聚类

K-Means

K-Means算法是无监督的聚类算法，它实现起来比较简单，聚类效果也不错，因此应用很广泛。假设有一组数据$\{x_1, x_2, ..., x_m\}$，区别于有监督学习的情况这些数据不再有标记y_i。我们通常假设存在隐变量$\{z_1, z_2, ..., z_n\}$，能够使得数据再一定的目标函数下被正确分类。再k-means算法中，z_i表示聚类中心.。

流程：

1. 初始化聚类中心$\{z_1, z_2, ..., z_n\}$
2. repeat until convergence
    1. 对数据分类，$C_i := \arg\min_j ||x_i - z_j||_2$
    2. 更新聚类中心，$z_j = \frac{\sum_{i=1}^m \mathbb{I}\{C_i=j\}x_i}{\sum_{i=1}^m \mathbb{I}\{C_i=j\}}$

容易证明上述过程实际上是对目标函数：

$$J(C_i, z) = \sum_{i=1}^m ||x_i - z_{C_i}||_2^2$$

分别对变量x_i, z使用坐标上升法求解，这不是个凸优化的问题，所以有可能会得到局部最优解。

EM算法

EM算法是一种迭代法，当样本中具有无法观测的隐变量时，他可以求解目标的极大似然估计，或最大后验概率

Jensen不等式

对于$f(x)$是凹函数，

$$f(E[x]) \geq Ef([x])$$

当且仅当$x_i = x_2 = ... = x_m$时等号成立，凸函数则反之。

假设有观测数据集$\{x_1, x_2, ..., x_m\}$，服从概率分布$p(x;\theta)$，存在隐变量$\{z_1, z_2, ..., z_n\}$，我们的目标是求参数θ，可以写成对数似然函数，

$$\begin{aligned}\ell(\theta) = \sum_{i=1}^m \ln P(x;\theta) = \sum_{i=1}^m \ln \sum_{z_j} P(x, z_j; \theta) = \sum_{i=1}^m \ln \sum_{z_j} Q_i(z_j) \frac{P(x, z_j;\theta)}{Q_i(z_j)} \\ \geq \sum_{i=1}^m \sum_{z_j} Q_i(z_j) \ln \frac{P(x, z_j;\theta)}{Q_i(z_j)}\end{aligned}$$

E-Step

为使不等式等号成立，可以得到到隐变量的概率分布

$$Q(z_j) \sim P(x, z_j; \theta) = P(z_j|x; \theta) = \frac{P(x, z_j; \theta)}{\sum_j P(x, z_j; \theta)}$$

M-Step

对对数似然函数的下界函数做极大似然估计，求解参数θ

$$\theta = \arg\max_{\theta} \sum_{i=1}^{m} \sum_{z_j} Q_i(z_j) \ln \frac{P(x, z_j; \theta)}{Q_i(z_j)}$$

重复E-M步直到收敛

文本聚类

我们有数据集$\{x^1, x^2, \dots, x^m\}$，其中$x^i \in \{0, 1\}$，每个数据表示$x_k^i = \mathbb{I}\{word_k \in document_i\}$，存在隐变量$\{z^1, z^2, \dots, z^n\}$，隐变量满足二项分布，$\phi(z^i = 1) = \phi_1$。先验概率

$$P(x_k^i = 1|z^i = j) = \phi_{kj}^i$$

E-Step:

$$q_{ij} = Q_i(z^j) = P(z^j|x^i; \phi_{kj}^i, \phi_j) = \frac{\prod_{k=1}^d (\phi_{kj}^i)^{x_k^i} (1 - \phi_{kj}^i)^{1-x_k^i} \phi_j}{\sum_j \prod_{k=1}^d (\phi_{kj}^i)^{x_k^i} (1 - \phi_{kj}^i)^{1-x_k^i} \phi_j}$$

在第一步时我们可以初始化参数ϕ_{kj}^i, ϕ_1，得到隐变量的概率分布。

M-Step:

下界函数的极大似然估计，

$$\ell' = \sum_{i=1}^{m} \sum_{z_j} q_{ij} \ln \frac{P(x, z_j; \theta)}{q_{ij}} = \sum_{i=1}^{m} \sum_{j} \left(q_{ij} \ln \frac{\prod_{k=1}^d (\phi_{kj}^i)^{x_k^i} (1 - \phi_{kj}^i)^{1-x_k^i} \phi_j}{q_{ij}} \right)$$

对参数求导

$$\frac{\partial \ell'}{\partial \phi_{kj}^i} = \sum_{i=1}^{m} q_{ij} \left(\frac{x_k^i}{\phi_{kj}^i} - \frac{1 - x_k^i}{1 - \phi_{kj}^i} \right) = 0$$

$$\phi_{kj}^i = \frac{\sum_{i=1}^m q_{ij} x_k^i}{\sum_{i=1}^m q_{ij}}, \quad \phi_1 = \frac{\sum_{i=1}^m q_{i1}}{\sum_{i=1}^m \sum_j q_{ij}} = \frac{\sum_{i=1}^m q_{i1}}{m}$$

高斯混合模型

类似于高斯判别式分析，这里假设有隐变量$\{z^1, z^2, \dots, z^n\}$，满足多项分布$P(z^k) = \phi_k, \sum_{k=1}^n \phi_k = 1$。

有数据集$\{x^1, x^2, \dots, x^m\}$，则有

$$P(x^i|z = z^k) = N(x^i; \mu_k, \Sigma_k)$$

E-Step:

$$q_{ik} = Q_i(z^k) = \frac{N(x^i; \mu_k, \Sigma_k)\phi_k}{\sum_{k=1}^n N(x^i; \mu_k, \Sigma_k)\phi_k}$$

M-Step:

下界函数的极大似然估计,

$$\ell' = \sum_{i=1}^{m}\sum_{k} q_{ik} \ln \frac{N(x^i;\mu_k,\Sigma_k)\phi_k}{q_{ik}}$$

同理对各个参数求导

$$\mu_j = \frac{\sum_{i=1}^m q_{ij}x^i}{\sum_{i=1}^m q_{ij}}, \quad \Sigma_j = \frac{\sum_{i=1}^m q_{ij}(x_i-\mu_j)(x_i-\mu_j)^T}{\sum_{i=1}^m q_{ij}}, \quad \phi_k = \frac{\sum_{i=1}^m q_{ik}}{m}$$

因子分析模型

在高斯混合模型中，当$d >> m$时，协方差矩阵是奇异的。因子分析模型(factor analysis model)可以很好的解决这个问题。

多元高斯分布

多元高斯分布的条件分布和边缘分布，假设一个随机向量$x=[x_1,x_2], x_1\in R^r, x_2 \in R^s$，x满足分布

$$x \sim N(\mu,\Sigma), \quad \mu=[\mu_1,\mu_2]^T, \quad \Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$$

其中$\mu_1\in R^r, \mu_2 \in R^s, \Sigma_{11}\in R^{r*r}, \Sigma_{12}\in R^{r*s}, ets$。

边缘分布:x_1的边缘分布$x_1 \sim N(\mu_1,\Sigma_{11})$

条件分布:条件分布$x_1|x_2 \sim N(\mu_{1|2},\Sigma_{1|2})$，其中

$$\begin{aligned}\mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)\\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\end{aligned}$$

假设一个关于(x,z)的联合密度，其中$z\in R^K$是隐变量:

$$z\sim N(0,I), \quad x|z \sim N(\mu+\Lambda z,\Phi)$$

其中$\mu\in R^m, \Lambda \in R^{m*K}, \Phi\in R^{m*m}$。

令$x=\mu+\Lambda z+\epsilon, \epsilon\sim N(o,\Phi)$，由此看出z和x满足联合高斯分布:

$$\begin{bmatrix}x\\ z\end{bmatrix} \sim N(\mu_{zx},\Sigma), \quad \mu_{zx} = \begin{bmatrix}E(x)\\ E(z)\end{bmatrix} = \begin{bmatrix}\mu\\ 0\end{bmatrix}, \quad \Sigma = \begin{bmatrix}\Sigma_{xx} & \Sigma_{zx}\\ \Sigma_{xz} & \Sigma_{zz}\end{bmatrix}$$

计算谐方差矩阵

$$\Sigma_{zz} = I, \Sigma_{zx} = E[(z-E[z])(x-E[x])^T] = \Lambda^T, \quad \Sigma_{xx} = \Lambda\Lambda^T + \Phi$$

最终可以得到x的边缘分布

$$x\sim N(\mu, \Lambda\Lambda^T+\Phi)$$

直接对x进行极大似然估计将会很麻烦。但是可以使用EM算法求解。

E-Step:

$$\begin{aligned}\mu_{z|x} &= \mu_z + \Sigma_{zx}\Sigma_{xx}^{-1}(x-\mu_x) = \Lambda^T(\Lambda\Lambda^T+\Phi)^{-1}(x-\mu)\\ \Sigma_{z|x} &= \Sigma_{zz} - \Sigma_{zx}\Sigma_{xx}^{-1}\Sigma_{xz} = I - \Lambda^T(\Lambda\Lambda^T+\Phi)^{-1}\Lambda\end{aligned}$$

所有可以得到

$$q_{ij} = Q_i(z=j) = P(z=j|x^i;\mu,\Phi,\Lambda) \sim N(\mu_{z|x}, \Sigma_{z|x})$$

M-Step:

$$\begin{aligned}
\ell' &= \sum_{i=1}^{m} \sum_{j} q_{ij} \ln \frac{P(x^i|z=j;\mu,\Phi,\Lambda)P(z=j)}{q_{ij}} \\
&= \sum_{i=1}^{m} E_{z\sim Q_i} \ln \frac{P(x^i|z=j;\mu,\Phi,\Lambda)P(z=j)}{q_{ij}} \\
&= \sum_{i=1}^{m} E_{z\sim Q_i} \ln \frac{N(\mu+\Lambda z,\Phi)P(z=j)}{q_{ij}}
\end{aligned}$$

我们知道条件概率$x|z \sim N(\mu+\Lambda z,\Phi)$，对上式各个参数求导。

在因子分析模型的讨论中可以发现，因子分析模型就是一种降维的概率模型。