第5章 GMM-HMM

洪青阳

纲要

5.1 概率统计

5.2 高斯分布

5.3 GMM

5.4 GMM-HMM

5.5 GMM-HMM的训练

5.6 模型自适应

5.7 本章小结

5.1 概率统计

- 概率统计用来研究和揭示随机现象的的统计规律，应用范围很广，包括气象预报、水文预报、生物统计、保险、金融等领域。

观察序列(H：正面；T：反面)：

$O=\{o_1, o_2, \ldots, o_T\}$

H, H, T, …, T

硬币投掷试验

概率理论

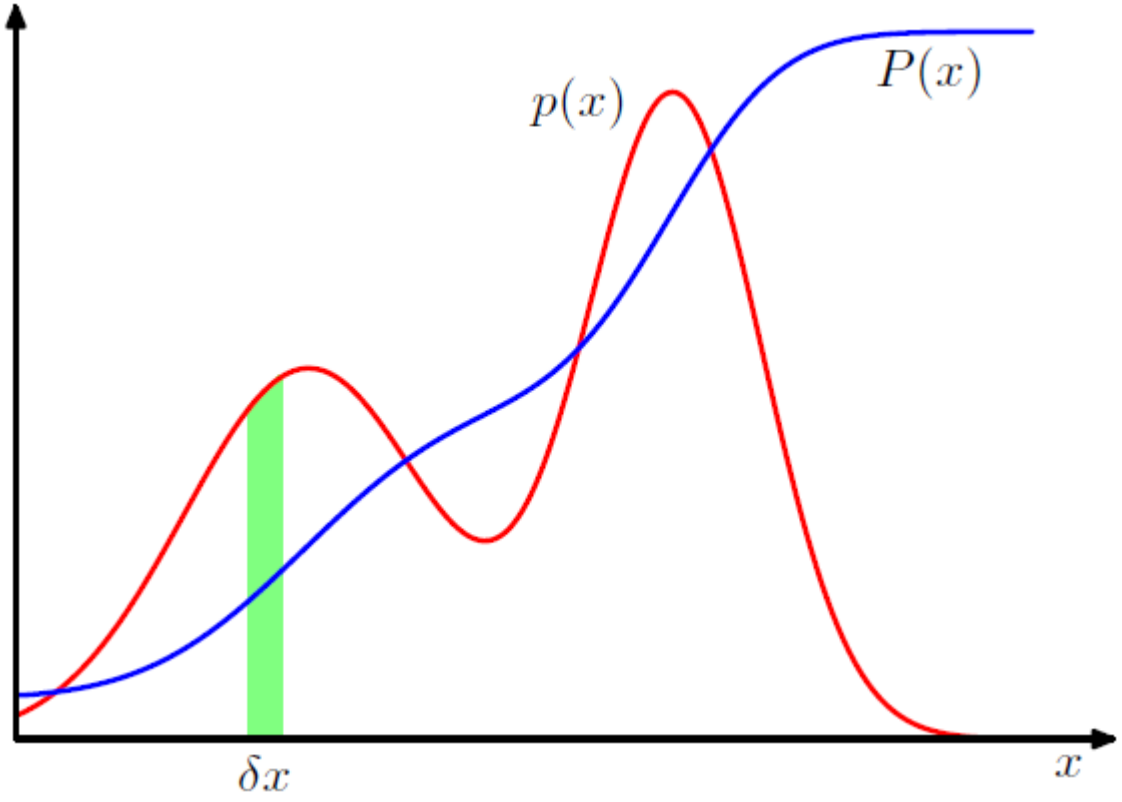
- 概率和：$p(X) = \sum_Y p(X, Y)$
- 联合概率：$p(X, Y) = p(Y|X)p(X)$
- 概率的对称性：$p(X, Y) = p(Y, X)$
- 贝叶斯(Bayes)定理：

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$$p(X) = \sum_Y p(X|Y)p(Y)$$

概率密度

$$p\big(x \in (a,b)\big) = \int_a^b p(x)d(x)$$



$$p(x) \geq 0$$

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

概率密度$p(x)$是积分函数$P(x)$的导数

连续变量的联合概率

$$p(x) = \int p(x, y) dy$$

$$p(x, y) = p(y|x)p(x)$$

期望(Expectation)

- 离散变量的期望

$$E[f] = \sum_x p(x)f(x)$$

- 连续变量的期望

$$E[f] = \int p(x)f(x)dx$$

- 条件期望

$$E_x[f|y] = \sum_x p(x|y)f(x)$$

方差(Covariance)

- 变量x的方差:

$$var[x] = E[x^2] - E[x]^2$$

- 变量x和y的方差:

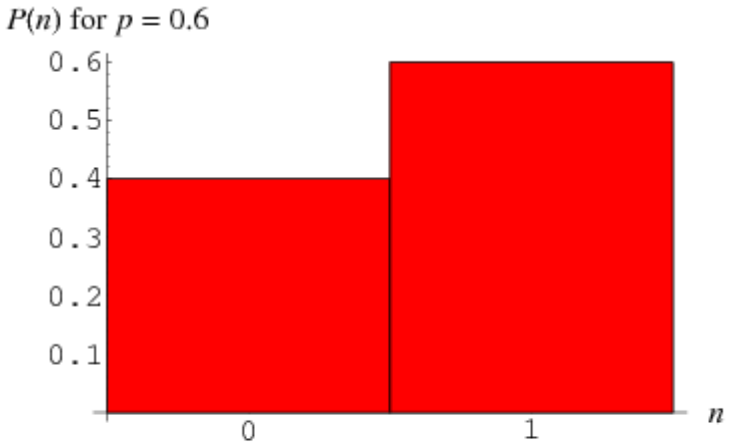
$$cov[x, y] = E_{x,y}[\{x - E[x]\}\{y - E[y]\}] = E_{x,y}[xy] - E[x]E[y]$$

- 两个矢量x和y的方差

$$cov[x, y] = E_{x,y}[\{x - E[x]\}\{y^T - E[y^T]\}] = E_{x,y}[xy^T] - E[x]E[y^T]$$

伯努利分布(Bernoulli distribution)

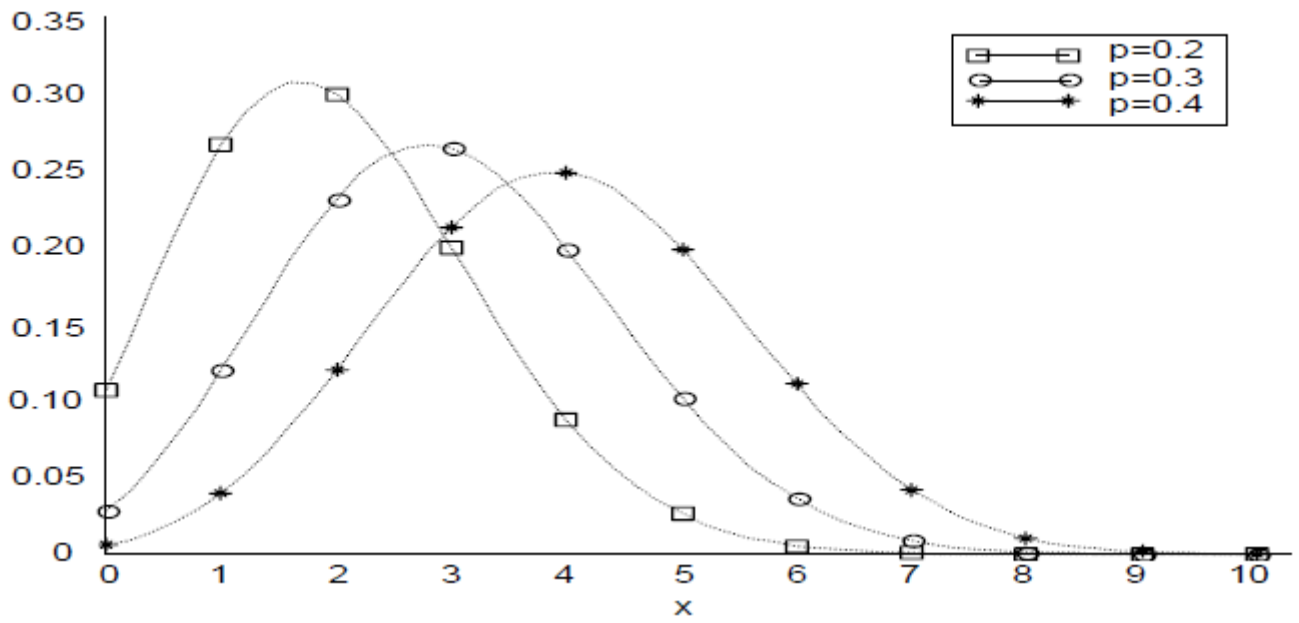
- 伯努利分布是一种离散分布，有两种可能的结果：
  - 1表示成功，出现的概率为p(其中$0 < p < 1$)。
  - 0表示失败，出现的概率为$q = 1 - p$。



二项式分布(Binomial Distribution)

- 二项式分布即**重复n次独立的伯努利试验**。当试验次数n为1时，二项式分布就是伯努利分布。

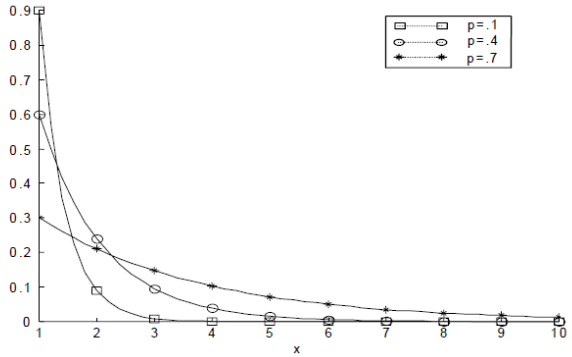
$$P(X=x)=f(x|n,p)=\binom{n}{x}p^x(1-p)^{n-x}$$



可用于描述**硬币投掷试验(正面/反面)**

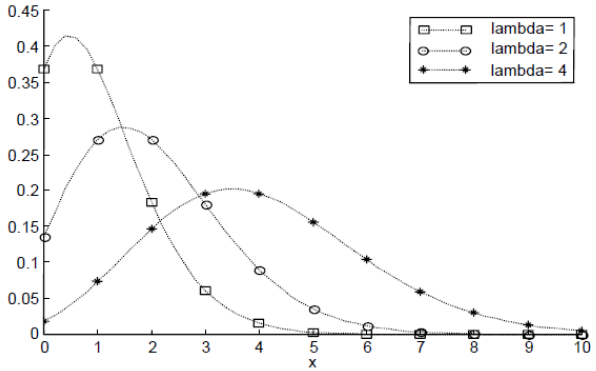
更多数学分布

$$P(X=x)=f(x\mid p)=p^{x-1}(1-p)\quad x=1,2,\ldots\ \text{and}\ 0<p<1$$



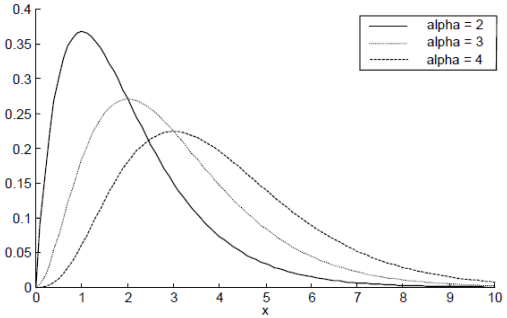
几何分布

$$P(X=x)=f(x\mid\lambda)=\begin{cases}\dfrac{e^{-\lambda}\lambda^{x}}{x!} & \text{for } x=0,1,2,\ldots\\ 0 & \text{otherwise}\end{cases}$$



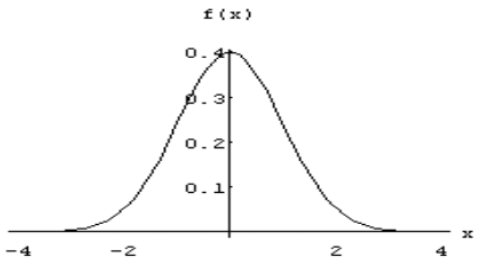
泊松分布

$$f(x\mid\alpha,\beta)=\begin{cases}\dfrac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} & x>0\\ 0 & x\le 0\end{cases}\qquad \Gamma(\alpha)=\int_0^{\infty}x^{\alpha-1}e^{-x}dx$$



伽马分布

$$f(x|\mu,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{1/2}}exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



高斯分布

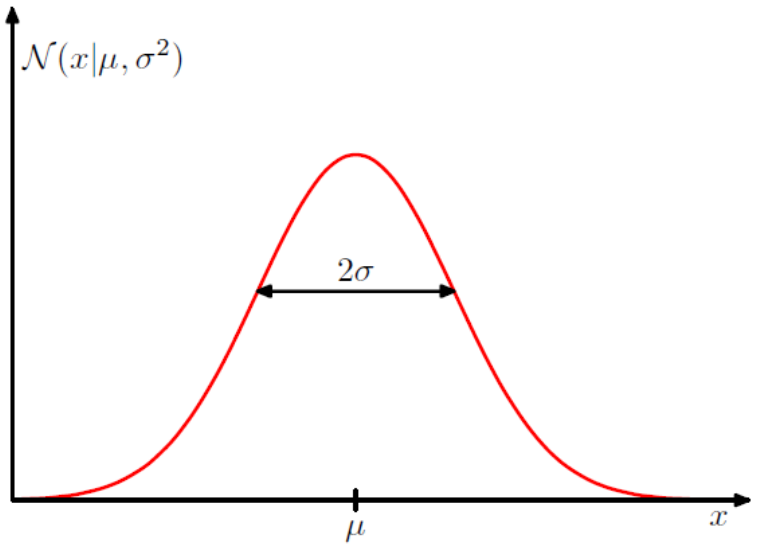
5.2 高斯分布(Gaussian Distribution)

- 又称正态分布(Normal Distribution)
- 高斯函数

$$N(x,\mu,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{1/2}}exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

其中μ是均值(mean)，σ^2是方差(covariance).

$$\int_{-\infty}^{\infty}N(x|\mu,\sigma^2)dx=1$$



高斯分布的期望和方差

- 期望

$$E[x] = \int_{-\infty}^{\infty} N(x|\mu, \sigma^2) x dx = \mu$$

$$E[x^2] = \int_{-\infty}^{\infty} N(x|\mu, \sigma^2) x^2 dx = \mu^2 + \sigma^2$$

- 方差

$$var[x] = E[x^2] - E[x]^2 = \sigma^2$$

二维高斯分布

$$p(x_1, x_2) = p(x_1)p(x_2) = N\left(x_1 \middle| \mu_1, \sigma_1^2\right) N\left(x_2 \middle| \mu_2, \sigma_2^2\right)$$

$$= \frac{1}{\left(2\pi\sigma_1^2\right)^{1/2}} exp\left\{-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2\right\}$$

$$\times \frac{1}{\left(2\pi\sigma_2^2\right)^{1/2}} exp\left\{-\frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2\right\}$$

$$= \frac{1}{2\pi\left(\sigma_1^2\sigma_2^2\right)^{1/2}} exp\left(-\frac{1}{2}\left(\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2}\right)\right)$$

D维高斯分布

D维变量$\boldsymbol{x} = \{x_1, x_2, \cdots, x_D\}$的高斯分布

$$N\left(\boldsymbol{x}|\mu, \sigma^2\right) = \frac{1}{(2\pi)^{D/2}} \frac{1}{\left(\prod_{d=1}^{D} \sigma_d^2\right)^{1/2}} exp\left\{-\frac{1}{2}\sum_{d=1}^{D}\frac{(x_d - \mu_d)^2}{\sigma_d^2}\right\}$$

$$= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\sum|^{1/2}} exp\left\{-\frac{1}{2}(\boldsymbol{x} - \mu)\sum^{-1}(\boldsymbol{x} - \mu)^T\right\}$$

其中D维μ是均值，$\sum$是$D \times D$协方差矩阵，$|\sum|$ 是$\sum$ 的行列式。

对角矩阵：$|\sum| = \prod_{d=1}^{D} \sigma_d^2$

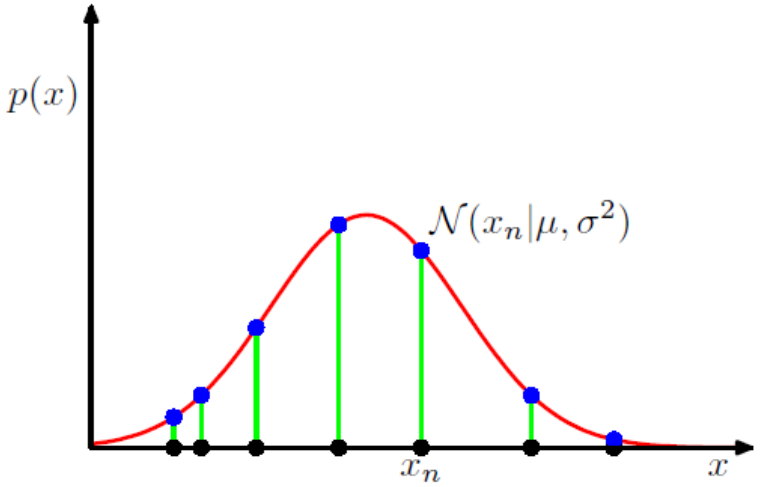
观察系列的概率计算

- 给定观察值序列

$$\boldsymbol{x} = (x_1, \cdots, x_N)$$

- 计算联合概率

$$p(\boldsymbol{x}|\mu, \sigma^2) = \prod_{n=1}^{N} N(x_n|\mu, \sigma^2)$$



对数概率

- 为简化计算，便于计算机处理(防止精度溢出)，一般采用对数概率(似然率)

$$\ln p\left(\boldsymbol{x}\middle|\mu,\sigma^2\right)=\sum_{n=1}^{N}\ln\left(N\left(x_n\middle|\mu,\sigma^2\right)\right)$$

$$=\sum_{n=1}^{N}\ln\left(\frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}exp\left\{-\frac{1}{2}\sum_{d=1}^{D}\frac{(x_{nd}-\mu_d)^2}{\sigma_d^2}\right\}\right)$$

$$=\sum_{n=1}^{N}\frac{1}{2}\left(-D\ln(2\pi)-\ln|\Sigma|-\sum_{d=1}^{D}\frac{(x_{nd}-\mu_d)^2}{\sigma_d^2}\right)$$

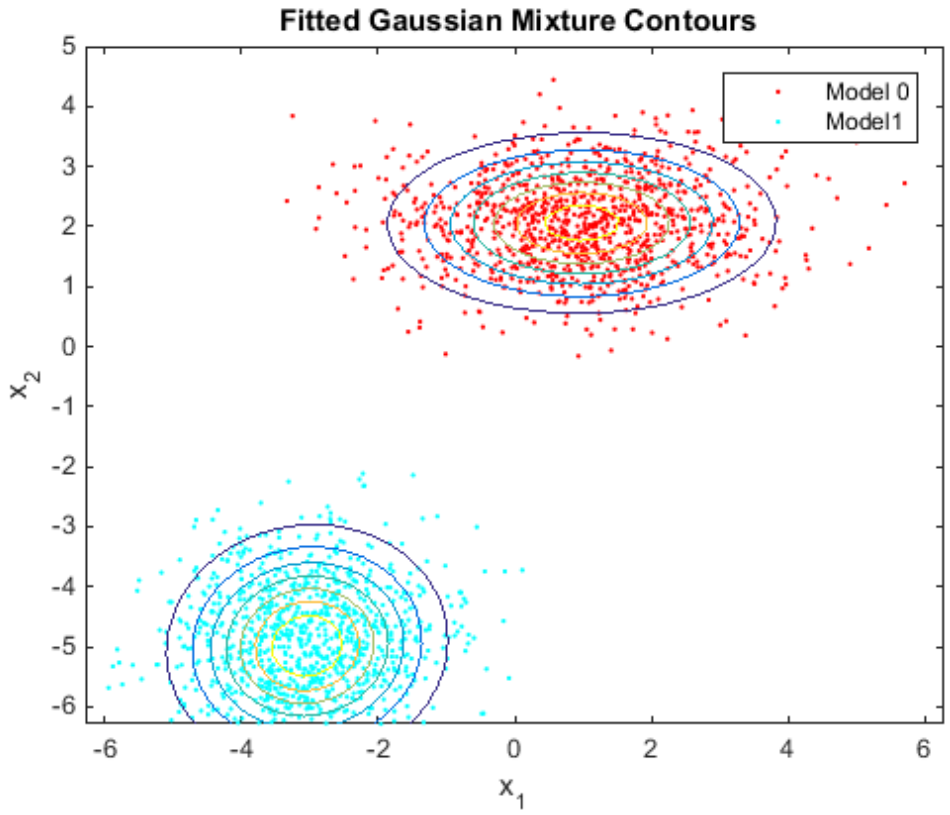
最大似然(Maximum Likelihood, ML)估计

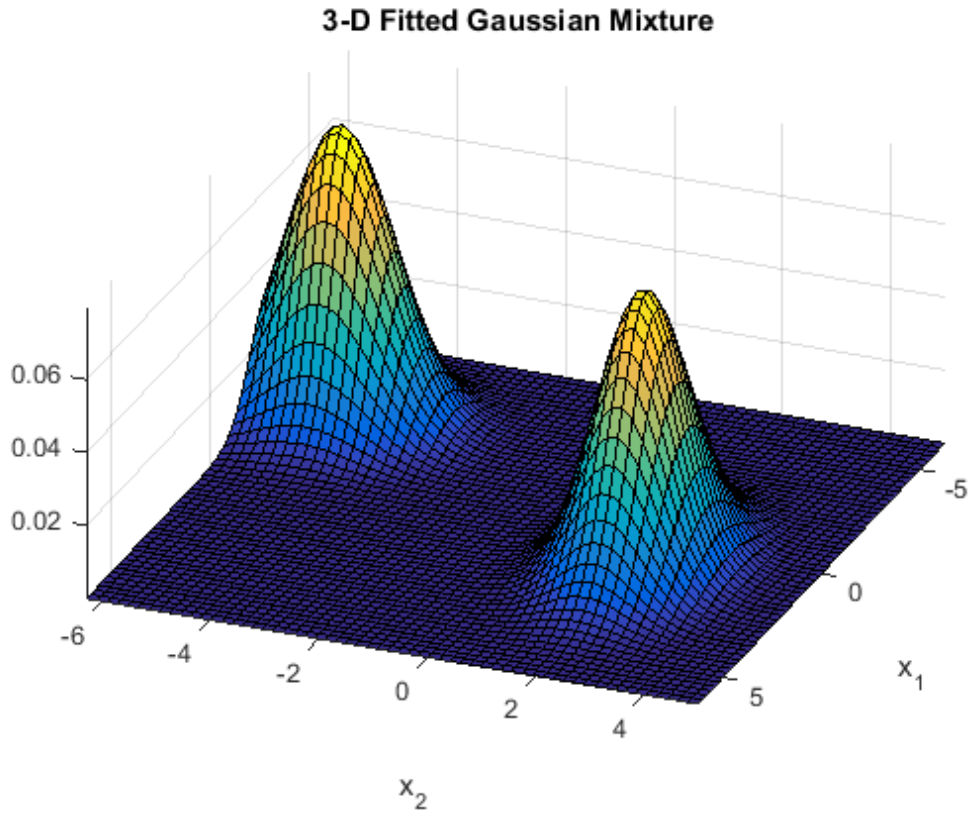
$$\begin{cases} \dfrac{\partial \ln p(\mathbf{x}|\mu,\sigma^2)}{\partial \mu} = \dfrac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu) = 0 \\ \dfrac{\partial \ln p(\mathbf{x}|\mu,\sigma^2)}{\partial \sigma^2} = -\dfrac{N}{2\sigma^2} + \dfrac{1}{2\sigma^4}\sum_{n=1}^{N}(x_n-\mu)^2 = 0 \end{cases}$$

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\sigma_{ML}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2$$

5.3 高斯混合模型（GMM）





高斯混合模型（GMM）

令$\lambda=\{\mu,\Sigma\}$, K阶高斯GMM的概率密度函数如下：

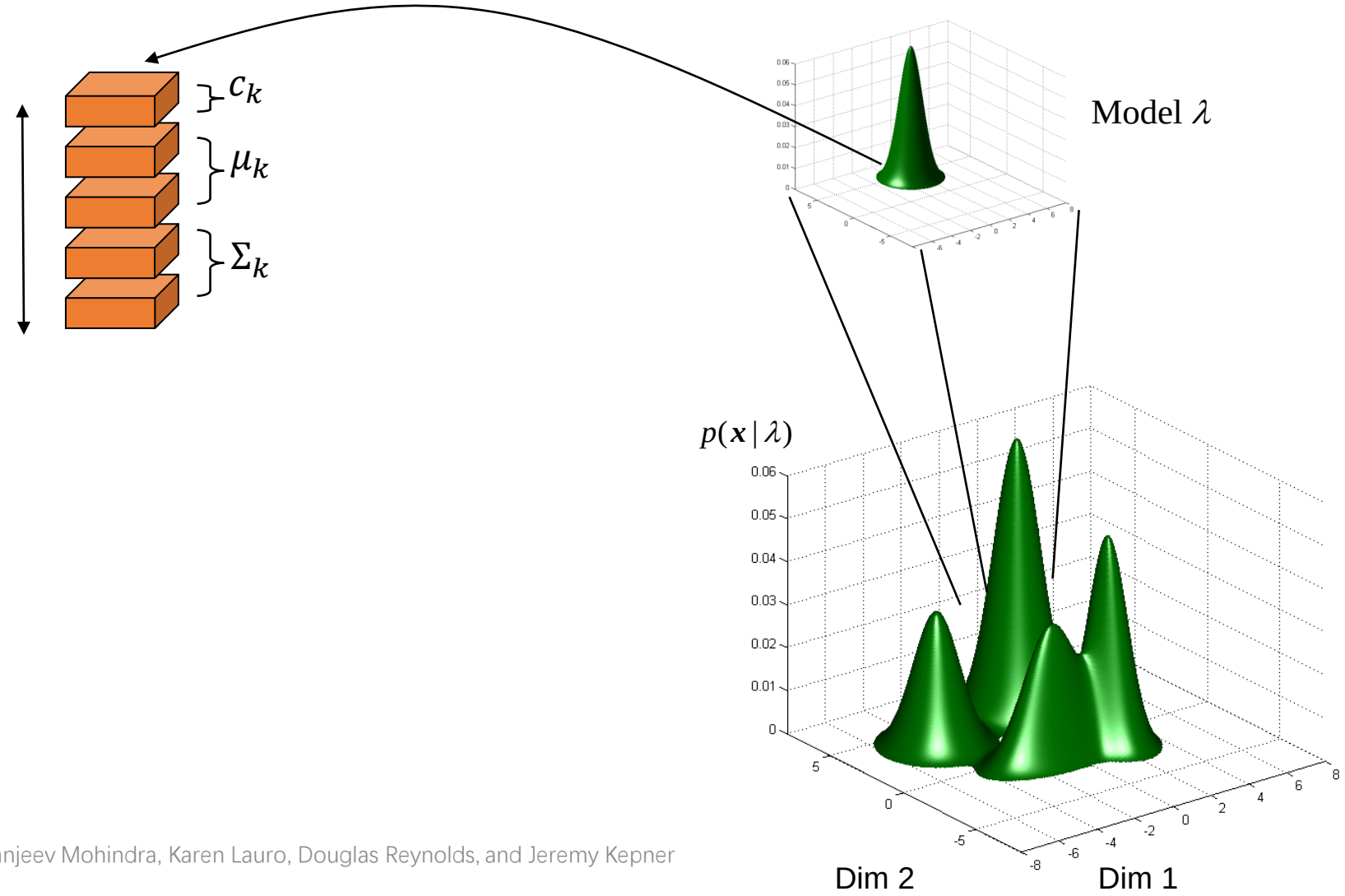
$$P(x|\lambda)=\sum_{k=1}^{K}P(x,k|\lambda)=\sum_{k=1}^{K}P(k)P(x|k,\lambda)=\sum_{k=1}^{K}c_k N(x|\mu_k,\Sigma_k)$$

其中$\sum_{k=1}^{K}c_k=1$

单高斯函数：

$$N(x|\mu_k,\Sigma_k)=\frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}}\exp\left\{-\frac{(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}{2}\right\}$$

高斯混合模型（GMM）



Nicolas Malyska, Sanjeev Mohindra, Karen Lauro, Douglas Reynolds, and Jeremy Kepner

GMM的初始化—*K*-means算法

ALGORITHM 5.1: The *K*-means Algorithm

Step 1: 初始化

把训练数据（特征向量）平均分配K组，计算每组的高斯均值μ_k。

Step 2: 最近邻分类

针对每个特征向量x_n，通过计算欧式距离，寻找与之最靠近的第k个高斯，并把该特征向量分配给这个高斯。

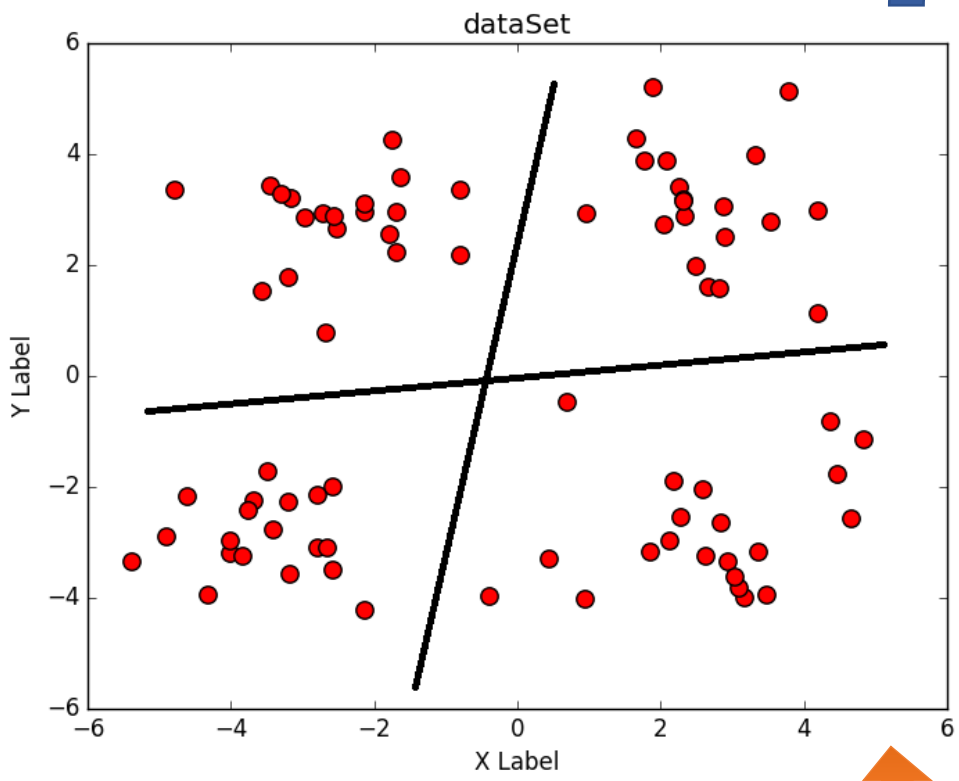
Step 3: 更新中心点

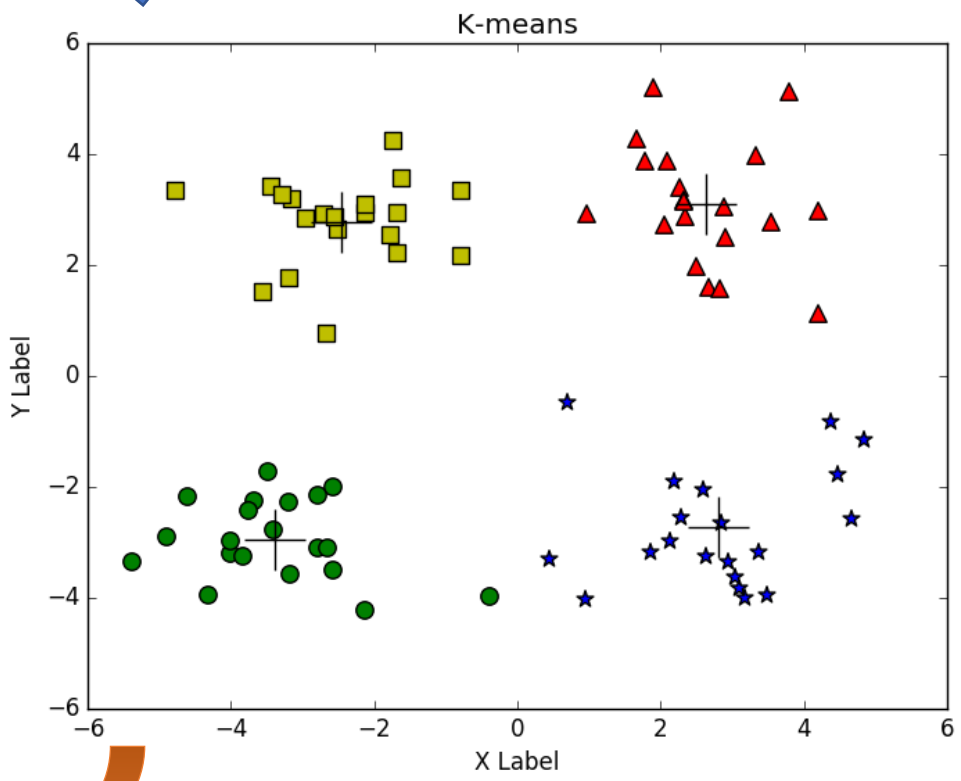
通过求平均，更新每个分布的中心点，得到对应高斯的均值。

Step 4: 迭代

重复步骤2和3，直到整体的平均距离低于预设的阈值。

GMM的初始化—*K*-means算法





GMM的似然函数

- 根据权重系数c_k限制，加入拉格朗日算子：

$$\ln p(\mathbf{x}|c,\mu,\Sigma)+\lambda\left(\sum_{k=1}^{K}c_k-1\right)$$

$$=\sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}c_k N(x_n|\mu_k,\Sigma_k)\right\}+\lambda\left(\sum_{k=1}^{K}c_k-1\right)$$

- 分别对μ_k,Σ_k,c_k求最大似然函数。

对μ_k求最大似然函数

- 对μ_k求偏导并令导数为0

$$-\sum_{n=1}^{N}\frac{c_k N(x_n|\mu_k,\Sigma_k)}{\sum_{k=1}^{K} c_k N(x_n|\mu_k,\Sigma_k)}\Sigma_k(x_n-\mu_k)=0$$

- 两边同除Σ_k，重新整理得到

$$\mu_k=\frac{\sum_{n=1}^{N}\gamma(n,k)x_n}{\sum_{n=1}^{N}\gamma(n,k)}$$

其中

$$\gamma(n,k)=\frac{c_k N(x_n|\mu_k,\Sigma_k)}{\sum_{k=1}^{K} c_k N(x_n|\mu_k,\Sigma_k)}$$

对$\sum_k$求最大似然函数

- 同样对$\sum_k$求偏导并令导数为0，得到

$$\sum_k = \frac{\sum_{n=1}^{N} \gamma\left(n,k\right)(x_n-\mu_k)(x_n-\mu_k)^T}{\sum_{n=1}^{N} \gamma\left(n,k\right)}$$

其中

$$\gamma(n,k) = \frac{c_k N(x_n|\mu_k,\sum_k)}{\sum_{k=1}^{K} c_k N(x_n|\mu_k,\sum_k)}$$

对c_k求最大似然函数

- 对c_k求偏导并令导数为0

$$\sum_{n=1}^{N} \frac{N(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} c_k N(x_n|\mu_k, \Sigma_k)} + \lambda = 0$$

得到

$$c_k = \frac{\sum_{n=1}^{N} \gamma(n,k)}{\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(n,k)}$$

EM算法估计GMM参数

ALGORITHM 5.2: The EM Algorithm

1.定义高斯数K，采用K-means算法，对每个高斯设置c_k,μ_k,$\sum_k$ 的初始值;

2.**E-step**

根据当前的c_k,μ_k,$\sum_k$计算后验概率

$$\gamma(n,k)=\frac{c_k N(x_n|\mu_k,\sum_k)}{\sum_{k=1}^{K} c_k N(x_n|\mu_k,\sum_k)}$$

3.**M-step**

根据E-step中计算的$\gamma(n,k)$，更新c_k,μ_k,$\sum_k$

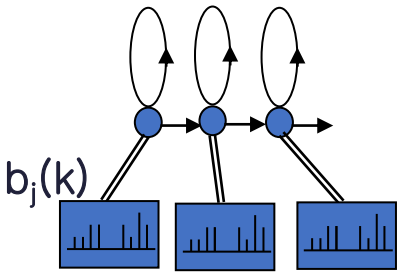
4.计算对数似然函数

$$\ln p(\mathbf{x}|c,\mu,\sum)=\sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K} c_k N(x_n|\mu_k,\sum_k)\right\}$$

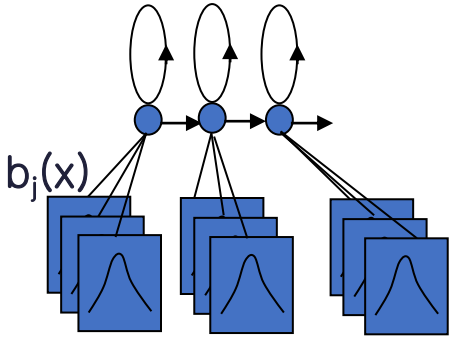
5.检查对数似然函数是否收敛，若不收敛，则返回第2步。

HMM分类

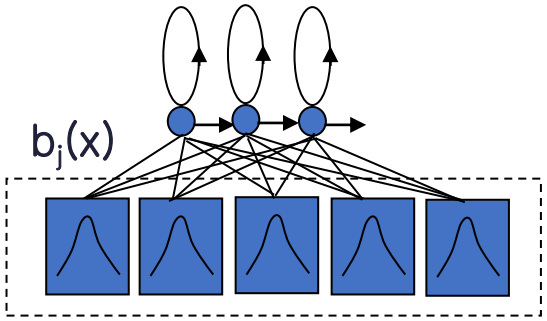
- 离散HMM:
  - ✓离散的符号作为观察值。



- 连续HMM:
  - ✓ 观察值为连续概率密度函数
  - ✓ 每个状态有不同的一组概率密度函数。



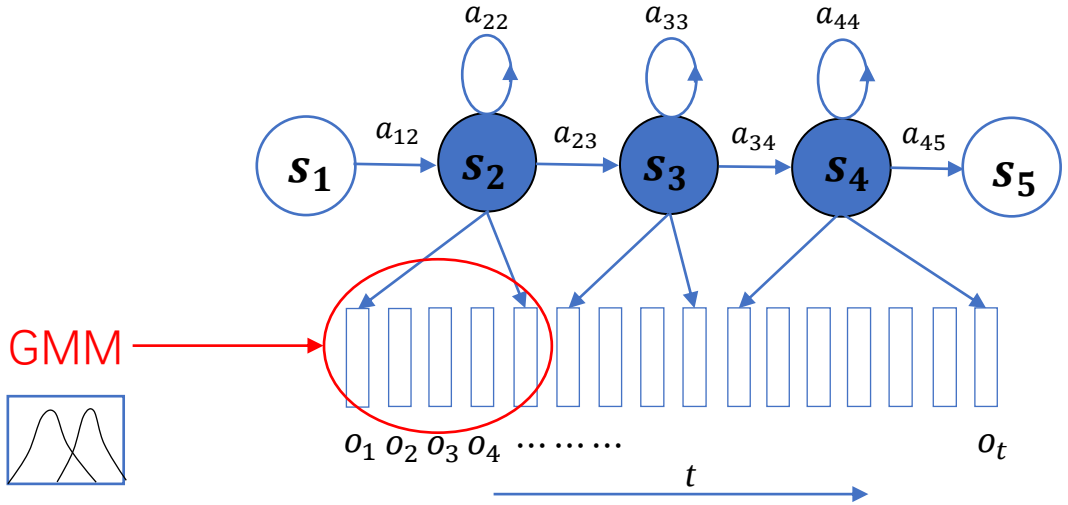
- 半连续HMM
  - ✓ 观察值为连续概率密度函数
  - ✓ 所有状态共享一组概率密度函数。



5.4 GMM-HMM

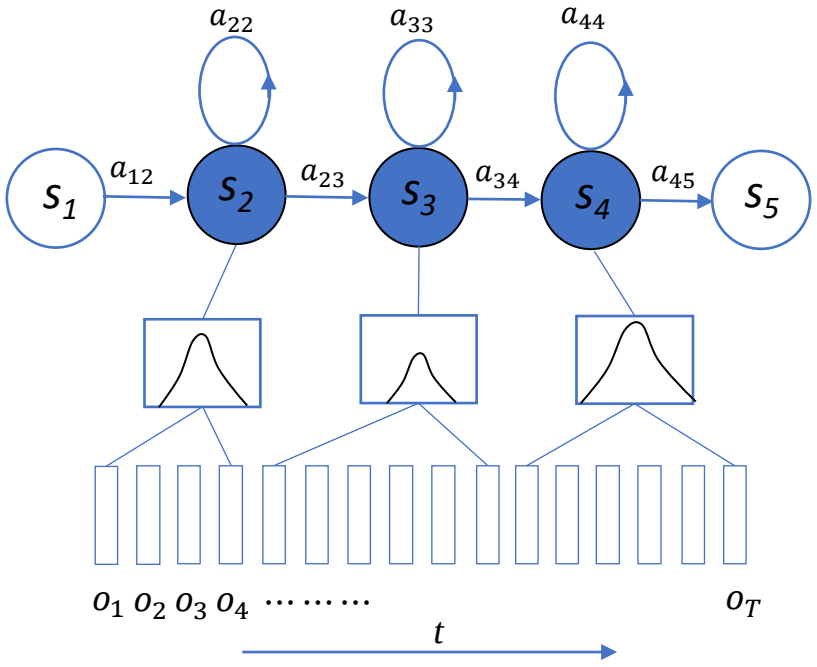
- 高斯混合概率密度函数(pdf):
  - 每个状态j表示为若干函数$N(o_t)$的线性组合
  - $N(o_t)$是连续高斯概率密度函数

$$b_j(o_t) = \sum_{k=1}^{K} c_{jk} N\left(o_t | \mu_{jk}, \Sigma_{jk}\right)$$



GMM-HMM

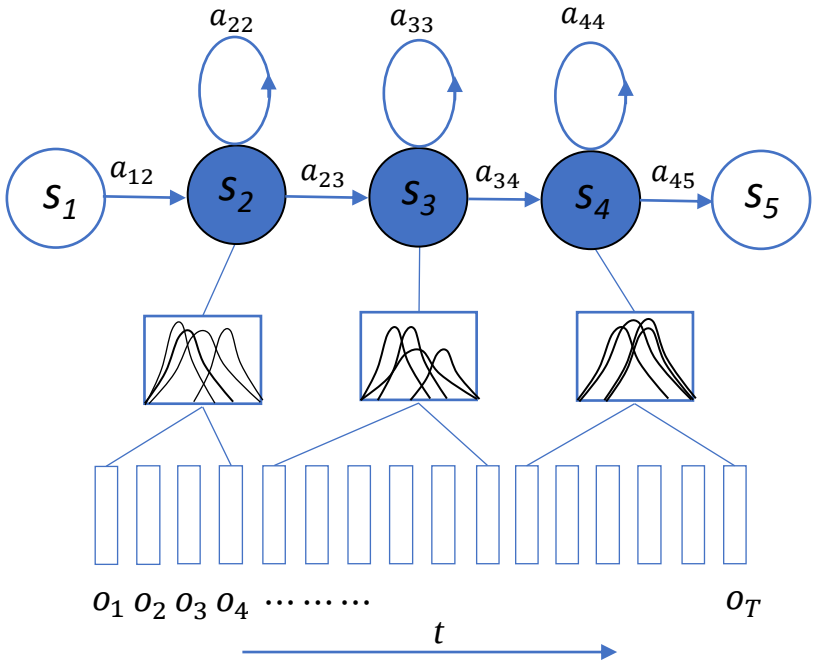
每个发射状态1个高斯




```
<BEGINHMM>
<NUMSTATES> 5
<STATE> 2
   <MEAN> 6
      0.1 0.2 0.3 0.8 0.4 0.4
   <VARIANCE> 6
      1.0 2.0 0.5 1.0 3.0 0.8
<STATE> 3
   <MEAN> 6
      0.3 0.2 0.4 0.7 0.5 0.6
   <VARIANCE> 6
      2.0 3.0 4.0 1.5 1.0 2.5
<STATE> 4
   <MEAN> 6
      0.2 0.2 0.4 0.9 0.7 0.8
   <VARIANCE> 6
      0.5 3.0 2.0 3.0 2.0 0.5
<TRANSP> 5
   0.0 1.0 0.0 0.0 0.0
   0.0 0.6 0.4 0.0 0.0
   0.0 0.0 0.5 0.5 0.0
   0.0 0.0 0.6 0.4 0.0
   0.0 0.0 0.0 0.0 0.0
<ENDHMM>
```

GMM-HMM

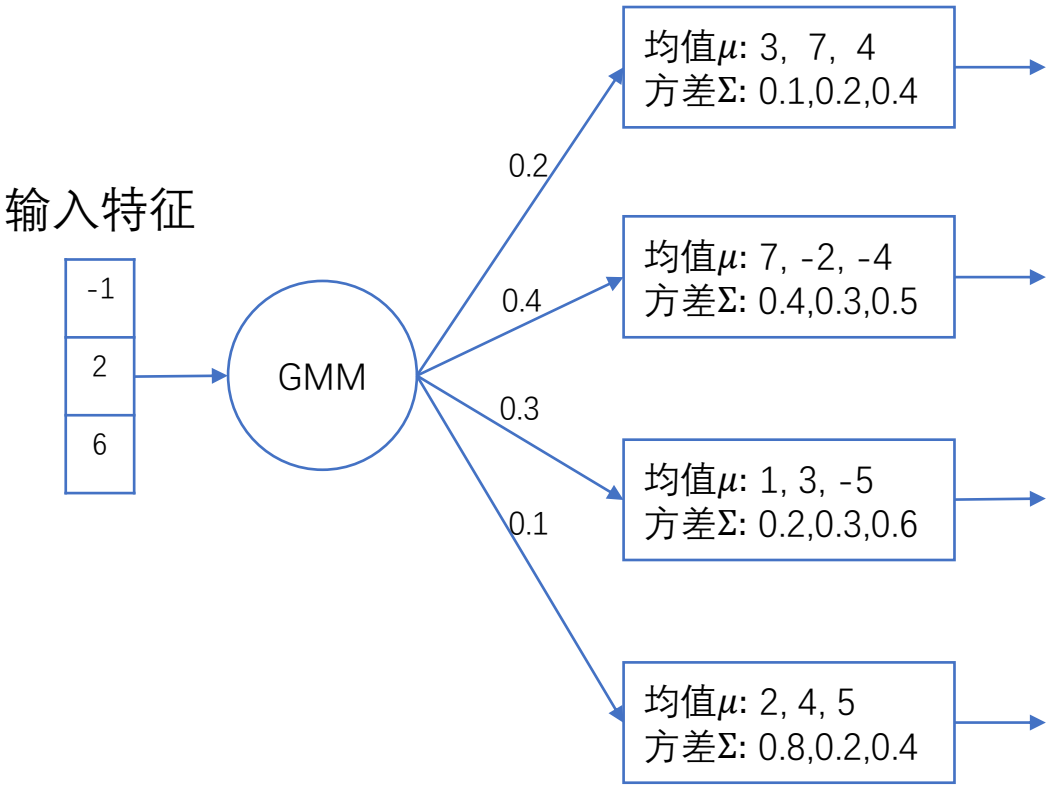
每个发射状态4个高斯




```
<STATE> 2
<NUMMIXES> 4
<MIXTURE> 1 0.2
   <MEAN> 6
      0.1 0.2 0.3 0.8 0.4 0.4
   <VARIANCE> 6
      1.0 2.0 0.5 1.0 3.0 0.8
<MIXTURE> 2 0.3
   <MEAN> 6
      0.1 0.2 0.3 0.8 0.4 0.4
   <VARIANCE> 6
      1.0 2.0 0.5 1.0 3.0 0.8
<MIXTURE> 3 0.3
   <MEAN> 6
      0.1 0.2 0.3 0.8 0.4 0.4
   <VARIANCE> 6
      1.0 2.0 0.5 1.0 3.0 0.8
<MIXTURE> 4 0.2
   <MEAN> 6
      0.1 0.2 0.3 0.8 0.4 0.4
   <VARIANCE> 6
      1.0 2.0 0.5 1.0 3.0 0.8
```

GMM概率计算

$$N(\boldsymbol{x}|\mu,\sigma^2)=\frac{1}{(2\pi)^{D/2}}\frac{1}{(\prod_{d=1}^{D}\sigma_d^2)^{1/2}}exp\left\{-\frac{1}{2}\sum_{d=1}^{D}\frac{(x_d-\mu_d)^2}{\sigma_d^2}\right\}$$



$$\frac{1}{(2\pi)^{3/2}}\frac{1}{(0.1\times0.2\times0.4)^{1/2}}exp\left\{-\frac{1}{2}\left(\frac{(-1-3)^2}{0.1}+\frac{(2-7)^2}{0.2}+\frac{(6-4)^2}{0.4}\right)\right\}$$

$$\frac{1}{(2\pi)^{3/2}}\frac{1}{(0.4\times0.3\times0.5)^{1/2}}exp\left\{-\frac{1}{2}\left(\frac{(-1-7)^2}{0.4}+\frac{(2+2)^2}{0.3}+\frac{(6+4)^2}{0.5}\right)\right\}$$

$$\frac{1}{(2\pi)^{3/2}}\frac{1}{(0.2\times0.3\times0.6)^{1/2}}exp\left\{-\frac{1}{2}\left(\frac{(-1-1)^2}{0.2}+\frac{(2-3)^2}{0.3}+\frac{(6+5)^2}{0.6}\right)\right\}$$

$$\frac{1}{(2\pi)^{3/2}}\frac{1}{(0.8\times0.2\times0.4)^{1/2}}exp\left\{-\frac{1}{2}\left(\frac{(-1-2)^2}{0.8}+\frac{(2-4)^2}{0.2}+\frac{(6-5)^2}{0.4}\right)\right\}$$

5.5 GMM-HMM的训练

- 需要重估的参数
  - 起始概率
  - 转移概率
  - 各状态中不同pdf的权重
  - 各状态中不同pdf的均值和方差

GMM-HMM参数重估

若观察概率为混合高斯分布形式，即

$$b_j(o_t) = \sum_{k=1}^{K} c_{jk} N(o_t, \mu_{jk}, \Sigma_{jk})$$

定义针对状态j的第k个高斯的统计量如下：

$$\gamma_t^c(j,k) = \left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}\right]\left[\frac{c_{jk} N(\mathbf{o}_t^c, \mu_{jk}, \Sigma_{jk})}{\sum_{k=1}^{K} c_{jk} N(\mathbf{o}_t^c, \mu_{jk}, \Sigma_{jk})}\right]$$

$$= \begin{cases} \dfrac{1}{P(\mathbf{O}|\lambda)} \pi_j \beta_1(j) c_{jk} N(o_1^c, \mu_{jk}, U_{jk}), & t = 1 \\ \dfrac{1}{P(\mathbf{O}|\lambda)} \displaystyle\sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} \beta_t(j) c_{jk} N(o_t^c, \mu_{jk}, U_{jk}), & t > 1. \end{cases}$$

GMM-HMM参数重估

$$a_{ij} = \frac{\sum_{c=1}^{C}\sum_{t=1}^{T_c-1}\xi_t^c(i,j)}{\sum_{c=1}^{C}\sum_{t=1}^{T_c-1}\gamma_t^c(i)}$$

$$c_{jk} = \frac{\sum_{c=1}^{C}\sum_{t=1}^{T_c}\gamma_t^c(j,k)}{\sum_{k=1}^{K}\sum_{c=1}^{C}\sum_{t=1}^{T_c}\gamma_t^c(j,k)}$$

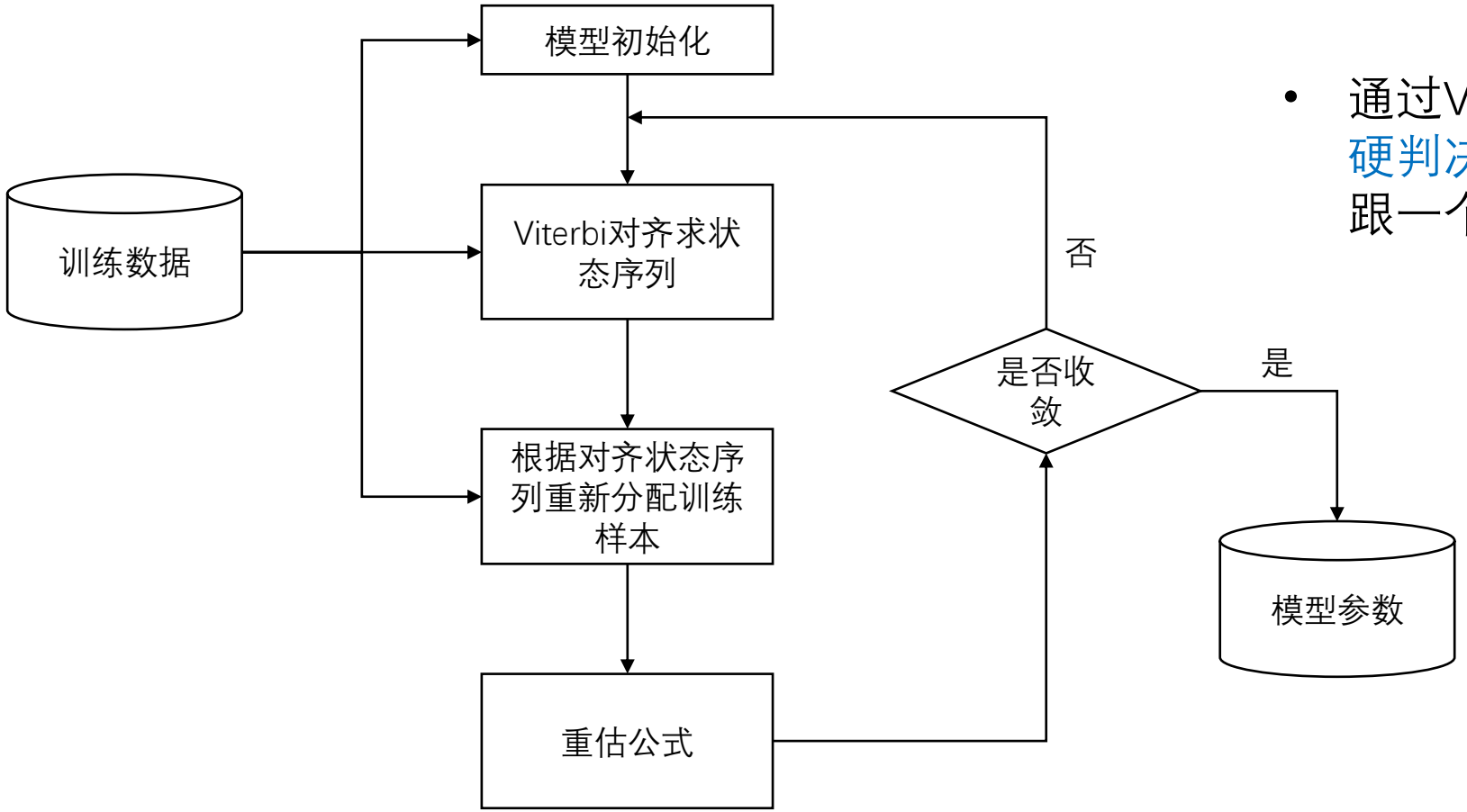
$$\mu_{jk} = \frac{\sum_{c=1}^{C}\sum_{t=1}^{T_c}\gamma_t^c(j,k)\mathbf{o}_t^c}{\sum_{c=1}^{C}\sum_{t=1}^{T_c}\gamma_t^c(j,k)}$$

$$\Sigma_{jk} = \frac{\sum_{c=1}^{C}\sum_{t=1}^{T_c}\gamma_t^c(j,k)(\mathbf{o}_t^c-\mu_{jk})(\mathbf{o}_t^c-\mu_{jk})'}{\sum_{c=1}^{C}\sum_{t=1}^{T_c}\gamma_t^c(j,k)}$$

- 四个参数的估计，均跟前向和后向概率关联，是一种软判决，即分配到状态的观察值多少是用概率来调节的。
- 这样的分配机制跟普通HMM的训练过程是类似的，即均为Baum-Welch算法。

其中C为训练条数，$\xi_t^c(i,j)$为第c条语音在t时刻从状态i到状态j的统计量。

Viterbi训练流程



- 通过Viterbi对齐，是一种硬判决，即每个观察值只跟一个状态对应。

5.6 模型自适应

- 自适应的必要性
  - 口音
  - 感冒...

- 经典的自适应算法：
  - 1994年提出的最大后验概率估计（Maximum A Posteriori Estimation，**MAP**）
  - 1995年提出的最大似然线性回归（Maximum Likelihood Linear Regression，**MLLR**）

5.6 模型自适应—MAP

- MAP算法
  - 本质上是重新训练一次，对原$\boldsymbol{B}$矩阵进行微调
  - 特点：简单，对每个HMM单独自适应，只需一次发音

$$\hat{\lambda} = \varepsilon\lambda + (1-\varepsilon)\lambda', 0 \le \varepsilon \le 1$$

  - 针对GMM的均值参数，基于原有均值μ_0和自适应数据$\boldsymbol{X} = \{x_1, x_2, \cdots, x_T\}$，得到新的自适应均值$\hat{\mu}$如下：

$$\hat{\boldsymbol{\mu}} = \frac{\tau\boldsymbol{\mu}_0 + \sum_{t=1}^{T}\gamma_t x_t}{\tau + \sum_{t=1}^{T}\gamma_t}$$

其中τ控制原有模型均值和自适应数据之间的平衡（一般取值$0 \le \tau \le 20$），γ_t是对应高斯在时刻t的统计量。

5.6 模型自适应—MLLR

- MLLR算法
  - 本质：将原模型的参数进行线性变换后再进行识别
  - 特点：少量语音可以对所有模型进行自适应，只要得到线性变换矩阵即可

$$\hat{\boldsymbol{\mu}} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}$$

  - 假如声学特征是d维向量，则$\boldsymbol{A}$为$d \times d$维矩阵，$\boldsymbol{b}$为d维向量。定义$\boldsymbol{W} = [\boldsymbol{b}\boldsymbol{A}]$，$\boldsymbol{\eta} = [1\boldsymbol{\mu}^T]^T$，则有

$$\hat{\boldsymbol{\mu}} = \boldsymbol{W}\boldsymbol{\eta}$$

其中转换矩阵$\boldsymbol{W}$采用自适应数据估计得到，并且可以由所有或部分高斯共用。

5.7 本章小结

- 重点讲解了GMM的定义，特别是与HMM的结合，用于特征序列的建模。
- 本章的难点在于输入的多维特征如何跟HMM的每个状态结合起来，并用该状态的GMM参数计算其观察概率。
- 通过课程实践，加强对GMM-HMM的理解。