EM算法: ②

① 若数据分布可以建模为方差相同的高斯分布（混合），我可以使用 KMeans
若方差不相等，还想建模成高斯混合模型 → 就需要用EM算法来求参

② 直观理解高斯混合模型(GMM)

随机变量由K个高斯分布混合而成，取各个高斯分布的概率为 $\pi_1, \pi_2 \cdots \pi_k$

第i个高斯分布均值为 $\mu_i$，方差为 $\Sigma_i$，根据样本 $X_1, X_2 \cdots X_n$ 估计参数 $\pi, \mu, \Sigma$

e.g. 样本

$$\begin{Bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{Bmatrix} = \begin{array}{cccc} \text{身高} & \text{体重} & \text{胸围} & \text{腰围} \\ \end{array}\begin{Bmatrix} X_{11} & X_{12} & X_{13} & X_{14} \\ X_{21} & X_{22} & X_{23} & X_{24} \\ X_{31} & X_{32} & X_{33} & X_{34} \\ X_{41} & X_{42} & X_{43} & X_{44} \end{Bmatrix} \xrightarrow{\text{协方差}} \Sigma_4 = \begin{Bmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{Bmatrix}$$

均值 ↓ $\mu = \{\mu_1, \mu_2, \mu_3, \mu_4\}^T$

元素是任意2个特征的协方差

③ 高斯混合模型求解使用EM的原因

对数似然函数: $l_{\pi,\mu,\Sigma}(x) = \sum_{i=1}^{N} \log\left[\prod_{k=1}^{K} \pi_k \boxed{N(x_i \mid \mu_k, \Sigma_k)}\right]$

$= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

隐变量

由于隐变量的存在，无法用求导的方法求最大似然，需要用 EM 方法求参

④ EM 算法步骤概要:

问题描述: 随机变量X由K个高斯分布混合而成，取各个高斯分布的概率为 $\varphi_1, \varphi_2, \cdots \varphi_k$，第i个高斯分布的均值为 $\mu_i$，方差为 $\Sigma_i$

根据样本 $X_1, X_2, \cdots X_n$ 估计参数 $\mu, \varphi, \Sigma$

E-step: 选取一组参数 $\phi, \mu, \Sigma$，计算该参数下，隐变量的条件概率值

$$w_j^{(i)} = Q_i(z^{(i)}=j) = P(z^{(i)}=j \mid x^{(i)}; \phi, \mu, \Sigma) \quad // \text{for each } i,j$$

M-step会更新

样本i属于主题zj的条件概率

$$= \frac{p(x^{(i)} \mid z^{(i)}=j; \mu, \Sigma)\, p(z^{(i)}=j; \phi)}{\sum_{l=1}^{k} p(x^{(i)} \mid z^{(i)}=l; \mu, \Sigma)\, p(z^{(i)}=l; \phi)}$$

M-Step：结合上步求出的隐变量条件概率 $w_j^{(i)} = Q_i(z^{(i)}=j)$ (对任意i, j)

求出似然函数的下界函数(本质上是某个期望函数)的最大值.

<a> 对数似然函数：$l(\theta) = \sum_{i=1}^{m} \log p(x;\phi) = \sum_{i=1}^{m} \log \sum_{z} p(x, z;\phi)$

<b> 对数似然函数的下界函数：$\sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\phi)}{Q_i(z^{(i)})}$

推导：令 $Q_i$ 是隐变量 $Z$ 的某一个分布，$Q_i \geq 0$，有：

$$l(\theta) = \sum_{i=1}^{m} \log \sum_{z} p(x, z;\phi) = \sum_{i=1}^{m} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)};\phi)$$

$$= \sum_{i=1}^{m} \log \sum_{z^{(i)}} Q_i(z^{(i)}) \boxed{\frac{p(x^{(i)}, z^{(i)};\phi)}{Q_i(z^{(i)})}}$$ ←对□求期望再取对数

$$\geq \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \boxed{\frac{p(x^{(i)}, z^{(i)};\phi)}{Q_i(z^{(i)})}}$$ ←对□取对数再求期望

根据Jensen不等式　　m个特征，m个θ

<c> 求下界函数的最大值

$$\theta := \arg\max_{\theta} \sum_{i}^{n} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\phi)}{Q_i(z^{(i)})}$$

求解过程

$$\sum_{i=1}^{n} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\phi)}{Q_i(z^{(i)})}$$ 条件概率展开

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} Q_i(z^{(i)}=j) \log \frac{p(x^{(i)} \mid z^{(i)}=j, \mu, \Sigma)\, p(z^{(i)}=j;\phi)}{Q_i(z^{(i)}=j)}$$

代入E-Step的结果　　高斯分布

⊛ $$= \sum_{i=1}^{n} \sum_{j=1}^{k} w_j^{(i)} \log \frac{1}{w_j^{(i)}} \cdot \left( \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_j|^{\frac{1}{2}}} e^{-\frac{1}{2}(x^{(i)}-\mu_i)^T \Sigma_j^{-1} (x^{(i)}-\mu_i)} \right) \cdot \Phi_j$$

s.t. 约束条件　$\varphi_1 + \varphi_2 + \cdots + \varphi_k = 1$　K个主题概率之和为1

对于 $\mu_l$，$\Sigma_j$ 求解：对⊛关于 $\mu_l$ 或 $\Sigma_j$ 求偏导，且令偏导为0即可

对于 $\Phi = \{\varphi_1, \varphi_2, \cdots \varphi_k\}$ 求解：⊛式与约束条件则构成一个约束条件求极值问题

用拉格朗日乘子法求解.

<4> 求得的结果

$$\mu_l = \frac{\sum_{i=1}^{n} \boxed{w_l^{(i)}} x^{(i)}}{\sum_{i=1}^{n} \boxed{w_l^{(i)}}}$$

i是样本下标　j是特征序号

$$\Sigma_j = \frac{\sum_{i=1}^{n} \boxed{w_j^{(i)}} (x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T}{\sum_{i=1}^{n} \boxed{w_j^{(i)}}}$$

$$\Phi_j = \frac{1}{n} \sum_{i=1}^{n} \boxed{w_j^{(i)}}$$

将M-Step更新出的 $\mu_l, \Sigma_j, \Phi_j$ 用于 E-Step 重新计算 $w_j^{(i)}$

$$\text{EStep}: \boxed{w_j^{(i)}} = Q_i(z^{(i)}=j) = P(z^{(i)}=j \mid x^{(i)}; \boxed{\phi, \mu, \Sigma})$$

$$= \frac{P(x^{(i)} \mid z^{(i)}=j; \mu, \Sigma)\, P(z^{(i)}=j; \phi)}{\sum_{l=1}^{k} P(x^{(i)} \mid z^{(i)}=l; \mu, \Sigma)\, P(z^{(i)}=l; \phi)}$$

直到收敛.

# pLSA模型（Probabilistic Latent Semantic Analysis）

<a> 概述

$$\xrightarrow[d_i]{P(d_i)} \text{D} \xrightarrow{P(z_k|d_i)} \text{Z} \xrightarrow{P(w_j|z_k)} \text{W}$$

多项分布 $P(z_k|d_i)$　　多项分布 $P(w_j|z_k)$

$P(d_i)$：文档$d_i$出现的概率

$P(z_k|d_i)$：文档$d_i$中主题$z_k$出现概率（隐变量）

$P(w_j|z_k)$：主题$z_k$出现单词$w_j$的概率.

与LDA类似，用文档生成的模拟过程来适合理解这个算法，即

以$P(d_i)$的概率选中文档$d_i$，以$P(z_k|d_i)$的概率选中主题$z_k$，以$P(w_j|z_k)$的概率产生一个词$w_j$ → 这些词还原出文档的词袋.

<b> 小公式

$$P(d_i, w_j) = \boxed{P(w_j|d_i)}\, P(d_i) \quad ※（后面用）$$

文档 词　代入

$$\boxed{P(w_j|d_i)} = \sum_{k=1}^{K} \underline{P(w_j|z_k)}\, \underline{P(z_k|d_i)}$$

目标：求这两个多项分布.

<c> 似然函数

$$L = \prod_{i=1}^{N}\prod_{j=1}^{M} P(d_i, w_j)^{\boxed{n(d_i, w_j)}}$$

$n(d_i, w_j)$：词$w_j$在$d_i$文档中出现次数

对数似然

$$l = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j)\log P(d_i, w_j) \quad \text{代入※}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j)\log \sum_{k=1}^{K} \underline{P(w_j|z_k)}\, \underline{P(z_k|d_i)}\, P(d_i)$$

未知，且由隐变量介入 → 无法求偏导

<d> 采用类似EM算法的方式求最大似然.

E-step：假定$P(z_k|d_i)$、$P(w_j|z_k)$已知，求隐变量$z_k$的后验概率.

$$P(z_k|d_i, w_j) = \frac{P(w_j|z_k)\, P(z_k|d_i)}{\sum_{l=1}^{K} P(w_j|z_l)\, P(z_l|d_i)}$$

M-step：在$(d_i, w_j, z_k)$已知的前提下，求关于参数$P(z_k|d_i)$、$P(w_j|z_k)$的似然期望的最大值，从而得到当前阶段$P(z_k|d_i)$、$P(w_j|z_k)$的最优解

（方法见下一页）

(d1) 变换对数似然函数，方便求解

$$\ell = \sum_i \sum_j n(d_i, w_j) \log \underline{P(d_i, w_j)}$$

$$= \sum_i \sum_j n(d_i, w_j) \log (P(w_j|d_i) \cdot P(d_i))$$

$$= \sum_i \sum_j n(d_i, w_j) (\log P(w_j|d_i) + \log P(d_i))$$

$$= \sum_i \sum_j n(d_i, w_j) \log P(w_j|d_i) + \boxed{\sum_i \sum_j n(d_i, w_j) \log P(d_i)}$$

从样本可直接计算 当作常数去掉

$$\ell_{new} = \sum_i \sum_j n(d_i, w_j) \log P(w_j|d_i)$$

$$\cancel{E(\ell_{new}) = \sum_i \sum_j n(d_i, w_j) \log P(w_j|d_i)}$$

在 $z_k$ 上的数学期望

$$E(\ell_{new}) = \sum_i \sum_j n(d_i, w_j) \sum_{k=1}^{K} P(z_k|d_i, w_j) \log P(w_j, z_k|d_i)$$

求最大值

$$= \sum_i \sum_j n(d_i, w_j) \sum_{k=1}^{K} P(z_k|d_i, w_j) \log P(w_j|z_k) P(z_k|d_i)$$

同时，有约束条件

$$\begin{cases} \sum_{j=1}^{M} P(w_j|z_k) = 1 \\ \sum_{k=1}^{K} P(z_k|d_i) = 1 \end{cases}$$

(d2) 这是一个约束条件求极值问题，用拉格朗日乘子法

解得

$$\begin{cases} P(w_j|z_k) = \dfrac{\sum_i n(d_i, w_j) \boxed{P(z_k|d_i, w_j)}}{\sum_{m=1}^{M} \sum_i n(d_i, w_j) \boxed{P(z_k|d_i, w_j)}} \\ P(z_k|d_i) = \dfrac{\sum_j n(d_i, w_j) \boxed{P(z_k|d_i, w_j)}}{\sum_{k=1}^{K} \sum_j n(d_i, w_j) \boxed{P(z_k|d_i, w_j)}} \end{cases}$$

E-Step：M-Step 的结果代回到 E-Step

$$P(z_k|d_i, w_j) = \frac{\boxed{P(w_j|z_k)} \boxed{P(z_k|d_i)}}{\sum_{l=1}^{K} \boxed{P(w_j|z_l)} \boxed{P(z_l|d_i)}}$$

反复迭代，直到收敛

<e> pLSA 与 LDA

pLSA 不需要先验信息就可以完成自学习，这是它的优势.

如果要有先验知识的影响呢? 这时使用 LDA (需要超参数)

LDA 的超参数.

① K: 主题个数

② $\alpha$: 主题~~自~~鲜明度

③ $\beta$: 主题突出度 (是否受无关词影响)