聚类模型

1、层次聚类

2、原型聚类-K-means

3、模型聚类-GMM

4、EM算法-LDA主题模型

5、密度聚类-DBSCAN

6、图聚类-谱聚类

一、层次聚类

一、聚类理论

传统来说，聚类是在训练样本的标签信息不知的情况下，学习样本内在的性质和规律，将有限的集合划分成c类。根据“方以类聚，物以群分”的思想，类内对象尽可能的相似，类间对象尽可能不相似。因此，吾师言：聚类中两个关键的问题是：何为类？何为类内相似，类间不相似？以下所有的聚类模型皆从这两点出发。

由于缺少样本标签，我们很难定义类和相似性，比如下面的问题：

按照颜色聚类可以分类三类，按照形状聚类可以分类两类，关键问题在于如何定义类，定义相似性。所以吾师还言：聚类一般不是一个任务的最终目标，而是一个预处理的过程。

聚类的评价指标有两种：

1）内部指标，指导思想是类内紧致性和类间分离性，比如Xie-Beni指标，DB指标

2）外部指标，假设数据集有标注，按有监督学习的评价指标进行评价

可以看出，外部指标有很大的问题，那就是聚类学到的数据规律不一定是标签，这对聚类算法的评价是不可靠的，但是对于只看结果，不评价模型的好坏是可以的，当然拿聚类的结果与有监督学习的结果对比是“无赖”的。

二、层次聚类

层次聚类的类表示可以看作是基于样本的，X_i表示属于第i的样本集合，即X_i作为第i类的类表示。类相似性度量可以用“欧式距离”。层次聚类分为两种，一种是自底向上的凝聚层次聚类，一种是自顶向下的分裂层次聚类。两者的区别在于前者一开始将每一个样本看作一类，通过不断的合并最相似的两个簇，直到c类；后者一开始将所有样本看作一类，通过最小化损失（类紧致）分裂为c类。

凝聚层次聚类：

输入：样本数据$D = x_1, x_2, \ldots, x_m$，相似性度量函数$s$，聚类簇数$k$

输出：k类样本

1）初始化每个样本为一个簇，$c_i = x_i, i = 1, 2, \ldots, m$

2）计算样本两两之间的距离$d(i, j), i = 1, 2..., m, j = 1, 2, \ldots, m$

3）通过相似性度量函数s，找出最相似的两个簇进行合并

最小距离：$s = d_{min}(c_i, c_j) = \min_{x \in c_i, z \in c_j} dist(x, z)$

最大距离：$s = d_{max}(c_i, c_j) = \max_{x \in c_i, z \in c_j} dist(x, z)$

平均距离：$s = d_{avg}(c_i, c_j) = \frac{1}{|c_i||c_j|} \sum_{x \in c_i} \sum_{z \in c_j} dist(x, z)$

4)直到簇数为k，否则循环2）

分裂层次聚类：

输入：样本数据$D = x_1, x_2, \ldots, x_m$，损失函数$s$，聚类簇数$k$

输出：k类样本

1）初始化所有样本为一个簇，$c_1 = \{x_i\}, i = 1, 2, \ldots, m$

2）计算样本两两之间的距离$d(i, j), i = 1, 2..., m, j = 1, 2, \ldots, m$

3）计算当前所有簇c_i，$i \in 1, 2..n < k$的损失函数，选择损失最大的簇进行二分

计算该簇下两点间距离$d(i, j), i = 1, 2..., m_i, j = 1, 2, \ldots, m_j$，选择簇中最远的两个点作为类中心将簇进行二分

4)直到簇数为k，否则循环2）

值得注意的是分裂层次聚类在进行二分时，可以采用kmeans进行二分，这样时间复杂度就不再是$O(m^2)$.

层次聚类算法特点：

1）可视化

2）采用计算样本两两之间的距离，时间复杂度为$O(m^2)$

3）凝聚和分裂的不可逆性

二、原型聚类-KMeans

KMeans的类表示是聚类中心点，以点x_i来表示类，相似性度量同样可以采用常用的距离度量。根据类紧致性准则定义失真函数为所有样本点到该样本所在类中心的失真程度和最小。

$$J(c)=\sum_{i=1}^{m}||x^{(i)}-c_j^{(i)}||^2, i=1,2..m, j=1,2..k$$

其中$c_j^{(i)}$表示第i个样本所属的类。可以看出Kmeans算法只考虑了类内相似性，没有考虑类间相似性。对于Kmeans算法的求解采用EM算法，先假设类中心$c_1^1,c_2^1,..,c_k^1$，然后根据相似性度量来划分所有样本点到k类中（Kmeans是一种硬划分），根据划分后的样本点重新更新k类的类中心$c_1^2,c_2^2,..,c_k^2$，不断的迭代至稳定（类中心不再变化）。

KMenas算法流程：

1）随机初始化类中心$c_1^1,c_2^1,..,c_k^1$（选择样本中的点，或者不是样本中的点）

2）重复以下步骤直到收敛

a）遍历所有的样本点$i=1,2...m$，根据相似性度量（欧式距离）将样本划分到最相似性的类

$$c^{(i)}=arg\min_j||x^{(i)}-c_j||^2, i=1,2..m, j=1,2..k$$

其中$c^{(i)}$表示第i个样本所属的类别，值为与类中心距离最近的一类j.

b）遍历所有样本，对每一类$c_1^t,c_2^t,..,c_k^t$，更新类中心（该类下所有样本的均值）

$$c_j^t:=\frac{\sum_{i=1}^m 1\{c^{(i)}=j\}x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)}=j\}}$$

至于Kmeans算法的收敛性可以从EM算法角度证明，因为每一次迭代都能保证失真函数不增，所以最终一定会趋于平衡，由于类别数有限，所以有限步收敛。

可以看出在Kmeans中，所有的类划分都是硬划分，下面介绍一点软化分的模糊C均值聚类。失真函数如下：

$$J_{FCM}=\sum_{j}^{k}\sum_{i=1}^{m}\mu_{ij}^{\gamma}||x^{(i)}-c_j^{(i)}||^2, i=1,2..m, j=1,2..k$$

其中μ_{ij}表示第i个样本属于j类的概率，且$\sum_{i=1}^k u_{ij}=1$，γ控制失真程度，当$\gamma=1$时，软化分也等同于硬划分，因为失真函数还是线性的，所以一般取$\gamma>1$。拉格朗日乘子可求解参数μ_{ij}和c_j

模糊C均值算法流程：

1）随机初始化类中心$c_1^1,c_2^1,..,c_k^1$

2）重复以下步骤直到收敛

a）遍历所有的样本点$i=1,2...m$，更新概率划分矩阵$\mu_{ij}\in R^{m\times k}$

$$\mu_{ij}=\frac{1}{\sum_{j=1}^k\left(\frac{||x_i-c^{(i)}||^2}{||x_i-c_j||^2}\right)^{\frac{1}{\gamma-1}}}$$

其中$c^{(i)}$表示第i个样本所属的类别.

b）遍历所有样本，对每一类$c_1^t,c_2^t,..,c_k^t$，更新类中心（该类下所有样本的均值）

$$c_j^t:=\frac{\sum_{i=1}^m \mu_{ik}^{\gamma}x^{(i)}}{\sum_{i=1}^m \mu_{ik}^{\gamma}}$$

Kmeans特点：

1）复杂度为mkt，线性复杂度

2）需要预先指定聚类个数，失真还是非凸，最小值为局部最小，对初始值选取敏感

3）硬划分会出现空类，即一个类下面无样本点

4）相似性度量决定其适合聚类的场景

三、模型聚类-高斯混合

高斯混合的类表示是一个高斯模型，相似性度量定义为服从类c_j高斯分布μ, Σ的概率（Kmeans的相似度量是聚距离度量），所以高斯混合聚类也可以看作是有参的密度聚类。高斯混合假设类之间服从伯努利分布，样本在某一类下服从高斯分布，也就是说每个样本独立服从多元高斯分布。为了使得所有样本的概率最大化，即最大化对数似然函数：

$$\begin{aligned} L(\Phi, \mu, \Sigma) &= log \prod_{i=1}^{m} P(x^{(i)}) \\ &= \sum_{i=1}^{m} log(P(x^{(i)}; \Phi, \mu, \Sigma)) \\ &= \sum_{i=1}^{m} log \sum_{z^{(i)}=1}^{k} (P(x^{(i)}|z^{(i)}; \mu, \Sigma) P(z^{(i)}; \Phi)) \end{aligned}$$

也就是说假设类之间服从一个伯努利分布：

$$P(z^{(i)} = c_j) = P(z^{(i)}; \Phi) = \Phi_j, \sum_{j=1}^{k} \Phi_j = 1, j = 1, 2..k$$

样本在类$z^{(i)}$下的条件概率服从高斯分布：

$$P(x^{(i)} = z^{(i)}|z^{(i)}) = \mathcal{N}(\mathbf{x}; \mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu) \right\}$$

那么样本$x^{(i)}$和类标签$z^{(i)}$的联合分布为：

$$\begin{aligned} P(x^{(i)}, z^{(i)}) &= P(x^{(i)}|z^{(i)}) P(z^{(i)}) \\ &= \sum_{i=1}^{m} log \sum_{z^{(i)}=1}^{k} (P(x^{(i)}|z^{(i)}; \mu, \Sigma) P(z^{(i)}; \Phi)) \end{aligned}$$

以上，当$z^{(i)}$已知时，即标签信息已知的话，类似于高斯判别分析（当然，高斯判别分析中多个高斯分布之间具有相同的协方差），对应的且只属于一类（类标已知），那么上式有：

$$= \sum_{i=1}^{m} log \left(P(x^{(i)}|z^{(i)}; \mu, \Sigma) + log P(z^{(i)}; \Phi) \right)$$

最大似然估计有参数：

$$\begin{aligned}\boldsymbol{\Phi}_j &= \sum_{i=1}^{m} \frac{1\{z^{(i)}=j\}}{m} \\ \mu_j &= \frac{\sum_{i=1}^{m} 1\{z^{(i)}=j\}x^{(i)}}{\sum_{i=1}^{m} 1\{z^{(i)}=j\}} \\ \boldsymbol{\Sigma}_j &= \frac{\sum_{i=1}^{m} 1\{z^{(i)}=j\}\left(x^{(i)}-\mu_j\right)\left(x^{(i)}-\mu_j\right)^T}{\sum_{i=1}^{m} 1\{z^{(i)}=j\}}\end{aligned}$$

可以看出Φ_j为每一类样本所占的比例，μ_j为该类下样本的均值，Σ_j为该类下样本的协方差。 考虑到高斯混合模型中的类划分是概率划分$w_j^{(i)}$，表示第i个样本属于第j类的概率。所以，高斯混合模型的所有参数都需要乘上类的划分概率$w_j^{(i)}$。

高斯混合模型流程：

1）初始化参数隐类别数$Z_j, j=1,2..k$，模型参数Φ, μ, Σ

2）采用EM算法，先假设参数，期望最大化，然后更新样本的划分概率更新参数

a）E-step：对所有样本$x_i, i=1,2..m$，根据参数划分每个样本类概率：

$$\begin{aligned}w_j^{(i)} &:= p(z^{(i)}=j|x^{(i)};\Phi,\mu,\Sigma) \\ &= \frac{p(x^{(i)}|z^{(i)}=j;\Phi,\mu,\Sigma)p(z^{(i)}=j;\Phi)}{\sum_{l=1}^{k} p(x^{(i)}|z^{(i)}=j;\Phi,\mu,\Sigma)p(z^{(i)}=j;\Phi)}\end{aligned}$$

b）M-step：根据划分后的类概率更新参数

$$\begin{aligned}\boldsymbol{\Phi}_j &= \sum_{i=1}^{m} \frac{w_j^{(i)}}{m} \\ \mu_j &= \frac{\sum_{i=1}^{m} w_j^{(i)} 1\{z^{(i)}=j\}x^{(i)}}{\sum_{i=1}^{m} 1\{z^{(i)}=j\}} \\ \boldsymbol{\Sigma}_j &= \frac{\sum_{i=1}^{m} w_j^{(i)}\left(x^{(i)}-\mu_j\right)\left(x^{(i)}-\mu_j\right)^T}{\sum_{i=1}^{m} w_j^{(i)}}\end{aligned}$$

这里可以看出，当w_j^i是已知的，即类标签已知，则直接进行参数估计等价于高斯判别分析，当$w_j^{(i)}$是硬划分，同Kmenas又是一致的。

四、EM算法

一、EM算法

EM算法是一种迭代算法，用于带隐变量的概率模型参数的极大似然估计，是无监督学习中一大类算法求解的算法。EM算法每次迭代由两步组成，E步：假设隐变量和特征变量的联合分布$P(x,z;\theta)$，求解样本关于隐变量z的概率函数（使Jensen不等式等号成立），M步：在已知**样本**(x,z)的联合分布（确定样本特征和类标），采用极大似然估计最大化似然函数求解参数θ。

在讨论EM算法之前，先介绍Jensen inequality（由凸函数性质导出）

假设f是定义在实数上的凸函数，由凸函数的定义有：

$$f(\lambda x^{(1)} + (1-\lambda)x^2) \leq \lambda f(x^{(1)}) + (1-\lambda)f(x^2)$$

严格凸函数则严格大于，凸函数的判定是其二阶可微的话，其Hesse矩阵半正定。对凸函数的性质推广有：

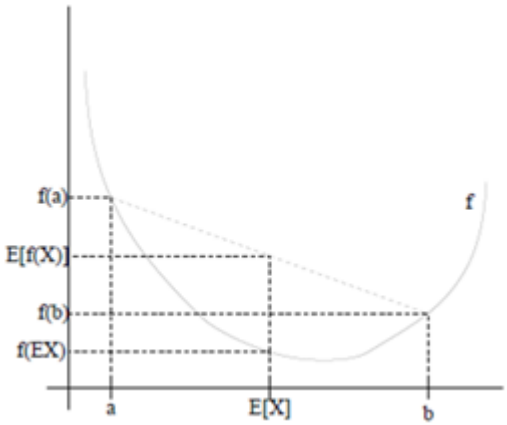
$$f(\sum_{i=1}^{k}(\lambda_i x^{(i)})) \leq \sum_{i=1}^{m}\lambda_i f(x^{(i)})$$

$$s.t. \sum_{i=1}^{m}\lambda_i = 1, \lambda_i \geq 0$$

当λ_i表示$f(x^{(i)}), x^{(i)}$的概率时，那么有：

$$f(E(x)) \leq E(f(x))$$

当且仅当：$p(f(x)=c)=1$，即$f(x)$为常数函数，等号成立。



反之，对于凹函数不等式方向相反。

现在来看EM算法，给定训练样本$\{x^{(1)}, x^{(2)}, ..x^{(m)}\}$，引入隐含的类别标签$z^{(i)}$，在有监督方法中，最大对数似然函数$L=p(z|x;\theta)$，同样这里最大化对数似然函数的$L=(x^{(i)};\theta)$在隐变量$z^{(i)}$的全期望：

$$\begin{aligned} L(\theta) &= \sum_{i=1}^{m} logP(x^{(i)};\theta) \\ &= \sum_{i=1}^{m} log\sum_{z^{(i)}} P(x^{(i)}, z^{(i)};\theta) \\ L(\theta, Q(z)) &= \sum_{i=1}^{m} log\sum_{z^{(i)}} Q_i(z^{(i)})\frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} \\ &\geq \sum_{i=1}^{m}\sum_{z^{(i)}} Q_i(z^{(i)}) log\frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} \end{aligned}$$

其中$Q_i(z^{(i)})$为样本的隐变量$z^{(i)}$的概率分布，$\sum_z Q_i(z^{(i)}) = 1, Q_i(z^{(i)}) \geq 0$。不同$Q(z)$选择，会得到EM在不同情况下的模型，比如高斯混合，朴素贝叶斯混，LDA等。

因为log函数是一个严格凹函数，由Jessen不等式有：

$$log(E(g(x)) \geq E(log(g(x))$$

$$log\left(\sum_{z^{(i)}} Q_i(z^{(i)}) \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}\right) \geq \sum_{z^{(i)}} Q_i(z^{(i)}) log \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

其中$g(x) = \frac{P(x^{(i)}, z^{(i)}|\theta)}{Q_i(z^{(i)})}$，因此当且仅当，$g(x) = c$，等号成立。

因为$\sum_z Q_i(z^{(i)}) = 1, Q_i(z^{(i)}) \geq 0$，所以$Q_i(z^{(i)})$可以看做是样本关于隐变量$z$的概率分布，等于$x, z$联合概率归一化，即比上联合概率对$z$的全期望：

$$\begin{aligned} Q_i(z^{(i)}) &= \frac{p(x^{(i)}, z^{(i)}; \theta)}{\sum_z p(x^{(i)}, z^{(i)}; \theta)} \\ &= \frac{p(x^{(i)}, z^{(i)}; \theta)}{p(x^{(i)}; \theta)} \\ &= p(z^{(i)}|x^{(i)}; \theta) \end{aligned}$$

因此，EM算法的第一步就是计算在给定θ下隐变量z的条件概率。

当已知$Q_(z)$之后，且Jessen不等式等号成立，回过头来再最大化似然函数：

$$\begin{aligned} arg\max_\theta L(\theta, Q(z)) &= \sum_{i=1}^m log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \\ &= \sum_{i=1}^m \sum_{z^{(i)}} Q_i(z^{(i)}) log \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \end{aligned}$$

因此，EM算法的极大似然估计可以看作是坐标上升过程，第一步在给定的参数θ下最大化似然函数$L(Q(z); \theta)$，第二步则是在当前的$Q(z)$下最大化似然函数$L(\theta; Q(z))$。

收敛性：

$$\begin{aligned} L(\theta^{t+1}) &= \sum_{i=1}^m log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{P(x^{(i)}, z^{(i)}; \theta^{t+1})}{Q_i(z^{(i)})} \\ &\geq \sum_{i=1}^m \sum_{z^{(i)}} P(z^{(i)}|x^{(i)}, \theta^t) log \frac{P(z^{(i)}, x^{(i)}; \theta^{t+1})}{P(z^{(i)}|x^{(i)}; \theta^t)} \\ &\geq \sum_{i=1}^m \sum_{z^{(i)}} P(z^{(i)}|x^{(i)}; \theta^t) log \frac{P(z^{(i)}, x^{(i)}; \theta^t)}{P(z^{(i)}|x^{(i)}; \theta^t)}) \\ &= L(\theta^t) \end{aligned}$$

$$L(Q(z), \theta^{(i)}) \leq L(Q(z), \theta^{(i+1)}) \leq L(Q^*(z), \theta^{(i+1)})$$

下面的第一个不等号由由最大似然估计得到$\theta^{(i+1)}$，第二个不等号Jessen不等式得到$Q^*(z) = p(z^{(i)}|x^{(i)}; \theta)$，但是求解过程是先由Jessen不等式确定似然函数的下界，后最大似然函数下界。

EM算法的一般流程：

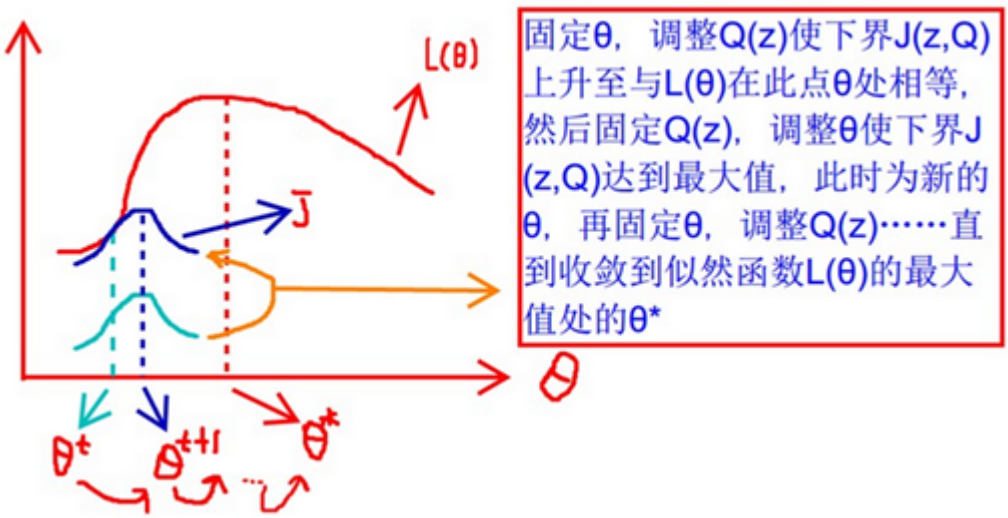
E-step：（固定参数下，求隐变量条件分布）

$$Q_i(z^{(i)}) = p(z^{(i)}|x^{(i)}; \theta)$$

M-step：（最大化似然函数下界）

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

EM求解的过程大致如图所示，是否能收敛到全局最优，取决于目标函数是否为凸函数：



从上图可以看出，EM算法的思想也是坐标上升的思想，即固定一组变量求解另一组变量，不断地迭代。比较特殊的，EM算法针对于带隐变量的概率模型参数的极大似然估计，求解过程又称“期望最大化”，第一步求期望，第二步最大化，这是带隐变量的概率模型特有的，不是EM算法的特点。

二、EM算法例子-高斯混合，朴素贝叶斯混合

高斯混合模型

为什么采用EM算法求解高斯混合模型，回顾高斯混合模型的目标函数，我们发现log函数在求和外面。特别的情况是当类标签已知，像高斯判别模型那么进行参数估计，然而在混合高斯模型中。而隐变量却是未知，所以我们很难直接优化，采用EM算法的Jessen不等式，我们将log函数放到里面，并使等号成立，对目标函数进行转化：

$$L(\theta)=\sum_{i=1}^{m}log\sum_{z^{(i)}=1}^{k}(P(x^{(i)},z^{(i)};\mu,\Sigma,\Phi))$$

$$L(\theta,)\sum_{i=1}^{m}\sum_{z^{(i)}}Q_i(z^{(i)})log\frac{P(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})}$$

其中$Q_i(z^{(i)})$条件概率分布$P(z|x;\theta)$。

下面从EM思想来看高斯混合模型，给出高斯混合模型最大似然估计计算出参数的推导过程：

E-step:（固定参数下，求隐变量条件分布）

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

其中$p(x|z)$服从高斯分布，$p(z)$服从伯努利分布。

M-step:（最大化似然函数下界）

$$
\begin{aligned}
arg\max_{\theta} L(\theta, Q(z)) &= \sum_{i=1}^{m} log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \\
&\geq \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) log \frac{P(x^{(i)}|z^{(i)}; \mu, \Sigma) P(z^{(i)}; \phi)}{Q_i(z^{(i)})} \\
&= \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} log \frac{\frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\} * \phi_j}{w_j^{(i)}}
\end{aligned}
$$

对μ, Σ, ϕ求导：

$$
\begin{aligned}
&\bigtriangledown_{\mu_j} \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} log \frac{\frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\} * \phi_j}{w_j^{(i)}} \\
&= - \bigtriangledown_{\mu_j} \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} \left\{ \frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) \right\} * \phi_j \\
&= \frac{1}{2} \sum_{i=1}^{m} w_j^{(i)} \bigtriangledown_{\mu_j} 2\mu_j^T \Sigma_j^{-1} x^{(i)} - \mu_j^T \Sigma_j^{-1} \mu_j \\
&= \Sigma_{i=1}^{m} w_j^{(i)} (\Sigma_j^{-1} - \Sigma_j^{-1} \mu_j)
\end{aligned}
$$

令导数为0，有：

$$
\mu_j = \frac{\sum_{i=1}^{m} w_j^{(i)} 1\{z^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{z^{(i)} = j\}}
$$

对目标函数求解参数ϕ，去掉无关项，有：

$$
\sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} log \phi_j
$$

由$\sum_{j=1}^{k} \phi_j = 1$，对其拉格朗日函数求导：

$$
\begin{aligned}
\bigtriangledown_{\phi_j} L(\phi, \beta) &= \bigtriangledown_{\phi_j} \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} log \phi_j + \beta((\sum_{j=1}^{k} \phi_j) - 1) \\
&= \sum_{i=1}^{m} \frac{w_j^{(i)}}{\phi_j} + \beta
\end{aligned}
$$

令其导数为0，有：

$$
\phi_j = \frac{\sum_{i=1}^{m} w_j^{(i)}}{-\beta}
$$

所以，$\phi_j \propto \sum_{i=1}^{m} w_j^{(i)}$，又$\sum_j \phi_j = 1$，所以得$-\beta = \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} = m$

所以：

$$\phi_j = \sum_{i=1}^{m} \frac{w_j^{(i)}}{m}$$

朴素贝叶斯混合

考虑文本聚类问题，采用One-hot编码，则一篇文档的类别可以看作是隐变量z服从伯努利分布，文档中词x在类别z下的条件概率也可看作是一个伯努利分布（如果文档采用词在字典中的位置表示，则对应一个N元伯努利分布）。所以有：

$$\phi_{z=1} = p(z=1), \phi_{z=0} = 1 - p(z=1)$$

$$\phi_{j=1|z=1} = p(x_j^{(i)} = 1|z=1), \phi_{j=0|z=1} = 1 - p(x_j^{(i)} = 1|z=1)$$
$$\phi_{j=1|z=0} = p(x_j^{(i)} = 1|z=0), \phi_{j=0|z=1} = 1 - p(x_j^{(i)} = 1|z=0)$$

其中由于概率和为1，所有的变量我们只需要求解一般，即$\phi_z, \phi_{z=1}, \phi_{z=0}$。

由于朴素贝叶斯中词（特征）的独立性假设有：

$$P(x^{(i)}, z^{(i)}) = \prod_{j=1}^{n} P(x_j^{(i)}, z^{(i)}), j = 1, 2..n$$

其中j为特征下标。

同样，最大化对数似然函数：

$$\sum_{i=1}^{m} log \prod_{j=1}^{n} P(x_j^{(i)}, z^{(i)}), j = 1, 2..n$$

套入EM框架：

E-step:（固定参数下，求隐变量条件分布）

$$w^{(i)} = p(z^{(i)} = 1|x^{(i)}; \phi_j) = \frac{p(x^{(i)}|z^{(i)} = 1)p(z^{(i)} = 1)}{\sum_{j=0}^{1} p(x^{(i)}|z^{(i)} = j)p(z^{(i)} = 1)}$$

由于该聚类是二聚类，$w^{(i)}$表示是属于其中一类的概率，样本$x^{(i)}$属于另一类的概率则为$1 - w^{(i)}$。

M-step:（最大化似然函数下界）

$$\begin{aligned} arg\max_{\theta} L(\theta, Q(z)) &= \sum_{i=1}^{m} log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \\ &= \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) log \frac{P(x^{(i)}|z^{(i)}; \mu, \Sigma)P(z^{(i)}; \phi)}{Q_i(z^{(i)})} \\ &= \sum_{i=1}^{m} \left[w^{(i)} log \frac{p(x^{(i)}, z^{(i)} = 1; \phi_z)}{w^{(i)}} + (1 - w^{(i)}) log \frac{p(x^{(i)}, z^{(i)} = 0; \phi_z)}{w^{(i)}} \right] \end{aligned}$$

其中，

$$p(x^{(i)}, z^{(i)} = 1; \phi_j) = \prod_{j=1}^{n} p(x_j^{(i)} | z^{(i)} = 1)\phi_z$$

$$p(x^{(i)}, z^{(i)} = 0; \phi_j) = \prod_{j=1}^{n} p(x_j^{(i)} | z^{(i)} = 0)\phi_z$$

似然函数对ϕ_z求导：

$$\begin{aligned}\bigtriangledown_{\phi_z} L(\theta, Q(z)) &= \bigtriangledown_{\phi_z} \sum_{i=1}^{m} \left[w^{(i)} log\phi_z + (1 - w^{(i)}) log(1 - \phi_z) \right] \\ &= \sum_{i=1}^{m} \left[\frac{w^{(i)}}{\phi_z} - \frac{1 - w^{(i)}}{1 - \phi_z} \right]\end{aligned}$$

令偏导为0，得：

$$\phi_z = \frac{\sum_{i=1}^{m} w^{(i)}}{m}$$

对$\phi_{j|z=1}$求偏导：

$$\bigtriangledown_{\phi_{j|z=1}} L(\phi_{j|z}, \phi_z) = \sum_{i=1}^{m} w^{(i)} log p(x_j^{(i)} | z^{(i)} = 1)$$

$$= \sum_{i=1}^{m} w^{(i)} log \left[(\phi_{j|z=1})^{I(x_j^{(i)}=1)} (1 - \phi_{j|z=1})^{(1 - I(x_j^{(i)}=1))} \right]$$

$$= \sum_{i=1}^{m} \left[\frac{w^{(i)} I(x_j^{(i)} = 1)}{\phi_{j|z=1}} - \frac{w^{(i)} (1 - I(x_j^{(i)} = 1))}{1 - \phi_{j|z=1}} \right]$$

令偏导为0，得：

$$\phi_{j|z=1} = \frac{\sum_{i=1}^{m} w^{(i)} I(x_j^{(i)} = 1)}{\sum_{i=1}^{m} w^{(i)}}$$

同理$\phi_{j|z=0}$，求导令其偏导为0有：

$$\phi_{j|z=0} = \frac{\sum_{i=1}^{m} w^{(i)} I(x_j^{(i)} = 0)}{\sum_{i=1}^{m} 1 - w^{(i)}}$$

可以看出，朴素贝叶斯混合和高斯混合惊人的相似。M-step的参数估计与有监督最大似然参数估计的结果一致。

三、LDA主题模型

在了解主题模型之前，我们顺着朴素贝叶斯混合模型在多元伯努利分布上的推广导出隐语义分析（pLSA）模型。在PLSA模型中，我们假设隐变量z的语义是主题，而一篇文档涉及多个主题，不同的主题下产生词的概率不同。那么一篇文档的生成的概率可以写作：

$$P(w_j, d_i) = P(d_i) \sum_{z} P(z_k | d_i) P(w_j | z_k)$$

其中$p(d_i)$表示第i篇文档被选中的概率，$p(z_k|d_i)$表示第i篇文档生成第k个主题的概率，$p(w_j|z_k)$表示第k个主题下产生词w_j的概率。其中后两个概率服从多项分布。采用EM算法，我们同样可以在E-step假设参数，来确定主题的条件概率；在M-step最大化似然函数的下界更新参数。

LDA同pLSA极为相似，不同的是pLSA是频率学派的角度来看待文档-主题-词的关系，而LDA是贝叶斯学派角度来看待文档-主题-词关系。频率学派认为数据服从参数一定的概率分布$p(x|\theta)$，贝叶斯学派则从数据中估计参数的概率$p(\theta|x)$，认为参数本身服从一个先验概率$p(\theta)$，由贝叶斯公式，最大化后验概率：

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta)$$

也就是说LDA比pLSA多了两个先验分布：

$$P(w_j, d_i|\vec{\alpha}, \vec{\beta}) = p(z_k|\theta_{(i)})p(\theta_{(i)}|\vec{\alpha})p(w_j^{(i)}|z_k, \phi_k)p(\phi_k|\vec{\beta})$$

其中i表示文档，k表示主题，j表示词。

LDA模型从贝叶斯角度引入先验概率的目的是构造似然函数的一个共轭先验分布，用实践α, β作为超参来调节似然函数模型。关于LDA的求解可以采用吉布斯采样和变分的EM算法，这里不细究，在专门介绍LDA时再详细学习。

四、因子分析

五、密度聚类-DBSCAN

DBSCAN的类表示是一簇密度可达的样本，相似性度量定义为密度可达，密度可达即为一类，属于硬划分。密度聚类是一种基于密度的聚类，其根据样本的空间分布关系进行聚类。一般来讲，用带参的模型来定义样本的分布可以看作是带参的密度估计，比如高斯混合模型，高斯判别分析；用无参的模型来描述样本的分布称为无参密度估计，比如直方图，核密度估计，山峰聚类，DBSCAN，meanshift。

假设我的样本集是$(x_1, x_2, \ldots, x_m)$，在DBSCAN中为了描述样本分布的关系，定义了如下几个概念：

1）ϵ-邻域：对于$x_j \in D$，其ϵ-邻域包含样本集D中与x_j的距离不大于ϵ的子样本集，即
$N_\epsilon(x_j) = \{x_i \in D|distance(x_i, x_j) \leq \epsilon\}$ 这个子样本集的个数记为$|N \in (x_j)|$

2）核心对象：对于任一样本$x_j \in D$，如果其ϵ-邻域对应的$N \in (x_j)$ 至少包含$MinPts$个样本，即如果
$|N \in (x_j)| \geq MinPts$，则$x_j$是核心对象

3）密度直达：如果x_i位于x_j的ϵ-邻域中，且x_j是核心对象，则称x_i由x_j密度直达。注意反之不一定成立，即此时不能说x_j由x_i密度直达, 除非且x_i也是核心对象。

4）密度可达：如果x_i由x_j密度直达，且x_j由x_k密度直达，那么x_i由x_k密度可达。密度可达满足封闭性

其中**密度可达**是相似性度量，由于密度可达具有封闭性，所以簇内的所有点与簇内的核心均密度可达，否则即不是一个簇，所以密度可达可以对样本进行聚类，其中密度可达涉及的参数有ϵ和$MinPts$和距离度量$distance(x_i, x_j)$。

5）噪声点：对于非核心点和不能由核心点密度可达的点即为噪声点

DBSCAN算法流程：

输入：样本集$D = (x_1, x_2, \ldots, x_m)$，邻域参数$(\epsilon, MinPts)$， 样本距离度量方式

输出：簇划分$C=\{c_1,c_2,..c_k,...,c_K\}$

1）初始化核心对象为$\Omega=\varnothing$，簇划分$C=\varnothing,k=0$，未访问样本集合$\Gamma=D$

2）变量所有样本点，找出$|N\in(x_j)|\geq MinPts$的核心对象x_j并入核心对象Ω

3）如果核心对象$\Omega=\varnothing||\Gamma=\varnothing$，算法结束，否则从核心对象集$\Omega$中随机选择一个核心对象$x_j$，初始化当前簇$c_{k++}=x_j$，当前核心对象集为$\Omega_{cur}=x_j$，循环：

a）从当前的核心对象集Ω_{cur}中选择x_j，循环：

aa）找出x_j密度直达的未访问的点x_i加入到c_k，把其中未访问的核心对象加入到Ω_{cur}

ab）更新当前核心对象集$\Omega_{cur}=\Omega_{cur}-x_j$

ac）更新核心对象集$\Omega=\Omega-\Omega_{cur}$，更新未访问样本集$\Gamma=\Gamma-c_k$

b）如果当前核心对象集$\Omega_{cur}=\varnothing$，结束内部循环

4）输出簇划分$C=\{c_1,c_2,..c_k,...,c_K\}$

可以看出，外层循环为簇的个数，内层循环是构建一个密度可达的簇。当算的执行完$\Gamma\neq\varnothing$，对应既不是核心点，也不是密度可达的点我们称为异常点或者噪声点。

DBSCAN的特点：

1）由于密度可达的定义，DBSCAN具有发现任意形状的簇划分

2）聚类的同时可发现异常点，抗噪性强

3）不需要预先指点类数，但ϵ和$MinPts$的直观性不强，参数选择麻烦

4）样本集较大时，聚类收敛时间较长，密度估计存在维度灾难问题

5）如果样本集的密度不均匀、聚类间距差差别很大时，聚类质量较差，这时用DBSCAN聚类一般不适合

六、图聚类-谱聚类

谱聚类是一种定义在图上的聚类算法，与其说是聚类算法，更是一种图的向量表示。基于向量表示之后，一般可以采用其他的聚类方法完成最后聚类结果。所以谱聚类的类表示既依赖于向量表示也与之后采用的聚类算法有关。

对于一个图G，我们一般用点的集合V和边的集合E来描述。即为$G(V,E)$。其中V即为我们数据集里面所有的点$(v_1,v_2,...v_m)$。 谱聚类根据图上节点之间的关系（关系度量：ϵ邻域，k近邻图，全连接图），构建一个邻接矩阵来描述m个节点之间的相似性：

$$W=\begin{bmatrix}w_{11}&w_{12}&.&w_{1m}\\w_{21}&.&&w_{2m}\\.&&w_{ij}&.\\w_{m1}&w_{m2}&.&w_{mm}\end{bmatrix}$$

由节点之间关系的对称性，显然相似性矩阵W是对称矩阵。现在，我们希望学习到节点的向 量表示$x_i,i=1,2..m$，使得相似性越大的两个节点i,j的向量表示$x_i,x_j\in R^n$的差异尽可能的小，因此，我们可以定义如下损失函数：

$$\min \sum_{i,j=1}^{m} w_{ij} \frac{|x_i - x_j|^2}{|x_i||x_j|}$$

即当w_{ij}大时，相似性越大，$|x_i - x_j|^2$尽可能小。上式经过如下变换，也就得到了谱聚类与拉普拉斯矩阵的关系：

$$\begin{aligned}
\frac{1}{2}\sum_{i,j=1}^{m} w_{ij}(x_i - x_j)^2 &= \frac{1}{2}(\sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}x_i^2 - 2\sum_{i,j=1}^{m} w_{ij}x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}x_j^2) \\
&= \sum_{i}^{m} d_i x_i^2 + \sum_{j=1}^{m} d_j x_j^2 - 2\sum_{i,j=1}^{m} w_{ij}x_i x_j \\
&= \sum_{i=1}^{n} d_i x_i^2 - \sum_{i,j=1}^{n} w_{ij}x_i x_j \\
&= tr(DXX^T) - tr(WXX^T) \\
&= tr(X^T DX) - tr(X^T WX) \\
&= tr(X^T LX)
\end{aligned}$$

其中d_i, d_j是w_{ij}按行求和（按列求和），因此矩阵D为w_{ij}的按行求和（按列求和)的对角矩阵。

$$\begin{bmatrix}
d_{11} & 0 & . & 0 \\
0 & . & & 0 \\
. & & d_{ii} & . \\
0 & 0 & . & d_{mm}
\end{bmatrix}$$

其中$X \in R^{m\times n}$其中$L = D - W$，我们称L为拉普拉斯矩阵。

因此，当我们约束$|x_i| = 1, i = 1, 2..m$时，我们的目标函数为：

$$\begin{aligned}
&\min tr(X^T LX) \\
\leftrightarrow &\sum_{i=1}^{K} X_k^T LX_k = \sum_{i=1}^{K} LX_k X_k^T \\
&s.t.\ X^T X = \mathbf{I}
\end{aligned}$$

其中$X_k \in R^{m\times 1}$表示所有样本在k维构成的向量，由$X^T X = \mathbf{I} \rightarrow X_k^T X_k = 1$.所以目标函数右乘$X_k$有$\sum_{i=1}^{K} LX_k = \lambda X_k$，因此，最小化目标函数等价$L$的前$K$个最小特征值相加，对应的$X$为前$K$个最下特征值对应的特征向量构成。就此目标函数求解问题转变为特征向量求解问题。

得到图节点的向量表示之后，后面就可以采用常用的聚类算法进行聚类，比如Kmeans。

谱聚类算法流程：

1）确定图上节点关系度量，得到相似性度量矩阵

2）根据相似性度量矩阵得到拉普拉斯矩阵

3）对拉普拉斯矩阵求解前K个最小特征值对应的特征向量，即为节点的向量表示

4）采用聚类算法对节点向量进行聚类

谱聚类特点：

1）相似性度量矩阵限制了数据的表示为m^2

2）谱聚类对相似性度量矩阵的向量表示存在损失

3）谱聚类的向量表示数学形式非常漂亮，代码实现方便

4）聚类的效果与相似性度量矩阵的计算，表示，以及最终采用的聚类算法有关