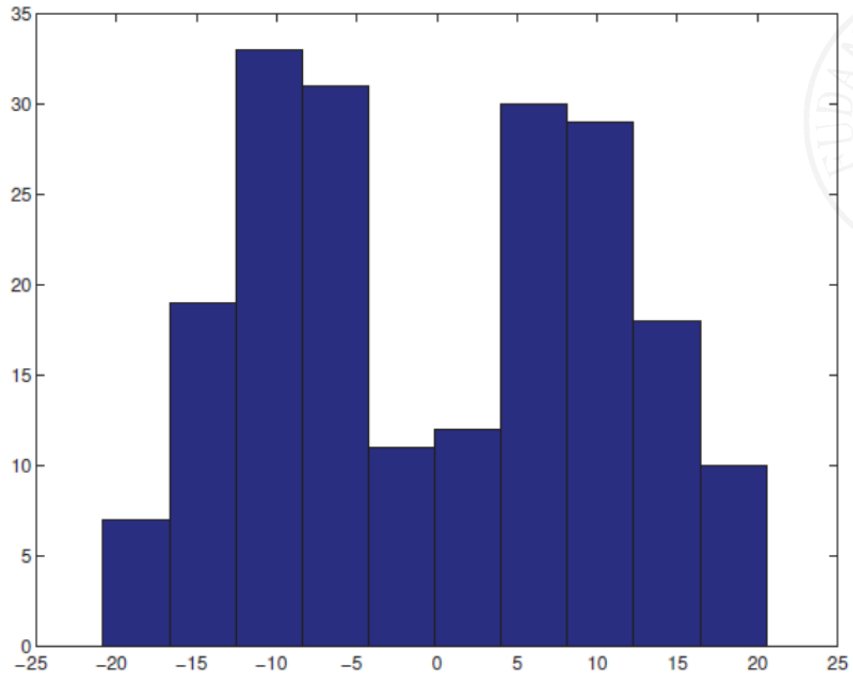


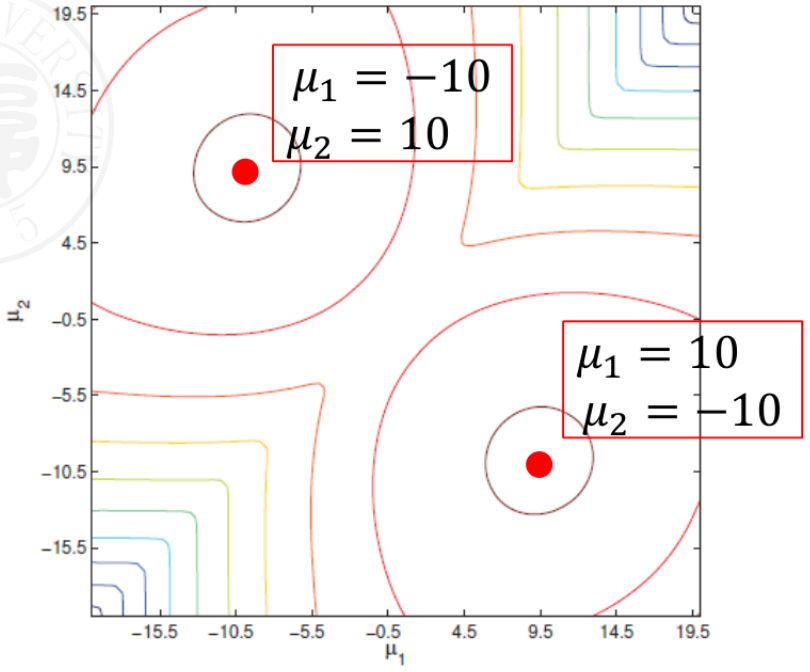
高斯混合模型

- 与生成分类器的差别
  - 训练阶段是否有类标签
    - 分类：已知类别信息y_i
    - 聚类：没有类别信息，使用隐变量z_i

参数估计

- 左图：2个一维高斯模型混合，采样200个点
- 右图：似然函数表面$\log p(\mathcal{D}|\mu_1, \mu_2)$





参数估计

- 极大似然估计 $\max_{\boldsymbol{\theta}} \log p(\mathcal{D}|\boldsymbol{\theta})$

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)$$

$$p(x_i, z_i|\theta) = p(z_i|\theta)p(x_i|z_i, \theta)$$

$$p(z_i = k|\theta) = \pi_k, \quad p(x_i|z_i = k, \theta) = \mathcal{N}(x_i|\mu_k, \Sigma_k)$$

$$\log p(\mathcal{D}|\boldsymbol{\theta}) = \sum_i \log \left[\sum_{\mathbf{z}_i} p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta}) \right]$$

- 最大后验概率估计 $\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathcal{D})$

$$\log p(\boldsymbol{\theta}|\mathcal{D}) \propto \log p(\mathcal{D}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

- 注意：这两个参数估计的解不唯一

  可能有 $K!$ 种不同的解（参见上一页右图，互换对称性）

EM算法

- 高斯混合模型隐变量的后验概率
  - 响应度（responsibility）r_{ik}
    - $p(z_i = k|\mathbf{x}_i, \theta)$：第i 个样本属于第k 个类的后验概率

$$r_{ik} \triangleq p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) = \frac{p(z_i = k|\boldsymbol{\theta})p(\mathbf{x}_i|z_i = k, \boldsymbol{\theta})}{\sum_{k'=1}^{K} p(z_i = k'|\boldsymbol{\theta})p(\mathbf{x}_i|z_i = k', \boldsymbol{\theta})}$$

EM算法

- E步：Q函数

$$
\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) &\triangleq \mathbb{E}\left[\sum_i \log p(\mathbf{x}_i, z_i|\boldsymbol{\theta})\right] \\
&= \sum_i \mathbb{E}\left[\log\left[\prod_{k=1}^{K}(\pi_k p(\mathbf{x}_i|\boldsymbol{\theta}_k))^{\mathbb{I}(z_i=k)}\right]\right] \\
&= \sum_i \sum_k \mathbb{E}\left[\mathbb{I}(z_i=k)\right] \log[\pi_k p(\mathbf{x}_i|\boldsymbol{\theta}_k)] \\
&= \sum_i \sum_k p(z_i=k|\mathbf{x}_i, \boldsymbol{\theta}^{t-1}) \log[\pi_k p(\mathbf{x}_i|\boldsymbol{\theta}_k)] \\
&= \sum_i \sum_k r_{ik} \log \pi_k + \sum_i \sum_k r_{ik} \log p(\mathbf{x}_i|\boldsymbol{\theta}_k)
\end{aligned}
$$

EM算法

- E步：Q函数
  - 计算

$$r_{ik} \triangleq p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}^{(t-1)})$$

$$r_{ik} = \frac{\pi_k p(\mathbf{x}_i|\boldsymbol{\theta}_k^{(t-1)})}{\sum_{k'} \pi_{k'} p(\mathbf{x}_i|\boldsymbol{\theta}_{k'}^{(t-1)})}$$

- M步：最大化Q函数
  - 类先验

$$\pi_k = \frac{1}{N}\sum_i r_{ik} = \frac{r_k}{N}$$

    - r_k：第k个类的加权数据量

EM算法

- M步：最大化Q函数
  - 第k个类的高斯模型参数

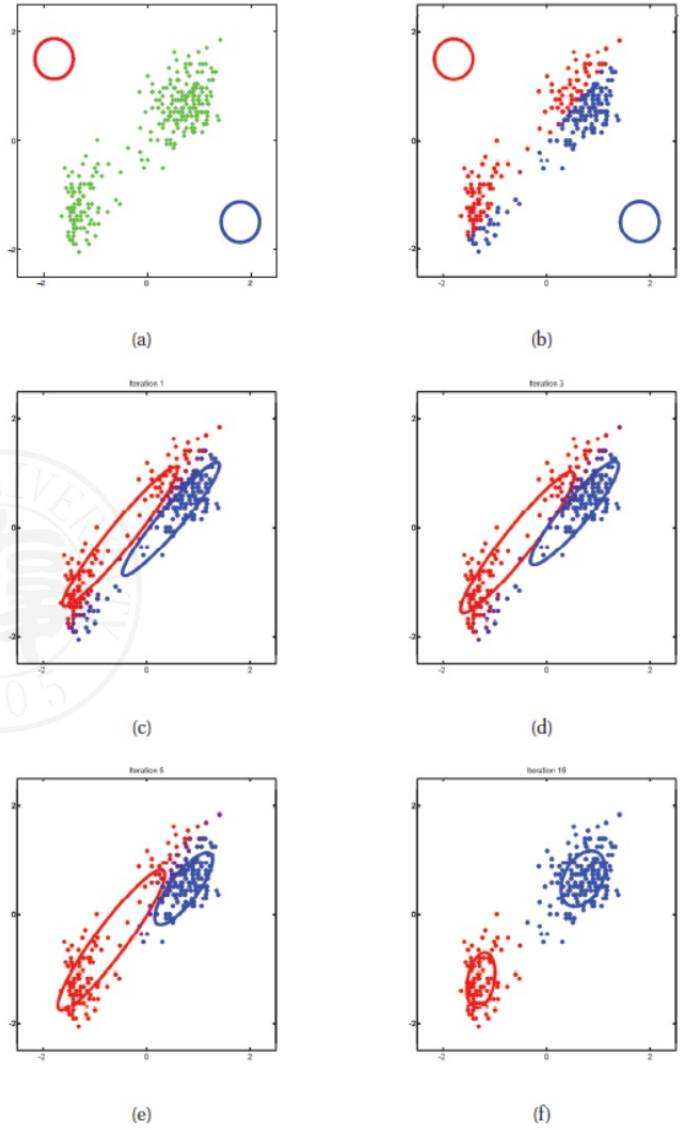
$$\boldsymbol{\mu}_k = \frac{\sum_i r_{ik}\mathbf{x}_i}{r_k}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_i r_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T}{r_k} = \frac{\sum_i r_{ik}\mathbf{x}_i\mathbf{x}_i^T}{r_k} - \boldsymbol{\mu}_k\boldsymbol{\mu}_k^T$$

  - 均值μ_k：第k个类的数据加权平均
  - 协方差Σ_k：第k个类的数据的加权散度矩阵

实例：EM算法

- (a)原始数据
- (b)模型参数赋初值
- (c)第1次迭代以后
- (d)第3次迭代以后
- (e)第5次迭代以后
- (f)第16次迭代以后



特例： K-means算法

- K-means算法是高斯混合模型的EM算法变种
  - 假设
    - 协方差矩阵$\Sigma_k = \sigma^2 I_D$已知
    - 类先验$\pi_k = 1/K$已知
    - 均值μ_k未知
  - EM算法
    - E步：delta-函数近似（硬分配）

$$z_i* = \text{argmax}_k \, p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta})$$

$$p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) \approx \mathbb{I}(k = z_i^*)$$

特例： K-means算法

- E步：计算z_i^*

$$\max_k p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta})$$

$$= \max_k \frac{p(z_i = k|\boldsymbol{\theta})p(\mathbf{x}_i|z_i=k,\boldsymbol{\theta})}{\sum_{k'=1}^{K} p(z_i = k'|\boldsymbol{\theta})p(\mathbf{x}_i|z_i=k',\boldsymbol{\theta})}$$

$$= \max_k p(z_i = k|\boldsymbol{\theta})p(\mathbf{x}_i|z_i = k, \boldsymbol{\theta})$$

$$= \max_k \frac{1}{K} \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_i - \mu_k)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mu_k)\right]$$

$$= \min_k (\mathbf{x}_i - \mu_k)^T (\mathbf{x}_i - \mu_k)$$

特例： K-means算法

- E步：计算z_i^*

$$z_i^* = \arg\min_k (\mathbf{x}_i - \mu_k)^T (\mathbf{x}_i - \mu_k)$$

$$= \arg\min_k \|\mathbf{x}_i - \mu_k\|_2^2$$

- M步：计算均值μ_k

$$\boldsymbol{\mu}_k = \frac{\sum_i r_{ik}\mathbf{x}_i}{r_k} = \frac{1}{N_k}\sum_{i:z_i=k}\mathbf{x}_i$$

特例：K-means算法

- 算法伪代码

Algorithm 11.1: K-means algorithm

initialize $\boldsymbol{\mu}_k$;
repeat
    Assign each data point to its closest cluster center: $z_i = \arg\min_k ||\mathbf{x}_i - \boldsymbol{\mu}_k||_2^2$;
    Update each cluster center by computing the mean of all points assigned to it: $\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i:z_i=k} \mathbf{x}_i$;
until *converged*;

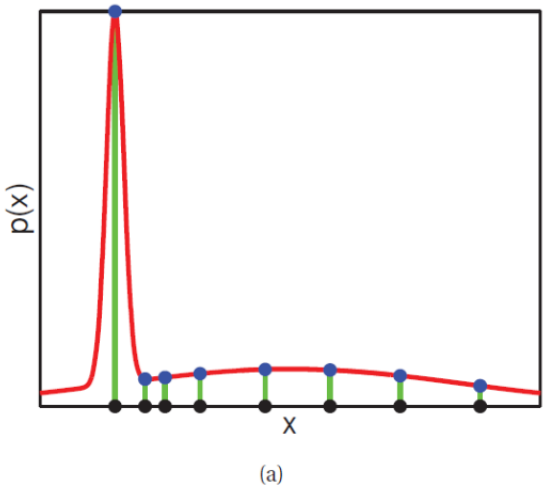
特例：K-means算法

- 高斯混合模型相比于K-means算法的优点：
  - 计算一个样本属于某类的概率值；
  - 不仅仅可以用于聚类，还可以用于概率密度估计；
  - 可以用于生成新的样本点

极大似然估计：过拟合现象

- 实例
  - K = 2
  - 协方差矩阵$\Sigma_k = \sigma_k^2 I_D$
  - EM算法初始化：$\mu_2 = \mathbf{x}_1$

$$\log p(\mathcal{D}|\boldsymbol{\theta}) = \sum_i \log \left[\sum_{\mathbf{z}_i} p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta}) \right]$$



$$\mathcal{N}(\mathbf{x}_1|\mu_2, \sigma_2^2 I_D) = \frac{1}{(2\pi)^{D/2}|\sigma_2^2 I_D|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_1 - \mu_2)^T (\sigma_2^2 I_D)^{-1}(\mathbf{x}_1 - \mu_2)\right]$$
$$= \frac{1}{(2\pi\sigma_2^2)^{D/2}}$$

- 当$\sigma_2 \to 0$时，$\mathcal{N}(\mathbf{x}_1|\mu_2, \sigma_2^2 I_D) \to \infty$

只要让某个成分收缩到一个数据点上，似然就可以无限增大。所谓的“最大似然解”根本不存在，会产生方差趋零的退化成分。

最大后验概率估计

- 后验概率

$$\log p(\boldsymbol{\theta}|\mathcal{D}) \propto \log p(\mathcal{D}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

- EM算法
  - E步：

$$Q'(\boldsymbol{\theta},\boldsymbol{\theta}^{(t-1)}) = \left[\sum_i \sum_k r_{ik} \log \pi_k + \sum_i \sum_k r_{ik} \log p(\mathbf{x}_i|\boldsymbol{\theta}_k)\right] + \log p(\boldsymbol{\pi}) + \sum_k \log p(\boldsymbol{\theta}_k)$$

    - 计算

$$r_{ik} \triangleq p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}^{(t-1)})$$

$$r_{ik} = \frac{\pi_k p(\mathbf{x}_i|\boldsymbol{\theta}_k^{(t-1)})}{\sum_{k'} \pi_{k'} p(\mathbf{x}_i|\boldsymbol{\theta}_{k'}^{(t-1)})}$$

最大后验概率估计

$$\mathcal{Q}'(\boldsymbol{\theta},\boldsymbol{\theta}^{(t-1)}) = \left[\sum_i \sum_k r_{ik} \log \pi_k + \sum_i \sum_k r_{ik} \log p(\mathbf{x}_i|\boldsymbol{\theta}_k)\right] + \log p(\boldsymbol{\pi}) + \sum_k \log p(\boldsymbol{\theta}_k)$$

- M步：
  - 类先验 $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$

$$\pi_k = \frac{r_k + \alpha_k - 1}{N + \sum_k \alpha_k - K}$$

  - 第k个高斯模型的先验

$$p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \mathrm{NIW}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\mathbf{m}_0, \kappa_0, \nu_0, \mathbf{S}_0)$$

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_k &= \frac{r_k \overline{\mathbf{x}}_k + \kappa_0 \mathbf{m}_0}{r_k + \kappa_0} \\
\overline{\mathbf{x}}_k &\triangleq \frac{\sum_i r_{ik} \mathbf{x}_i}{r_k} \\
\hat{\boldsymbol{\Sigma}}_k &= \frac{\mathbf{S}_0 + \mathbf{S}_k + \frac{\kappa_0 r_k}{\kappa_0 + r_k}(\overline{\mathbf{x}}_k - \mathbf{m}_0)(\overline{\mathbf{x}}_k - \mathbf{m}_0)^T}{\nu_0 + r_k + D + 2} \\
\mathbf{S}_k &\triangleq \sum_i r_{ik}(\mathbf{x}_i - \overline{\mathbf{x}}_k)(\mathbf{x}_i - \overline{\mathbf{x}}_k)^T
\end{aligned}$$

最大后验概率估计

- 特殊的超参数设置
  - 类先验 $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$

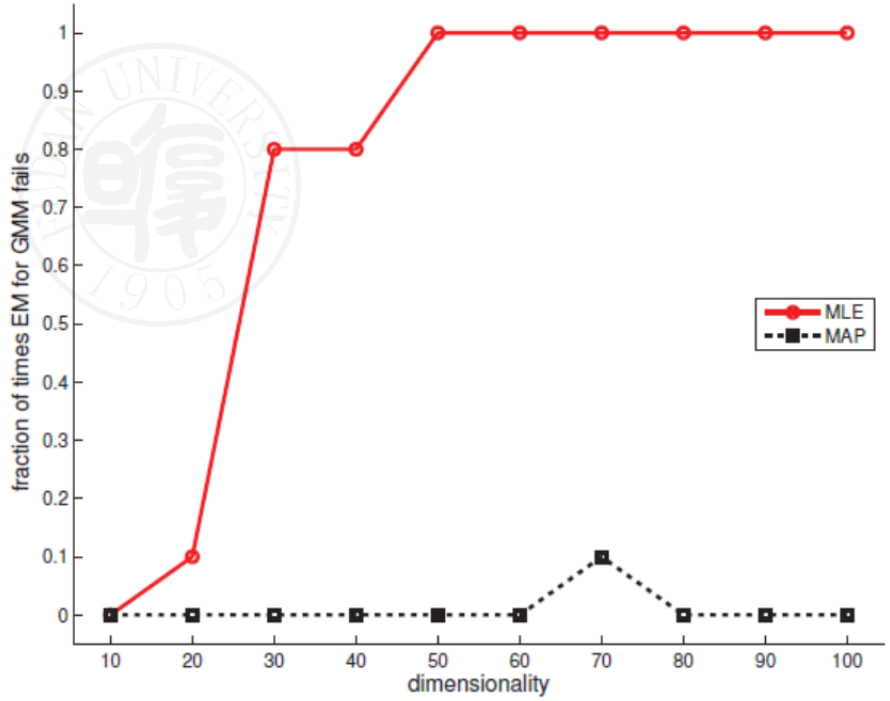
    $\alpha_k = 1, k = 1, \ldots, K$
  - 第k个高斯模型的先验
    - $\kappa_0 = 0$ $\mathbf{S}_0 = \frac{1}{K^{1/D}}\mathrm{diag}(s_1^2, \ldots, s_D^2)$ $s_j = (1/N)\sum_{i=1}^{N}(x_{ij} - \overline{x}_j)^2$ $\nu_0 = D + 2$

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_i r_{ik}\mathbf{x}_i}{r_k}$$

$$\hat{\boldsymbol{\Sigma}}_k = \frac{\mathbf{S}_0 + \mathbf{S}_k}{\nu_0 + r_k + D + 2}$$

MLE vs. MAP

- 比较极大似然估计和最大后验概率估计实验失败的次数
  - 使用合成数据
  - N=100
  - 5次重复实验

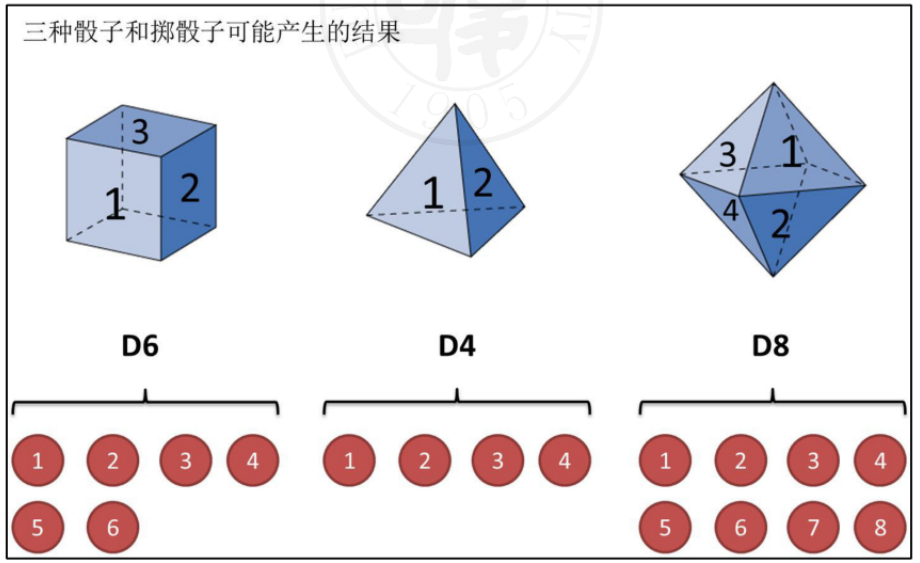


目录

- 生成分类器
- 单高斯模型
- 高斯混合模型
- 隐马尔可夫模型

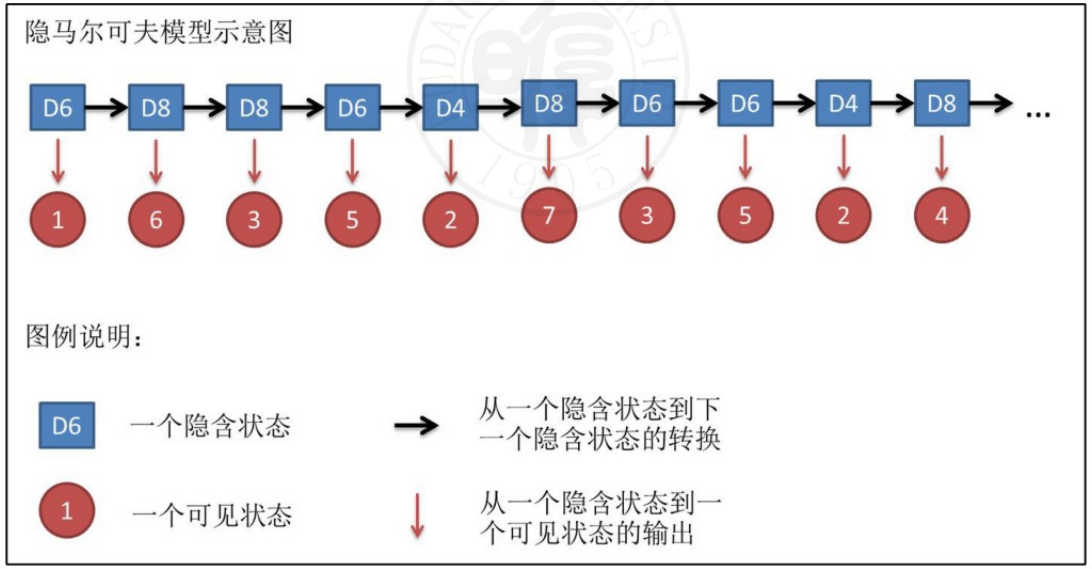
实例：扔骰子游戏

- 假设：有3个骰子 D6、D4、D8
  - 一开始，随机任选一个骰子；之后每次的骰子类型，依赖于上一次选的骰子
  - 掷骰子，得到1~8中的一个数字
- 重复N（N=10）次，得到一组数字1 6 3 5 2 7 3 5 2 4



实例：扔骰子游戏

- 可见状态链 x：投掷结果序列
- 隐含状态链 z：选用骰子的序列



实例：扔骰子游戏

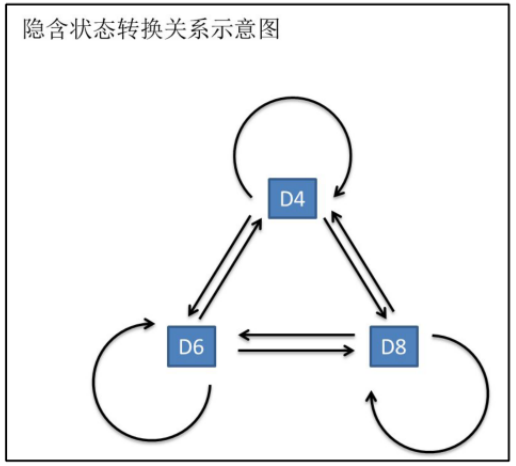
- 可用两种概率关系描述
  - 隐含状态之间：状态转移概率
  - 隐含与可见状态之间：观测概率
    - E.g., 一次投掷，获得1的概率

$$p(\mathbf{x}=1)=\sum_z p(z)p(\mathbf{x}|z)=\frac{1}{3}\times\frac{1}{6}+\frac{1}{3}\times\frac{1}{4}+\frac{1}{3}\times\frac{1}{8}$$

    - E.g., 两次投掷，获得1、6的概率

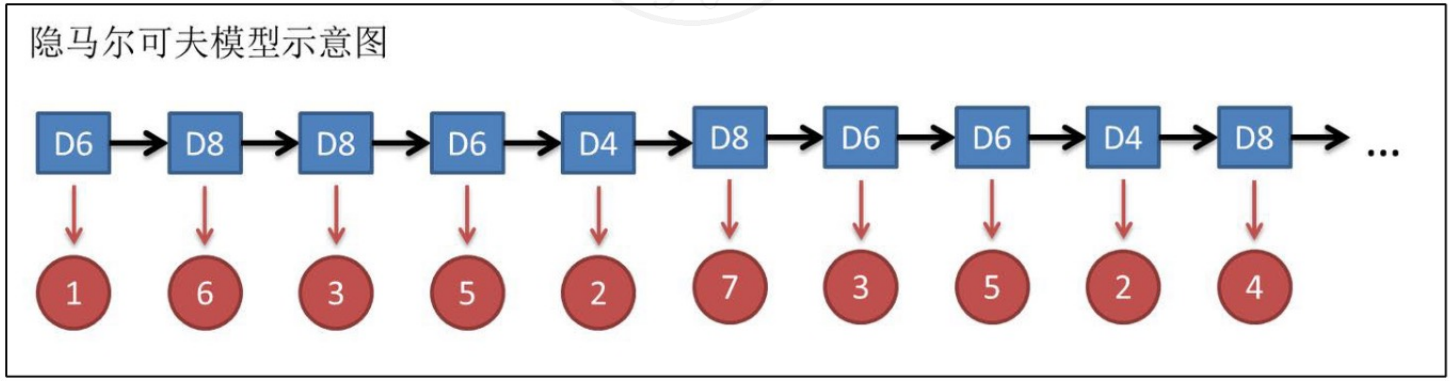
$$\begin{aligned}&p(\mathbf{x}=1\,6)=p(\mathbf{x}_1=1,\mathbf{x}_2=6)\\&=\sum_{z_1}\sum_{z_1}p(z_1)p(\mathbf{x}_1=1|z_1)p(z_2|z_1)p(\mathbf{x}_2=6|z_2)\end{aligned}$$

隐状态的切换



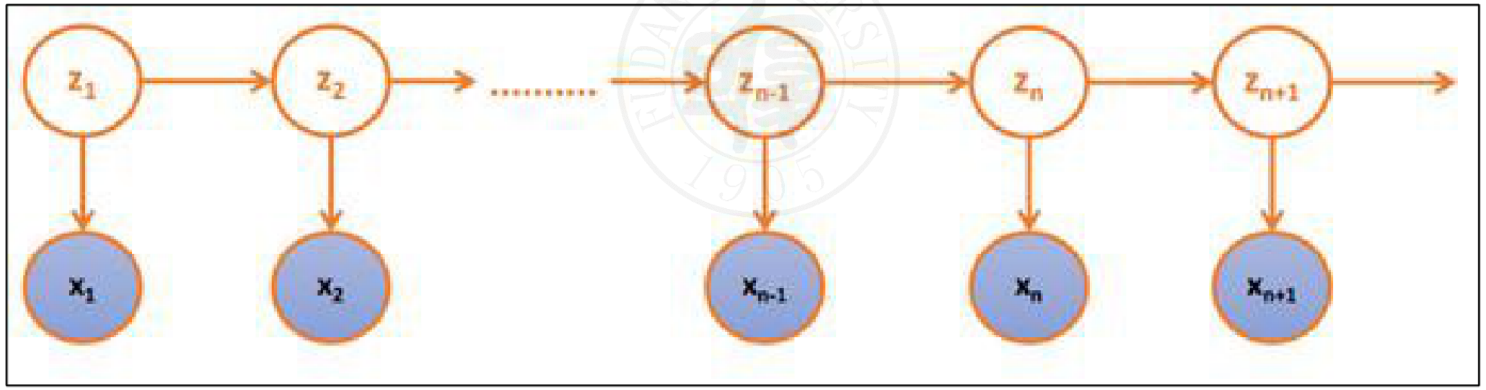
隐马尔可夫模型

- 隐马尔可夫模型 (Hidden Markov Model, HMM)
  - 关于时序的概率模型
  - 描述由一个隐藏的马尔可夫链随机生成不可观测的状态随机序列，再由各个状态生成一个观测，从而产生观测随机序列的过程
  - 序列的每个位置可看作一个时刻



隐马尔可夫模型

- HMM的概率图模型表示
  - 隐变量 z，观测变量 x



隐马尔可夫模型

- HMM模型参数：$\theta = (\pi, A, B)$
  - 隐含状态的初始概率分布π
  - 隐含状态的状态转移概率矩阵A

$$A(i,j) = p(z_t = j | z_{t-1} = i)$$

  - 观测概率矩阵B

$$p(\mathbf{x}_t | z_t = j)$$

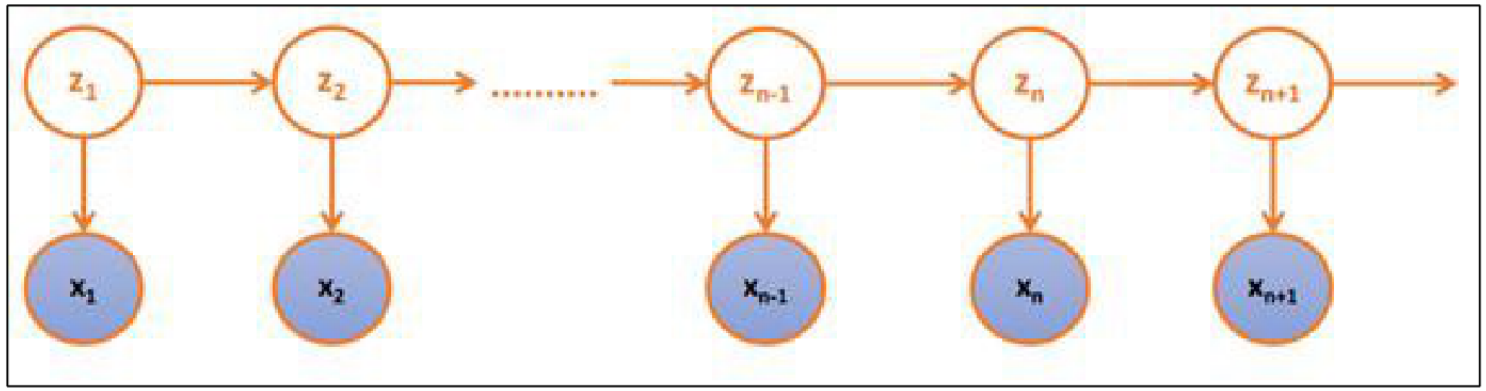
  - 依赖关系
    - 初始状态概率向量π和状态转移概率矩阵A决定状态序列z
    - 观测概率矩阵B决定观测序列x

隐马尔可夫模型

- 齐次马尔可夫性假设
  - 隐马尔可夫链 t 时刻的状态只和 $t-1$ 时刻的状态有关
  - 与其他时刻的状态及观测无关，也与时刻 t 无关

$$p(z_{1:T}) = p(z_1)p(z_2|z_1)p(z_3|z_1, z_2) \dots \dots p(z_T|z_1, \dots, z_{T-2}, z_{T-1})$$

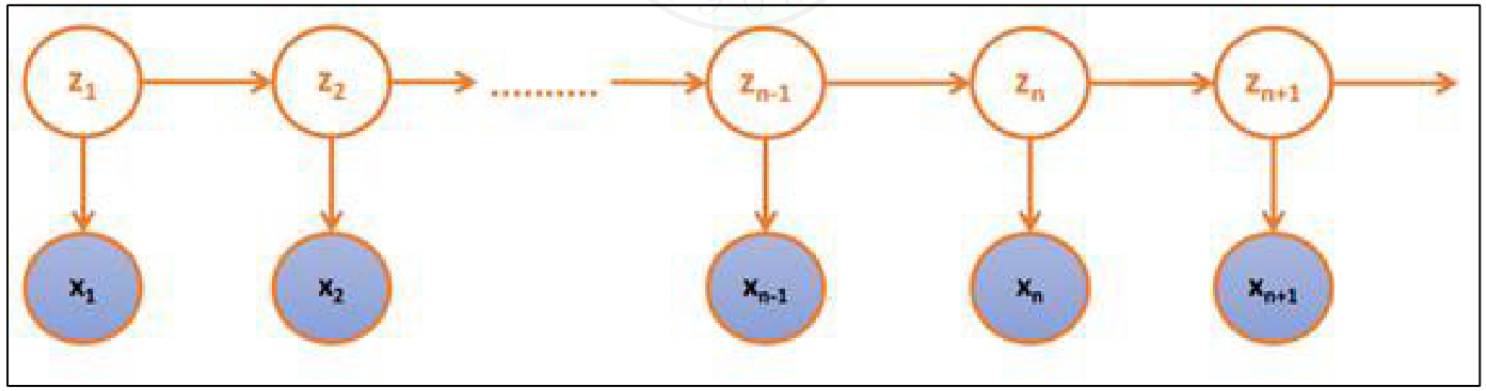
$$= p(z_1)p(z_2|z_1)p(z_3|z_2) \dots \dots p(z_T|\, z_{T-1})$$



隐马尔可夫模型

- 观测独立性假设
  - 观测变量只和当前时刻的状态有关，与其他时刻的观测和状态均无关

$$p(\mathrm{x}_{1:T}|z_{1:T}) = p(x_1|z_1)p(x_2|z_2)\ldots p(x_T|z_T)$$



隐马尔可夫模型

- 隐含状态链和观测序列的联合分布

$$p(\mathbf{z}_{1:T}, \mathbf{x}_{1:T}) = p(\mathbf{z}_{1:T})p(\mathbf{x}_{1:T}|\mathbf{z}_{1:T}) = \left[p(z_1)\prod_{t=2}^{T} p(z_t|z_{t-1})\right]\left[\prod_{t=1}^{T} p(\mathbf{x}_t|z_t)\right]$$

  - 离散状态空间
  - 离散或连续的观测变量
    - 离散观测变量：发射概率矩阵

$$p(\mathbf{x}_t = l|z_t = k, \boldsymbol{\theta}) = B(k, l)$$

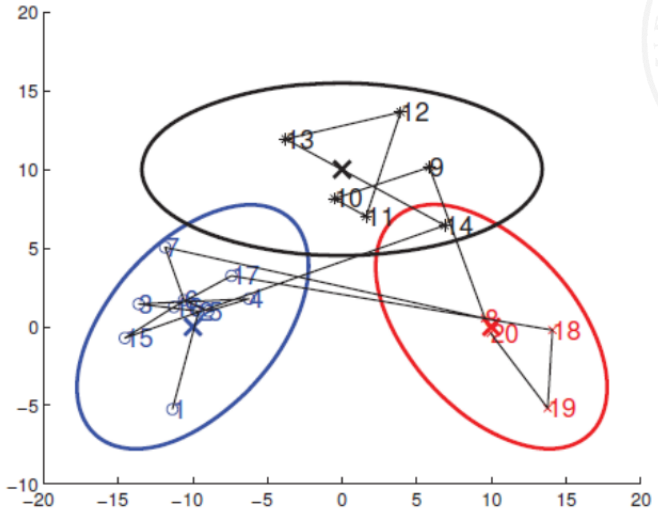
	X	Y	Z
z_1	0.4	0.1	0.5
z_2	0.1	0.5	0.4

    - 连续观测变量：高斯分布

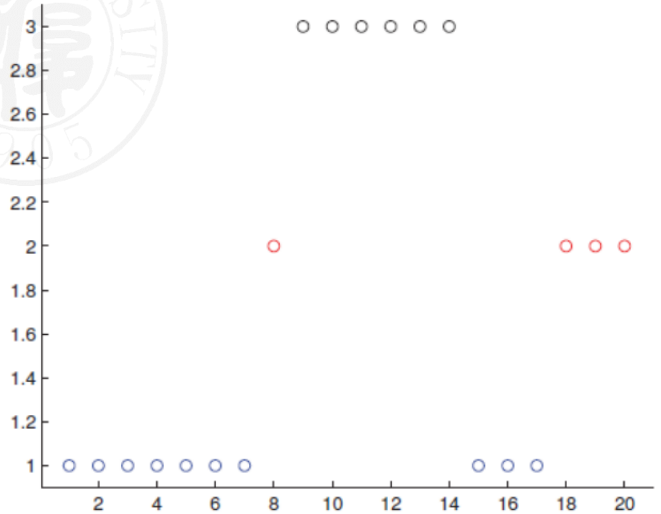
$$p(\mathbf{x}_t|z_t = k, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

隐马尔可夫模型

- 例子：
  - 从3个隐含状态采样的20个观测数据
  - 对应的隐含状态序列



(a)



(b)

隐马尔可夫模型的应用

- 在时间序列中建立观测与隐藏语义间的映射关系
- 语音识别
  - x：从语音信号中提取的声学特征
  - z：发出的音素或单词
- 行为识别
  - x：从视频帧中提取的姿态或运动特征
  - z：人的行为（如跑、走、坐等）

模型学习与推断

- 概率计算问题
  - 给定模型参数 θ 和观测序列 $\mathrm{x}_{1:T}$，计算生成该观测序列的概率 $p(\mathrm{x}_{1:T}|\theta)$
    - 前向算法：$p(z_t|\mathrm{x}_{1:t})$ —— Filtering（在线学习）
    - 后向算法：$p(z_t|\mathrm{x}_{1:T})$ —— Smoothing（离线学习）
- 学习问题
  - 给定观测序列 $\mathrm{x}_{1:T}$，学习使生成该序列的概率最大的模型参数 θ
    $\max p(\mathrm{x}_{1:T}|\theta)$
- 预测问题（解码问题）
  - 给定模型参数 θ 和观测序列 $\mathrm{x}_{1:T}$，求最有可能的隐含状态序列
    $\max p(\mathrm{z}_{1:T}|\mathrm{x}_{1:T},\theta)$

概率计算问题

- 前向算法
  - 预测：一步前预测密度 (one-step-ahead predictive density)

$$p(z_t = j|x_{1:t-1}, \theta) = \sum_i p(z_t = j|z_{t-1} = i, \theta) p(z_{t-1} = i|x_{1:t-1}, \theta)$$

$\alpha_{t-1}(i)$

  - 更新：t 时刻的信念状态 (belief state)

$$\begin{aligned}\alpha_t(j) \triangleq p(z_t = j|x_{1:t}, \theta) &= p(z_t = j|x_t, x_{1:t-1}, \theta) \\ &= \frac{1}{Z_t} p(x_t|z_t = j, x_{1:t-1}, \theta) p(z_t = j|x_{1:t-1}, \theta) \\ &= \frac{1}{Z_t} p(x_t|z_t = j, \theta) p(z_t = j|x_{1:t-1}, \theta),\end{aligned}$$

概率计算问题

- 前向算法
  - 更新：t 时刻的信念状态

$$\alpha_t(j) \propto p(x_t|z_t=j,\theta)p(z_t=j|x_{1:t-1},\theta)$$

$$= p(x_t|z_t=j,\theta)\sum_i p(z_t=j|z_{t-1}=i,\theta)p(z_{t-1}=i|x_{1:t-1},\theta)$$

$$= p(x_t|z_t=j,\theta)\sum_i p(z_t=j|z_{t-1}=i,\theta)\alpha_{t-1}(i)$$

$$\boldsymbol{\alpha}_t \propto \boldsymbol{\psi}_t \odot (\boldsymbol{\Psi}^T \boldsymbol{\alpha}_{t-1})$$

$$\psi_t(j) = p(x_t|z_t=j,\theta)$$

$$\Psi(i,j) = p(z_t=j|z_{t-1}=i)$$

概率计算问题

- 前向算法：观测序列概率计算
  - 计算$p(\mathbf{x}_{1:T}|\theta)$

$$\log p(x_{1:T}|\theta) = \log\left[p(x_1|\theta)\prod_{t=2}^{T} p(x_t|x_{1:t-1},\theta)\right]$$

$$= \log p(x_1|\theta) + \sum_{t=2}^{T} \log p(x_t|x_{1:t-1},\theta)$$

$$= \log \sum_{i=1}^{K} p(z_1 = i|\theta)p(x_1|z_1 = i,\theta) + \sum_{t=2}^{T} \log Z_t$$

$$Z_t \triangleq p(x_t|x_{1:t-1},\theta) = \sum_j p(z_t = j|x_{1:t-1},\theta)p(x_t|z_t = j,\theta)$$

一步前预测密度×当前观测

概率计算问题

- 隐状态后验概率计算
  - 基本思想

$$
\begin{aligned}
p(z_t = j | x_{1:T}, \theta) &= p(z_t = j | x_{1:t}, x_{t+1:T}, \theta) \\
&= \frac{p(z_t = j, x_{t+1:T} | x_{1:t}, \theta)}{p(x_{t+1:T} | x_{1:t}, \theta)} \\
&= \frac{p(x_{t+1:T} | z_t = j, x_{1:t}, \theta) p(z_t = j | x_{1:t}, \theta)}{p(x_{t+1:T} | x_{1:t}, \theta)} \\
&\propto p(z_t = j | x_{1:t}, \theta) p(x_{t+1:T} | z_t = j, x_{1:t}, \theta) \\
&\propto p(z_t = j | x_{1:t}, \theta) p(x_{t+1:T} | z_t = j, \theta).
\end{aligned}
$$

过去：信念状态（已知）

未来：后向变量（后向算法的计算目标）

这个计算很重要，在学习问题的EM算法会使用到。

概率计算问题

- 后向算法
  - 定义
    - 后向变量

$$\beta_t(j) = p(x_{t+1:T}|z_t = j, \theta)$$

    - 平滑后的后验分布

$$\gamma_t(j) = p(z_t = j|x_{1:T}, \theta)$$

    - 因此，

$$\gamma_t(j) \propto \alpha_t(j)\beta_t(j)$$

概率计算问题

- 后向算法
  - 计算后向变量

$$
\begin{aligned}
\beta_{t-1}(i) &= p(x_{t:T}|z_{t-1}=i,\theta) \\
&= \sum_j p(z_t=j, x_t, x_{t+1:T}|z_{t-1}=i,\theta) \\
&= \sum_j p(x_{t+1:T}|z_t=j, \cancel{x_t}, \cancel{z_{t-1}}=i,\theta)p(z_t=j,x_t|z_{t-1}=i,\theta) \\
&= \sum_j p(x_{t+1:T}|z_t=j,\theta)p(x_t|z_t=j,\cancel{z_{t-1}}=i,\theta)p(z_t=j|z_{t-1}=i,\theta) \\
&= \sum_j \beta_t(j)\psi_t(j)\Psi(i,j)
\end{aligned}
\quad \Rightarrow \quad \boxed{\boldsymbol{\beta}_{t-1} = \Psi(\boldsymbol{\psi}_t \odot \boldsymbol{\beta}_t)}
$$

概率计算问题

- 后向算法
  - 从 $t+1$ 逆推
  - 最后时刻 T，初始化

$$\beta_T(i) = p(x_{T+1}|z_T = i, \theta) = p(\emptyset|z_T = i, \theta) = 1$$

概率计算问题

- 后向算法：观测序列概率计算
  - 计算$p(\mathbf{x}_{1:T}|\theta)$

$$p(\mathrm{x}_{1:T}|\theta) = \sum_j p(z_1 = j|\theta)p(\mathrm{x}_{1:T}|z_1 = j, \theta)$$

$$= \sum_j p(z_1 = j|\theta)p(\mathrm{x}_1|z_1 = j, \theta)p(\mathrm{x}_{2:T}|\mathrm{x}_1, z_1 = j, \theta)$$

$$= \sum_j p(z_1 = j|\theta)p(\mathrm{x}_1|z_1 = j, \theta)\beta_1(j)$$

初始状态 × 当前观测 × 未来证据

学习问题

- 估计模型参数 $\theta = (\boldsymbol{\pi}, A, B)$
  - 隐含状态的初始概率分布 $\boldsymbol{\pi}$
  - 状态转移概率矩阵 $A = \Psi$

$$A(i,j) = p(z_t = j | z_{t-1} = i)$$

  - 观测概率矩阵 B
    - 离散观测：概率表
    - 连续观测：类条件概率密度函数 $p(x_t | z_t = j, \theta)$

  - 目标：$\arg\max_{\theta} p(x_{1:T} | \theta)$

学习问题

- 如果 z 可观测
  - 监督学习方法
  - 极大似然估计
- 如果 z 不可观测
  - 非监督学习方法
  - Baum-Welch 算法（EM算法）

极大似然估计

- 估计 $\boldsymbol{\pi}, A$
  - 给定观测数据集

$$\mathcal{D} = \{(\mathbf{z}_i, \mathbf{x}_i), i = 1, 2, \cdots, N | \mathbf{z}_i = z_{i,1:T_i}, \mathbf{x}_i = x_{i,1:T_i}\}$$

  - 生成1个长度为 T 的状态序列数据 $z_{1:T}$ 的似然函数

$$\begin{aligned} p(z_{1:T}|\theta) &= p(z_1)p(z_2|z_1)\cdots p(z_T|z_{T-1}) \\ &= \pi(z_1)\Psi(z_1, z_2)\cdots\Psi(z_{T-1}, z_T) \\ &= \prod_{j=1}^{K}(\pi_j)^{\mathbb{I}[z_1=j]}\prod_{t=2}^{T}\prod_{j=1}^{K}\prod_{k=1}^{K}(\Psi_{j,k})^{\mathbb{I}[z_{t-1}=j,z_t=k]} \end{aligned}$$

极大似然估计

- 估计 $\boldsymbol{\pi}, A$
  - 生成N个状态序列数据的对数似然函数
    - 记N个状态序列数据的集合为 $\mathcal{D}_Z$

$$\mathcal{D}_z = \{\mathbf{z}_1, \cdots, \mathbf{z}_N\}$$

    - 模型参数 θ 生成状态序列集 $\mathcal{D}_Z$ 的对数似然函数为所有样本似然的和

$$\begin{aligned}
\log p(\mathcal{D}_z|\theta) &= \sum_{i=1}^{N} \log p(\mathbf{z}_i|\theta) \\
&= \sum_{i=1}^{N} \log \left[\prod_{j=1}^{K} (\pi_j)^{\mathbb{I}[z_{i1}=j]} \prod_{t=2}^{T_i} \prod_{j=1}^{K} \prod_{k=1}^{K} (\Psi_{j,k})^{\mathbb{I}[z_{i,t-1}=j, z_{it}=k]} \right] \\
&= \sum_{i=1}^{N} \sum_{j=1}^{K} \mathbb{I}[z_{i1}=j] \log \pi_j + \sum_{i=1}^{N} \sum_{t=2}^{T_i} \sum_{j=1}^{K} \sum_{k=1}^{K} \mathbb{I}[z_{i,t-1}=j, z_{it}=k] \log \Psi_{j,k}
\end{aligned}$$

极大似然估计

- 估计 $\boldsymbol{\pi}, A$
  - 定义初始状态计数 N_j^1、状态转移计数 N_j^k

$$N_j^1 = \sum_{i=1}^{N} \mathbb{I}[z_{i1} = j], \quad N_{jk} = \sum_{i=1}^{N} \sum_{t=2}^{T_i} \mathbb{I}[z_{i,t-1} = j, z_{it} = k]$$

  - 参数估计

$$\hat{\pi}_j = \frac{N_j^1}{\sum_j N_j^1}, \quad \hat{A}_{jk} = \frac{N_{jk}}{\sum_k N_{jk}}$$

极大似然估计

- 估计 B
  - 如果观测变量 x 是离散的
    - 假设观测变量 x 服从多项分布

$$B(k,l) = p(x_t = l | z_t = k, \theta), k = 1, 2, \cdots, K, l = 1, 2, \cdots, L$$

    - 参数估计

$$\hat{B}_{kl} = \frac{N_{kl}^X}{N_k}$$

$$N_{kl}^X = \sum_{i=1}^{N} \sum_{t=1}^{T_i} \mathbb{I}[z_{i,t} = k, x_{i,t} = l]$$

$$N_k = \sum_{i=1}^{N} \sum_{t=1}^{T_i} \mathbb{I}[z_{i,t} = k]$$

      - N_{kl}^X表示在整个数据集D中，状态k生成观测变量l的出现次数
      - N_k表示在整个数据集D中，状态k出现的次数

极大似然估计

- 估计 B
  - 如果观测变量 x 是连续的
    - 假设观测变量 x 服从高斯分布

$$p(x_t|z_t=k,\theta)=\mathcal{N}(x_t|\mu_k,\Sigma_k)$$

    - 参数估计

$$\hat{\boldsymbol{\mu}}_k=\frac{\overline{\mathbf{x}}_k}{N_k},\quad \hat{\boldsymbol{\Sigma}}_k=\frac{(\overline{\mathbf{xx}})_k^T-N_k\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T}{N_k}$$

$$\overline{\mathbf{x}}_k \triangleq \sum_{i=1}^{N}\sum_{t=1}^{T_i}\mathbb{I}(z_{i,t}=k)\mathbf{x}_{i,t}$$

$$(\overline{\mathbf{xx}})_k^T \triangleq \sum_{i=1}^{N}\sum_{t=1}^{T_i}\mathbb{I}(z_{i,t}=k)\mathbf{x}_{i,t}\mathbf{x}_{i,t}^T$$

      - $\hat{\mu}_k$表示在整个数据集D中，由状态k生成的观测变量x_t的平均值
      - $\hat{\Sigma}_k$表示在整个数据集D中，由状态k生成的观测变量x_t的经验协方差矩阵

Baum-Welch算法

- 多数实际情况：
  - 隐状态 z 不可观测，MLE难以直接求解
- EM算法，先补充介绍后验隐状态联合概率计算

$$\xi_t(j,k;\theta^{\text{old}}) = P(z_t = j,\ z_{t+1} = k \mid x_{1:T};\theta^{\text{old}})$$

$\xi_t(j,k) = \frac{P(z_t=j, z_{t+1}=k, x_{1:T})}{P(x_{1:T})}$ 把观测拆成三段：$x_{1:t}$，x_{t+1}，$x_{t+2:T}$

$$P(z_t = j, z_{t+1} = k, x_{1:T}) = P(x_{1:t}, z_t = j)\ P(z_{t+1} = k \mid z_t = j)\ P(x_{t+1} \mid z_{t+1} = k)\ P(x_{t+2:T} \mid z_{t+1} = k)$$

$$P(z_t = j, z_{t+1} = k, x_{1:T}) = P(x_{1:t}, z_t = j)\ A^{\text{old}}_{jk}\ B^{\text{old}}_k(x_{t+1})\ \beta_{t+1}(k) = \alpha_t(j)\left(\prod_{s=1}^{t} c_s\right) A^{\text{old}}_{jk} B^{\text{old}}_k(x_{t+1})\beta_{t+1}(k)$$

$$\xi_t(j,k) \propto \alpha_t(j)\ A^{\text{old}}_{jk}\ B^{\text{old}}_k(x_{t+1})\ \beta_{t+1}(k)$$

Baum-Welch算法

- EM算法：为避免过于繁琐的符号，这里仅介绍一个观测序列的情形，多观测序列可类似拓展
- E步

后验下的对数似然期望

$$P(x_{1:T}, z_{1:T}; \theta) = P(z_1; \theta) \prod_{t=1}^{T-1} P(z_{t+1} \mid z_t; \theta) \prod_{t=1}^{T} P(x_t \mid z_t; \theta)$$

$$= \pi_{z_1} \prod_{t=1}^{T-1} A_{z_t, z_{t+1}} \prod_{t=1}^{T} B_{z_t}(x_t)$$

$$\log P(x_{1:T}, z_{1:T}; \theta) = \log \pi_{z_1} + \sum_{t=1}^{T-1} \log A_{z_t, z_{t+1}} + \sum_{t=1}^{T} \log B_{z_t}(x_t)$$

然后取关于 $z_{1:T} \mid x_{1:T}; \theta^{\text{old}}$ 的期望，得到 Q 函数：

$$Q(\theta, \theta^{\text{old}}) = \mathbb{E}_{z_{1:T} \mid x_{1:T}; \theta^{\text{old}}} \left[\log \pi_{z_1} \right]$$

$$+ \mathbb{E}_{z_{1:T} \mid x_{1:T}; \theta^{\text{old}}} \left[\sum_{t=1}^{T-1} \log A_{z_t, z_{t+1}} \right]$$

$$+ \mathbb{E}_{z_{1:T} \mid x_{1:T}; \theta^{\text{old}}} \left[\sum_{t=1}^{T} \log B_{z_t}(x_t) \right]$$

Baum-Welch算法

- EM算法
- E步

1. $\mathbb{E}\left[\log \pi_{z_1}\right] = \sum_j \gamma_1(j;\theta^{\text{old}}) \log \pi_j$

2. $\mathbb{E}\left[\sum_{t=1}^{T-1} \log A_{z_t,z_{t+1}}\right] = \sum_{t=1}^{T-1} \sum_j \sum_k \xi_t(j,k;\theta^{\text{old}}) \log A_{jk}$

3. $\mathbb{E}\left[\sum_{t=1}^{T} \log B_{z_t}(x_t)\right] = \sum_{t=1}^{T} \sum_j \gamma_t(j;\theta^{\text{old}}) \log B_j(x_t)$

合并后：

$$
\begin{aligned}
Q(\theta,\theta^{\text{old}}) = & \sum_j \gamma_1(j;\theta^{\text{old}}) \log \pi_j \\
& + \sum_{t=1}^{T-1} \sum_j \sum_k \xi_t(j,k;\theta^{\text{old}}) \log A_{jk} \\
& + \sum_{t=1}^{T} \sum_j \gamma_t(j;\theta^{\text{old}}) \log B_j(x_t)
\end{aligned}
$$

Baum-Welch算法

- EM算法
- M步 固定 γ, ξ，最大化 $Q(\theta, \theta^{\text{old}})$

$$Q(\theta, \theta^{\text{old}}) = \sum_j \gamma_1(j) \log \pi_j + \sum_{t=1}^{T-1} \sum_j \sum_k \xi_t(j,k) \log A_{jk} + \sum_{t=1}^{T} \sum_j \gamma_t(j) \log B_j(x_t)$$

拉格朗日乘数法求导可得：

$$\pi_j^{\text{new}} = \gamma_1(j; \theta^{\text{old}})$$

离散观测情形：

$$A_{jk}^{\text{new}} = \frac{\sum_{t=1}^{T-1} \xi_t(j,k;\theta^{\text{old}})}{\sum_{t=1}^{T-1} \sum_{k'} \xi_t(j,k';\theta^{\text{old}})} \qquad B_j^{\text{new}}(v) = \frac{\sum_{t=1}^{T} \mathbf{1}\{x_t = v\} \cdot \gamma_t(j;\theta^{\text{old}})}{\sum_{t=1}^{T} \gamma_t(j;\theta^{\text{old}})}$$

连续观测情形：

$$\boldsymbol{\mu}_j^{\text{new}} = \frac{\sum_{t=1}^{T} \gamma_t(j;\theta^{\text{old}}) \boldsymbol{x}_t}{\sum_{t=1}^{T} \gamma_t(j;\theta^{\text{old}})} \qquad \Sigma_j^{\text{new}} = \frac{\sum_{t=1}^{T} \gamma_t(j;\theta^{\text{old}}) \left(\boldsymbol{x}_t - \boldsymbol{\mu}_j^{\text{new}}\right) \left(\boldsymbol{x}_t - \boldsymbol{\mu}_j^{\text{new}}\right)^T}{\sum_{t=1}^{T} \gamma_t(j;\theta^{\text{old}})}$$

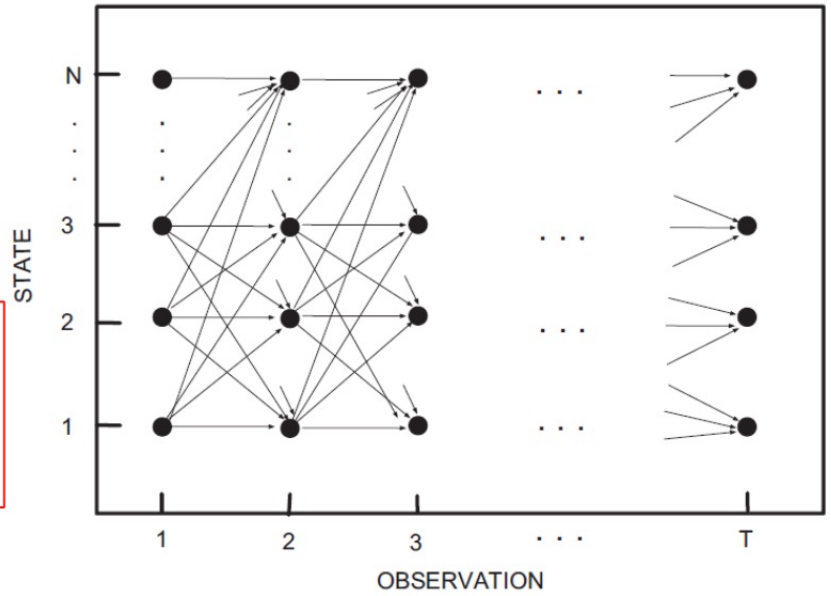
预测问题

- 给定模型参数和观测数据，找到最有可能的状态序列

$$\mathbf{z}^* = \arg\max_{\mathbf{z}_{1:T}} p(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$$

  - 等价于最大路径问题

$$\log \pi_1(z_1) + \log \phi_1(z_1) + \sum_{t=2}^{T} [\log \psi(z_{t-1}, z_t) + \log \phi_t(z_t)]$$



预测问题

- Viterbi算法
  - 用动态规划解概率最大路径，一个路径对应一个状态序列
  - 最优路径：如果最优路径在时刻t通过结点z_t^*，那么这一路径从结点z_t^*到终点z_T^*的部分路径，对于从z_t^*到z_T^*的所有可能的部分路径而言，必须是最优的
  - 为了找出最优路径的各个结点，从终结点开始，由后向前逐步求得结点z_T^*，$z_{T-1}^*, \cdots, z_1^*$，得到最优路径

预测问题

- 定义
  - $\delta_t(j)$：在时刻t隐藏状态为j的所有可能的状态转移路径$z_1, \ldots, z_t$中的概率最大值（为方便理解t时刻和t-1时刻的关系这里用联合概率，与后验概率是等价的，只差归一化常数）

$$\delta_t(j) = \max_{z_1,\ldots,z_{t-1}} \boldsymbol{P}(\boldsymbol{z_1}, \ldots, \boldsymbol{z_{t-1}}, \boldsymbol{z_t = j}, \boldsymbol{x_1}, \ldots, \boldsymbol{x_t})$$

$$\Rightarrow \quad \delta_t(j) = [\ \max_i\ \delta_{t-1}(i) \cdot \psi(i,j)\ ] \cdot \phi_t(j)$$

  - 在保证时刻t隐藏状态为j的前提下，t-1时刻的最优状态

$$a_t(j) = \arg\max_i\ \left[\delta_{t-1}(i) \cdot \psi(i,j)\right]$$

预测问题

- 初始化$\delta_1(j)$，并按照前面的公式依次计算$\delta_t(j)$以及$a_t(j)$

$$\delta_1(j) = \pi_j \phi_1(j)$$

- 计算T时刻最优隐含状态，然后通过$a_t(j)$逐步回溯前面t时刻最优状态

$$z_T^* = \arg\max_i \delta_T(i)$$

$$z_t^* = a_{t+1}(z_{t+1}^*)$$

Thanks!

Questions?