深蓝学院
shenlanxueyuan.com

音频语音与语言处理研究组
SLP
Audio, Speech and Language Processing Group NPU

语音识别：从入门到精通

第三讲：GMM以及EM算法

主讲人 孙思宁

西北工业大学博士

ssning2013@gmail.com

内容提要

1.潜变量模型

2.K-Means 聚类

- K-means 回顾
- K-means 应用

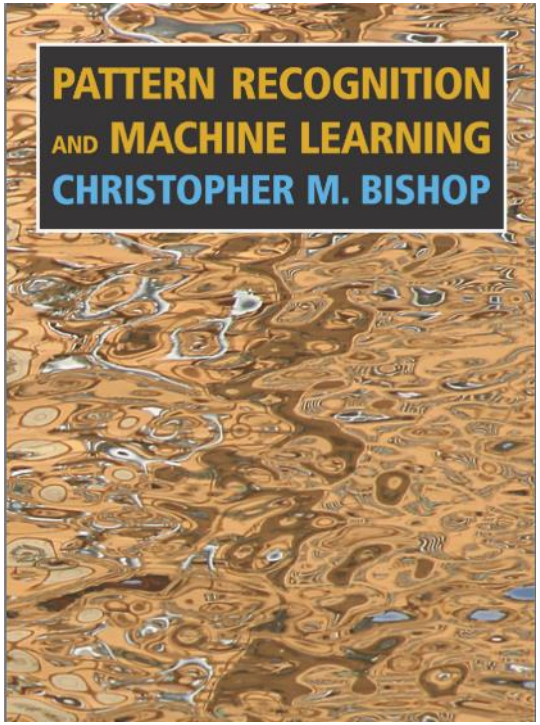
3.GMM 模型

- GMM模型基础
- GMM参数的EM估计

4.EM算法

- 深入EM算法
- EM算法的通用解释

5.实践



如果你只想完成作业...

- **你只需要知道什么是GMM以及GMM的参数估计过程**

给定一个GMM模型，优化目标是寻找使似然函数最大的各个高斯成分的均值向量、协方差矩阵和混合系数

1. 初始化　　初始化参数μ_k, Σ_k, π_k

2. E步　　使用当前参数计算后验概率

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)}$$

3. M步　　使用后验重新估计参数

$$\mu_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n$$

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k^{new})(x_n - \mu_k^{new})^T$$

$$\pi_k^{new} = \frac{N_k}{N}, \qquad N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

4. 重算似然　　重新计算似然函数，重复2-4，直至满足收敛条件

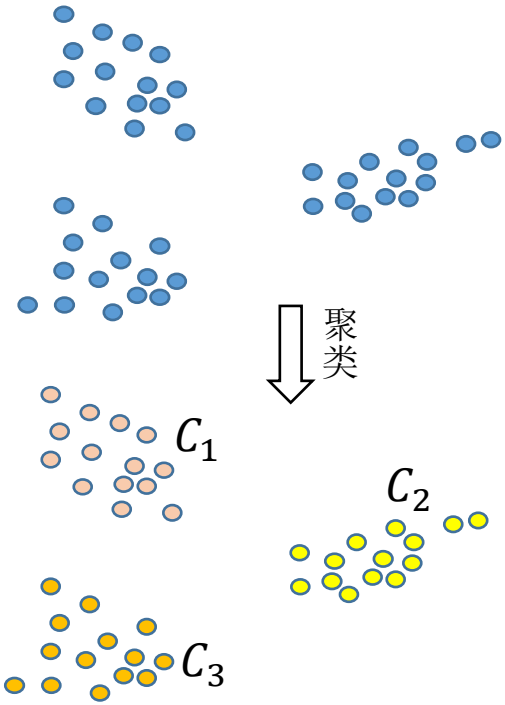
$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

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

潜变量模型

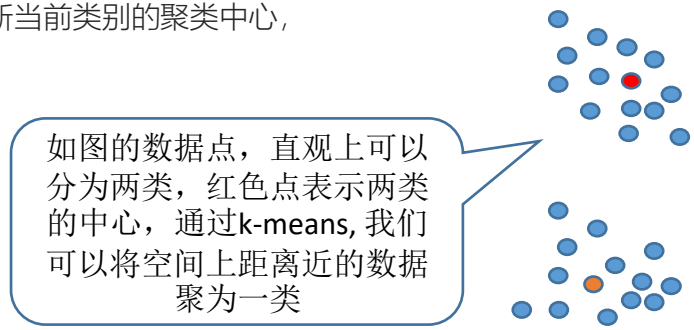
- **观测变量（observed variable）**
  - 直接可以观测到的变量
- **潜（隐）变量（latent variable ）**
  - 无法直接被观测到，需要通过模型和观测变量进行推断
  - 利用潜变量来解释观测变量的数学模型，称为潜变量模型，GMM、HMM都是潜变量模型
  - 潜变量模型将**不完全数据**（只有观测数据）的边缘分布转换成容易处理的**完全数据**（观测数据+潜变量）的联合分布

右图给出一个示例，示例中，潜变量是类别，是未知的，观测变量是数据点，只通过观测变量，如何推断出哪些观测属于同一个类别?



K-means聚类

- **问题定义：**
  - 给定一个含有N个数据点的集合$\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$，$\mathbf{x}_\mathrm{n} \in R^D$,聚类的目标是将此$N$个数据点聚类到$K$个类别中，且假设$K$值已经给定。
- **K-means思路**
  1. 引入K个D维均值向量$\boldsymbol{\mu}_k, k = 1,2, \dots, K$， $\boldsymbol{\mu}_k$即为第k个类别的聚类中心
  2. 计算数据点$\mathbf{x}_\mathrm{n}$和所有类中心$\boldsymbol{\mu}_k$的距离（如欧式距离），类中心距离此数据点最近的类别，即为当前数据点的类别
  3. 根据新的聚类结果，使用当前聚集到各个类别的数据的均值来更新当前类别的聚类中心，
  4. 返回第2步，直到满足一定的停止准则



K-means聚类

- **引入潜变量**
  - 对于每一个数据点$\mathbf{x_n}$，引入一个二进制指示因子 $r_{nk} \in \{0,1\}$, 如果$\mathbf{x_n}$属于第k个类别，则 r_{nk}=1，否则， r_{nk} =0， r_{nk}即为潜变量
  - 定义目标函数 $J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$
  - 优化目标：寻找合适的 $\{r_{nk}\}$ 和 $\{\boldsymbol{\mu}_k\}$，使得目标函数J最小
- **模型优化：两阶段迭代优化方法 （初识EM）**
  - 选择初始的$\boldsymbol{\mu}_k$值，并保持$\boldsymbol{\mu}_k$值固定，最小化J关于r_{nk}（E步）

为什么这一步是期望？
是对哪个变量的期望？

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

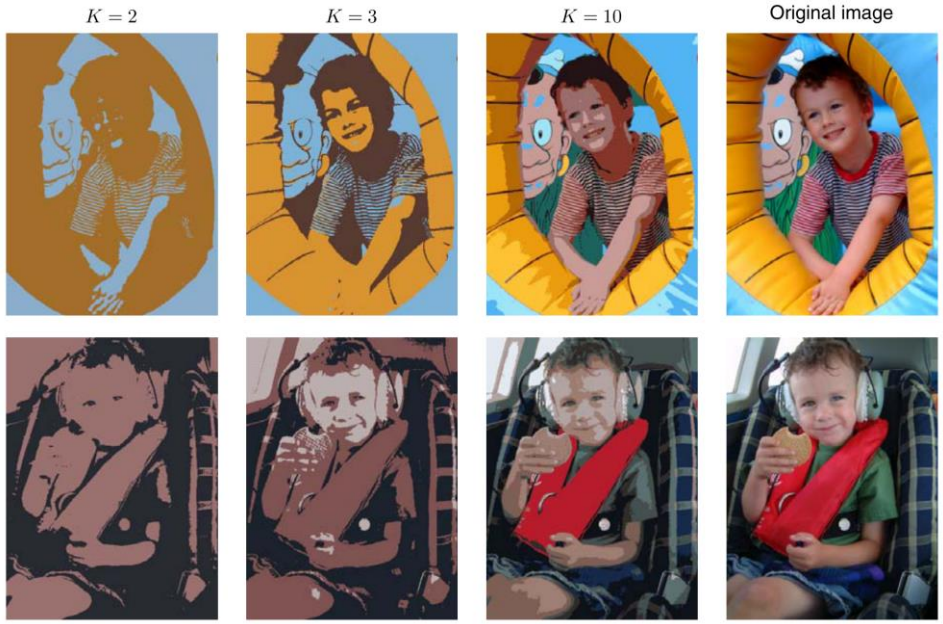
  - 保持r_{nk}固定，最小化J关于$\boldsymbol{\mu}_k$ （M步）

$$\frac{\partial J}{\partial \boldsymbol{\mu}_k} = 2\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \rightarrow \boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\,\mathbf{x}_n}{\sum_n r_{nk}}$$

$\sum_n r_{nk}$为当前第k类中的数据点数

K-means聚类

- **K-means聚类应用**
  - 图像分割和压缩



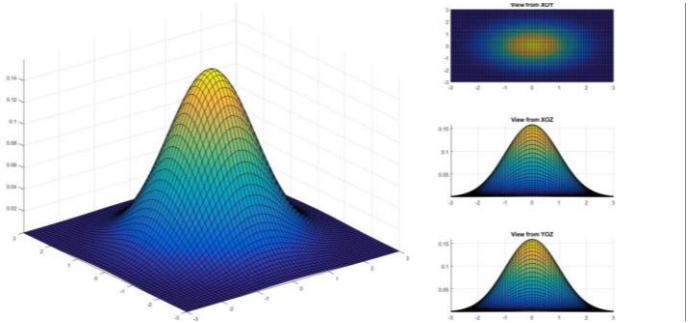
GMM模型

- **高斯分布**
  - D维随机变量的高斯分布为：

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

其中，$\boldsymbol{\mu} \in \boldsymbol{R^D}$,为高斯分布的均值向量，

$\boldsymbol{\Sigma} \in \boldsymbol{R^{D\times D}}$，为高斯分布的协方差矩阵



- **为什么选择高斯分布**
  - 高斯分布在自然界的数据中广泛存在
  - 中心极限定理：在适当的条件下，大量相互独立随机变量的均值经适当标准化后依分布收敛于正态分布

GMM模型

- **最大似然估计**
  - 假设随机变量 X服从分布$p(X|\boldsymbol{\theta})$，即$X\sim p(X|\boldsymbol{\theta})$，其中，$\boldsymbol{\theta}$为待估计的参数，如果可以获得$N$个互相独立的$X$的采样点$\{\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_N\}$，则似然函数的定义为

$$p(\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_N)=\prod_{n=1}^{N}p(\mathbf{x}_n|\boldsymbol{\theta})$$

在实际使用中，一般采用对数似然函数：

$$\ln p(\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_N|\boldsymbol{\theta})=\sum_{n=1}^{N}\ln p(\mathbf{x}_n|\boldsymbol{\theta})$$

参数$\boldsymbol{\theta}$的最大似然估计为：

$$\boldsymbol{\theta}=\arg\max_{\boldsymbol{\theta}}\ln p(\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_N|\boldsymbol{\theta})$$

- **高斯模型的最大似然估计**

$$\boldsymbol{\mu}_{ML}=\frac{1}{N}\sum_{n}\mathbf{x}_n\ ,\quad \boldsymbol{\Sigma}_{ML}=\frac{1}{N}\sum_{n}(\mathbf{x}_n-\boldsymbol{\mu}_{ML})(\mathbf{x}_n-\boldsymbol{\mu}_{ML})^T$$

GMM模型

- **高斯混合分布**

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

$$p(\mathbf{x}) = \sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \quad (1)$$

其中

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

$\{\pi_k\},\{\boldsymbol{\mu}_k\},\{\boldsymbol{\Sigma}_k\}$为待估计参数

- **π_k的解释（直观理解：第k个高斯所占的比重）**
  - 引入一个K维 one-hot(只有一维为1，其余维度为0) 向量$\boldsymbol{z} = [z_1,\dots,z_k,\dots,z_K], z_k \in \{0,1\}, \sum_k z_k = 1$，概率$P(z_k = 1)$为向量$\boldsymbol{z}$的第$k$维为1的**先验概率**，

$$p(z_k = 1) = \pi_k \quad (2)$$

  - 向量$\boldsymbol{z}$的分布可以表示为

$$p(\boldsymbol{z}) = \prod_{k=1}^{K}\pi_k^{z_k} \quad (3)$$

等价于$\mathrm{p(z)} = \pi_k, \text{where } z_k = 1$

GMM模型

- **条件分布** $p(\mathbf{x}|z_k=1)=\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ $p(\mathbf{x}|\mathbf{z})=\prod_{k=1}^{K}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)^{z_k}$ (4)
- **联合分布**

$$p(x,z)=\prod_{k=1}^{K}\pi_k^{z_k}N\left(x|u_k,\Sigma_k\right)^{z_k}$$

- **边缘分布**
  - 使用贝叶斯公式，对潜变量求和，得到

$$p(\mathbf{x})=\sum_{\mathbf{z}}p(\mathbf{x},\mathbf{z})=\sum_{\mathbf{z}}p(\mathbf{z})p(\mathbf{x}|\mathbf{z})=\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \quad (5)$$

  - 对于每一个观测$\mathbf{x}_n$，都有一个潜变量$\mathbf{z}_n$和其对应，上述公式将变量$\mathbf{x}$和潜变量$\mathbf{z}$联系起来，并且引入了联合分布$p(\mathbf{x},\mathbf{z})$，如前所述，完成了将观测数据的边缘分布转换成观测和潜变量的联合分布，之后，我们将重点处理联合分布$p(\mathbf{x},\mathbf{z})$。
- **后验分布**

$$\gamma(z_k)\equiv p(z_k=1|\mathbf{x})=\frac{p(z_k=1)p(\mathbf{x}|z_k=1)}{\sum_{j=1}^{K}p(z_j=1)p(\mathbf{x}|z_j=1)}=\frac{\pi_k\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}. \quad (6)$$

$\gamma(z_k)$为得到观测x后，$z_k=1$的**后验概率**，理解成第k个高斯成分对于生成观测x的贡献值

GMM模型

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

- **GMM的对数似然函数**

$$\ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\}. \qquad (7)$$

其中， $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1^T\\ \vdots \\ \mathbf{x}_N^T\end{bmatrix}$ 同时给出潜变量矩阵定义 $\mathbf{Z} = \begin{bmatrix}\mathbf{z}_1^T\\ \vdots \\ \mathbf{z}_N^T\end{bmatrix}$

- **GMM模型参数估计的EM算法（最大似然准则）**
  - 似然函数$\ln(p(X|\pi,\boldsymbol{\mu},\boldsymbol{\Sigma})$分别对参数$\pi,\boldsymbol{\mu},\boldsymbol{\Sigma}$求导

$$0 = \sum_{n=1}^{N}\underbrace{\frac{\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_j\pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k)$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \qquad N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}$$

拉格朗日乘子法，求解λ

$$0 = \sum_{n=1}^{N}\frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_j\pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)} + \lambda$$

等号两边同时乘以π_k，并对k求和

$$0 = \sum_{n=1}^{N}\sum_{k}\frac{\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_j\pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)} + \lambda\sum_k\pi_k$$

$$0 = \sum_{n=1}^{N}\frac{\sum_k\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_j\pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)} + \lambda$$

$$\lambda = -N$$

$$\ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)$$

$$0 = \sum_{n=1}^{N}\frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_j\pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)} + \lambda$$

$$\pi_k = \frac{N_k}{N}$$

GMM模型

- **GMM模型参数估计的EM算法总结**
  - 上述的参数估计方法并不是一个严格的解析解，因为公式中的后验概率$\gamma(z_{nk})$是依赖于每个高斯的待估计参数的，但是上述推导过程给出了一个迭代的估计参数的过程，并能保证似然逐步增加（后续证明）

给定一个GMM模型，优化目标是寻找使似然函数最大的各个高斯成分的均值向量、协方差矩阵和混合系数

1. 初始化　　初始化参数$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$

2. E步　　使用当前参数计算后验概率

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{6}$$

3. M步　　使用后验重新估计参数

$$\boldsymbol{\mu}_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \tag{8}$$

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k^{new})(\mathbf{x}_n - \boldsymbol{\mu}_k^{new})^T \tag{9}$$

$$\pi_k^{new} = \frac{N_k}{N}, \qquad N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

4. 重算似然　　重新计算似然函数，重复2-4，直至满足收敛条件

GMM模型

- **GMM模型和K-means模型的联系**
  - K-means可以看成GMM模型一个特殊情况，假设公式(1) 中，每个组件的高斯分布都具有相同的协方差矩阵，并且有$\boldsymbol{\Sigma} = \epsilon\mathbf{I}$，$\mathbf{I}$为单位矩阵，高斯分布可以简化为

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi\epsilon)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x} - \boldsymbol{\mu}_k\|^2\right\}$$

公式(6)中的后验概率变为

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

$$\gamma(z_{nk}) = \frac{\pi_k \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x}_\mathrm{n} - \boldsymbol{\mu}_k\|^2\right\}}{\sum_j \pi_j \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x}_\mathrm{n} - \boldsymbol{\mu}_\mathrm{j}\|^2\right\}}$$

直观上来理解，当$\epsilon \to 0$时，$-\frac{1}{2\epsilon}\|\mathbf{x}_\mathrm{n} - \boldsymbol{\mu}_\mathrm{j}\|^2 \to -\infty$，$\exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x}_\mathrm{n} - \boldsymbol{\mu}_\mathrm{j}\|^2\right\} \to 0$，对于分母中，假如第m项$\|\mathbf{x}_\mathrm{n} - \boldsymbol{\mu}_{\boldsymbol{m}}\|^2$最小,那么分母上j=m这一项将在$\epsilon \to 0$的时候以最慢的速度趋于0，因此，只有当分子上$k = m$时，$\gamma(z_{nk}) \to 1$，$k \neq m$时，$\gamma(z_{nk}) \to 0$，很明显，此时$\gamma(z_{nk}) \to r_{nk}$

K-means是一种硬对齐方式，某个数据点只能对应到某个类别上，GMM是一种软对齐方式，使用后验概率来表示某个数据点由某个类别产生的概率！

EM算法

- **深入理解EM算法**
  - EM算法的目标是寻找潜变量模型的最大似然解，为了对EM算法做进一步解释，我们将所有待估计的参数用 $\boldsymbol{\theta}$ 表示（对于GMM，$\boldsymbol{\theta}=\{\pi_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}$）
  - 对数似然函数用完全数据的联合概率表示为：

$$\ln p(\mathbf{X}|\boldsymbol{\theta})=\ln\left\{\sum_{\mathbf{Z}}p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})\right\} \quad (10)$$

  - 使用EM算法，一般认为完全数据的联合概率分布的似然 $\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})$容易计算
  - 实际上，完全数据集合$\{\mathbf{X},\mathbf{Z}\}$ 无法获取，但是潜变量$\mathbf{Z}$的后验概率分布$p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})$可以进行估计，在E步，我们计算完全数据的似然$\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})$ 在$\mathbf{Z}\sim p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old})$时，关于变量$\mathbf{Z}$的期望

$$\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{old})=\sum_{\mathbf{Z}}p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old})\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})=E_{\mathbf{Z}\sim p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\mathrm{old}})}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})] \quad (11)$$

为什么会这么设计$\mathcal{Q}$?

M步，寻找使$\mathcal{Q}$函数最大的新的参数值

$$\boldsymbol{\theta}^{\mathrm{new}}=\arg\max_{\boldsymbol{\theta}}\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\mathrm{old}}) \quad (12)$$

$\mathcal{Q}$函数的设计，个人认为是EM算法的重点，当我们尝试使用EM算法来解决我们自己的问题时，也是需要明确$\mathcal{Q}$函数

EM算法

- **EM算法的通用步骤**

给定完全数据$\{X, Z\}$的联合概率分布$p(X, Z|\theta)$，待学习参数θ，优化的目标是寻找θ来最大化似然函数$p(X|\theta)$

1. 初始化　　初始化参数θ^{old}

2. E步　　计算潜变量的后验概率$p(Z|X, \theta^{old})$

3. M步　　使用后验重新估计参数

$$\mathcal{Q}(\theta, \theta^{old}) = \sum_{Z} p(Z|X, \theta^{old}) \ln p(X, Z|\theta) = E_{Z\sim p(Z|X,\theta^{old})}[\ln p(X, Z|\theta)]$$

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}).$$

4. 重算似然　　重新计算似然函数，重复2-4，更新参数$\theta^{old} \leftarrow \theta^{new}$,直至满足收敛条件

EM算法

- **重新考虑GMM模型参数估计**
  - 在开始介绍GMM模型，我们已经引入的GMM的的潜变量$\mathbf{Z}$，然而在之前（第9页）的推导中，并没有用到完全数据的联合概率分布，而是直接对不完全数据$\mathbf{X}$的对数似然进行了求解
  - 根据前述的EM算法的通用步骤，首先考虑完全数据的似然函数

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}} \tag{13}$$

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\} \tag{14}$$

  - 根据公式(4)，计算$\mathbf{Z}$的后验概率

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_{nk}} \tag{15}$$

  - 完全数据的对数似然关于潜变量的期望值

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\} \tag{16}$$

公式（16）对参数进行求导将简单很多，相比于公式（7），对数操作应用于单高斯分布，将极大简化求导

EM算法

- **EM算法的通用解释**
  - EM算法的一般假设是直接优化观测数据的似然$p(\mathbf{X}|\boldsymbol{\theta})$十分复杂，但是优化完全数据的似然$p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})$比较容易
  - 引入一个关于变量$\mathbf{Z}$的任意一个分布$q(\mathbf{Z})$

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{Z} q(\mathbf{Z}) \ln p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})} \right\}$$
$$= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})} \quad (16)$$

为什么这里可以引入任意一个$\mathbf{Z}$的分布？因为我们只知道$\mathbf{Z}$服从某一个分布，但是并不知道具体服从什么样的真实分布，但是后面的推导会证明，无论$\mathbf{Z}$的真实分布是什么都不会影响我们的推导

令

$$\mathcal{L}(q,\boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \quad (17)$$

$\mathcal{L}(q,\boldsymbol{\theta})$是$q(\mathbf{Z})$和$\boldsymbol{\theta}$的泛函

$$\mathrm{KL}(q||p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})} \quad (18)$$

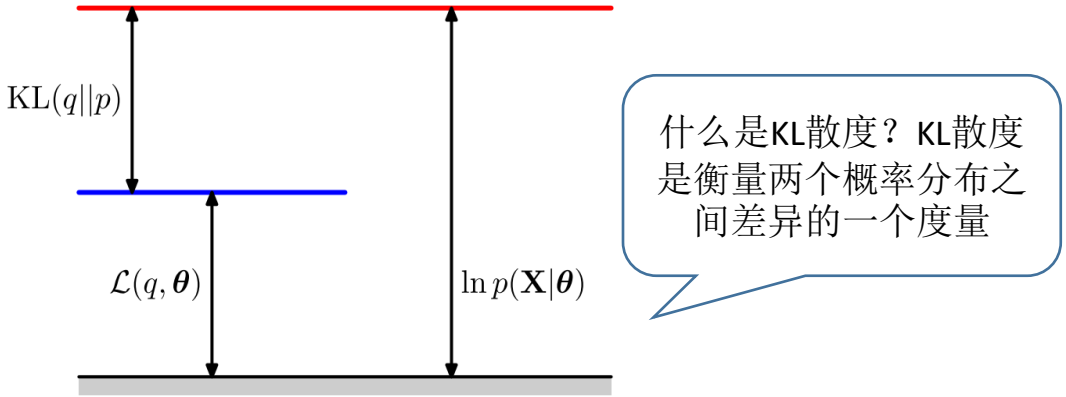
$\mathrm{KL}(q||p) \geq 0$
当且仅当$q = p$时等号成立

公式（16）可以表示为

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q,\boldsymbol{\theta}) + \mathrm{KL}(q||p) \quad (19)$$

EM算法

- **理解公式(19)** $\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q,\boldsymbol{\theta}) + \mathrm{KL}(q||p)$
  - $\mathrm{KL}(q||p) \geq 0$，故$\ln p(\mathbf{X}|\boldsymbol{\theta}) \geq \mathcal{L}(q,\boldsymbol{\theta})$，只有当后验分布$p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})$和$q(\mathbf{Z})$相等时，等号成立
  - $\mathcal{L}(q,\boldsymbol{\theta})$可以看作是$\ln p(\mathbf{X}|\boldsymbol{\theta})$的 下界！**如果我们可以无法直接提升$\ln p(\mathbf{X}|\boldsymbol{\theta})$的准确值，我们可以提升其下界**



公式(19)示意图

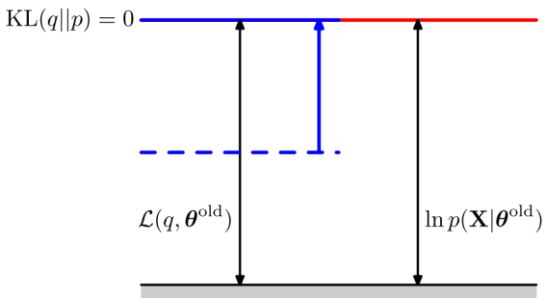
EM算法

- **理解公式(19)**

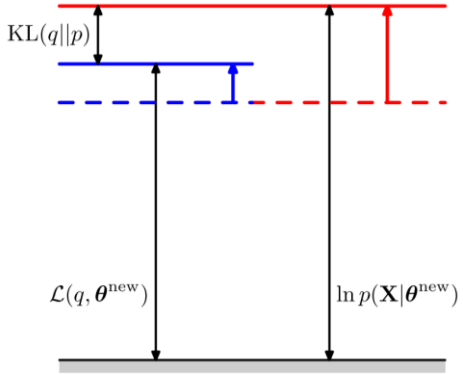
E步：寻找使$\mathcal{L}(q,\boldsymbol{\theta})$最大的$q(\mathbf{Z})$，当$q(\mathbf{Z})=p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})$时，$\mathrm{KL}(q||p)=0$，此时$\mathcal{L}(q,\boldsymbol{\theta})$最大

  - $\mathrm{KL}(q||p)\geq 0$，故$\ln p(\mathbf{X}|\boldsymbol{\theta})\geq\mathcal{L}(q,\boldsymbol{\theta})$，只有当后验分布$p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})$和$q(\mathbf{Z})$相等时，等号成立
  - $\mathcal{L}(q,\boldsymbol{\theta})$可以看作是$\ln p(\mathbf{X}|\boldsymbol{\theta})$的 下界！**如果我们可以无法直接提升$\ln p(\mathbf{X}|\boldsymbol{\theta})$的准确值，我们可以提升其下界**

E步：寻找使$\mathcal{L}(q,\boldsymbol{\theta})$最大的$q(\mathbf{Z})$，当$q(\mathbf{Z})=p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})$时，$\mathrm{KL}(q||p)=0$，此时$\mathcal{L}(q,\boldsymbol{\theta})$最大



M步：固定$q(\mathbf{Z})$，寻找使$\mathcal{L}(q,\boldsymbol{\theta})$增加的新参数$\boldsymbol{\theta}^{new}$，因为参数更新，$q(\mathbf{Z})$和$p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{new})$不再相等，此时$\mathrm{KL}(q||p)>0$，因此导致了$\ln p(\mathbf{X}|\boldsymbol{\theta}^{new})$的增加



EM算法

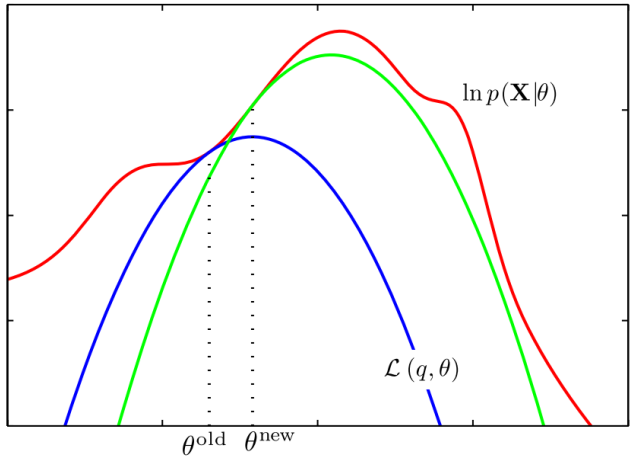
- **$\mathcal{L}(\mathbf{q},\boldsymbol{\theta})$与$\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\mathbf{old}})$的关系**
  - 当$q(\mathbf{Z})=p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old})$，$\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{old})$等价于$\mathcal{L}(q,\boldsymbol{\theta})$

$$\mathcal{L}(q,\boldsymbol{\theta})=\sum_{\mathbf{Z}}\boldsymbol{p}(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old})\ \ln \boldsymbol{p}(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})-\sum_{\mathbf{Z}}\boldsymbol{p}(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old})\ \ln \boldsymbol{p}(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old})$$
$$=\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{old})+const$$

因此，在之前的EM算法中，E步所计算的$\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{old})$实际上等价于计算$\mathcal{L}(q,\boldsymbol{\theta})$

- **从参数空间理解EM算法**

左图中，红线表示观测数据的对数似然，蓝线表示在原始参数下，下边界$\mathcal{L}(\mathbf{q},\boldsymbol{\theta})$的变化，在M步，最大化下边界$\mathcal{L}(\mathbf{q},\boldsymbol{\theta})$关于参数θ，得到新的参数，此时得到的新的下边界绿线所示，继续进行E步和M步，直到达到观测数据的最大似然。



本章总结

- GMM模型是一种非常重要的模型，它将在之后的语音识别应用中，同HMM模型一起来刻画语音数据的分布
- GMM模型的参数，通过EM算法估计，EM算法是一种通用的算法
- EM算法的假设是不完全数据的似然函数难以处理，但是完全数据的似然函数简单，便于处理
- 理解EM算法，关键是$Q(\theta, \theta^{old})$的设计思想
- 本章没有涉及图模型，但是在PRML第9章中，使用图模型帮助读者进行理解，但是对于没有接触过图模型的同学太过陌生，有兴趣同学参看书中第8章

作业

本次作业，实现基于GMM的0(o)-9孤立词识别系统

提供的数据：本次课程提供了330句训练预料，每个英文单词（0-9，o）含有30个句子用于训练对应的GMM，所有的训练数据和测试数据的原始音频路径、对应的抄本text（标注，0-9，o）、特征（feats.scp, feats.ark) 都在train和test文件夹下。原始音频的39维MFCC特征已经通过kaldi提取给出，代码中也给出了读取kaldi格式特征的代码。feats.scp 里面存储的是某句话的特征数据的真实文件和位置，特征实际存储在二进制文件feats.ark中（可以忽略kaldi特征部分，我们已经提供了特征读取代码，读取后可在python环境中查看）

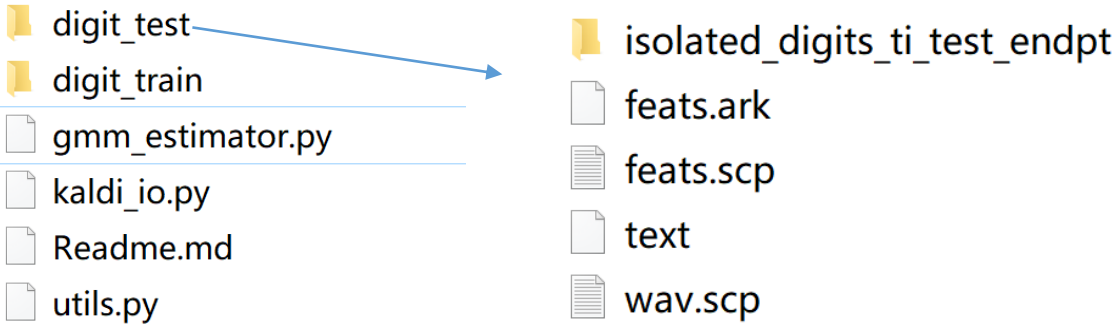
使用提供的特征，完善代码中GMM参数估计部分，并且用测试数据对其进行测试，统计错误率。每一个孤立词建立一个GMM模型，高斯成分个数（K）可以自定，特征维度是39维。

https://github.com/nwpuaslp/ASR_Course/tree/master/03-GMM-EM

作业

数据说明

https://github.com/nwpuaslp/ASR_Course/tree/master/03-GMM-EM



感谢各位聆听！