问题

k-means和GMM（高斯混合模型）都是聚类算法，这两者其实也有一定的相似之处，值得我们探究一下。通过之前的整理，我们对k-means算法已经有了一定的了解，这里就着重补充一下GMM的内容以及它们之间的区别与联系。

高斯混合模型（GMM）

● 定义：高斯混合模型是指具有如下形式的概率分布模型：

$$P(x|\theta)=\sum_{k=1}^{K}\alpha_k\phi(x|\theta_k) \tag{1}$$

其中，α_k是高斯混合系数，$\alpha_k \geq 0$ 且 $\sum_{k=1}^{K}\alpha_k=1;\theta_k=(\mu_k,\sigma_k^2);$

$\phi(x|\theta_k)$是第k个高斯分布模型的概率密度函数，具体形式如下：

$$\phi(x|\theta_k)=\frac{1}{\sqrt{2\pi}\sigma_k}exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$$

GMM聚类

高斯混合模型（GMM）聚类的思想和 k-means 其实有点相似，都是通过迭代的方式将样本分配到某个簇类中，然后更新簇类的信息，不同的是GMM是基于概率模型来实现的，而 k-means 是非概率模型，采用欧氏距离的度量方式来分配样本。

GMM聚类主要思想和流程：

每个GMM由K个混合成分组成，每个混合成分都是一个高斯分布，α_k 为相应的混合系数。GMM模型假设所有的样本都根据高斯混合分布生成，那么每个高斯分布其实就代表了一个簇类。具体流程如下：

1. 先初始化高斯混合模型的参数 $\{(\alpha_k,\mu_k,\sigma_k^2)\mid 1\leq k\leq K\}$，训练一个GMM模型需要估计这些参数，如何估计后面会介绍。
2. 对每个样本，固定各个高斯分布，计算样本在各个高斯分布上的概率（即该样本是由某个高斯分布生成而来的概率）。
3. 然后固定样本的生成概率，更新参数以获得更好的高斯混合分布。
4. 迭代至指定条件结束。

> 上面的1-4给出了GMM算法的大致思想，虽然简略了一些，但对比k-means算法的思想一起来看应该也很容易理解。
>
> ● k-means初始化K个均值，GMM初始化K个高斯分布和相应的混合系数；
>
> ● k-means计算样本到各个簇中心的欧氏距离并选择最小距离来划分样本应该属于哪个簇类，而GMM给出的是样本由某个高斯分布生成而来的概率，比如有80%的概率是由A分布生成的，有20%的概率是B分布生成的。这一点在医学诊断上很有意义，比如相比于k-means算法会很硬性地认为某位病人得了肿瘤或正常，GMM给出病人有51%的概率患有肿瘤这样的结果往往会更有参考意义。
>
> ● 两者都是采用迭代的方式来不断更新参数，以求得最优解（都是局部最优解）。

EM算法估计GMM参数

上面提到，要训练一个GMM模型，就需要估计每个高斯分布的参数 $\{(\alpha_k, \mu_k, \sigma_k^2) \mid 1 \leq k \leq K\}$，才能知道每个样本是由哪个高斯混合成分生成的，也就是说，数据集的所有样本是可观测数据，$\{(\alpha_k, \mu_k, \sigma_k^2) \mid 1 \leq k \leq K\}$ 这些是待观测数据(隐变量)，而估计待观测数据常用的算法就是EM算法。

下面给出EM算法估计高斯混合模型的参数的步骤，详细的推导过程可以参考《统计学习方法》第9.3节的内容：

1. 给定数据集$D = \{x_1, x_2, \ldots, x_m\}$，初始化高斯混合分布的模型参数 $\{(\alpha_k, \mu_k, \sigma_k^2) \mid 1 \leq k \leq K\}$。
2. **E步：**遍历每个样本，对每个样本 x_i，计算其属于第k个高斯分布的概率：

$$\gamma_{ik} = \frac{\alpha_k \phi(x_i|\theta_k)}{\sum_{k=1}^{K} \alpha_k \phi(x_i|\theta_k)}, \quad \text{其中，} \theta_k = (\mu_k, \sigma_k^2) \tag{2}$$

3. **M步：**更新各个高斯分布的参数为$\{(\hat{\alpha}_k, \hat{\mu}_k, \hat{\sigma}_k^2) \mid 1 \leq k \leq K\}$：

$$\begin{aligned} \hat{\alpha}_k &= \frac{\sum_{i=1}^{m} \gamma_{ik} x_i}{\sum_{i=1}^{m} \gamma_{ik}} \\ \hat{\mu}_k &= \frac{\sum_{i=1}^{m} \gamma_{ik} (x_i - \mu_k)^2}{\sum_{i=1}^{m} \gamma_{ik}} \\ \hat{\sigma}_k^2 &= \frac{\sum_{i=1}^{m} \gamma_{ik}}{m} \end{aligned} \tag{3}$$

4. 重复2-3步，直至收敛。

> 注意，EM算法通过迭代的方式估计GMM模型的参数，得到的是**局部最优解**而不是全局最优。

在了解了EM算法后，让我们再来看看高斯混合聚类是怎么操作的吧。。。

在迭代收敛后，遍历所有的样本，对于每个样本 x_i，计算它在各个高斯分布中的概率，将样本划分到概率最大的高斯分布中（每个高斯分布都相当于是一个簇类，因此可以理解为是将每个样本划分到相应的类别中，不过实际上是给出属于每个类别的概率而非属于某个类别）。

k-means和GMM算法的区别与联系

终于要回到正题了，不过相信从上面的分析看下来，应该对这两种算法的区别与联系已经有了大致理解了吧，下面就再来总结一下：

区别：

① k-means算法是非概率模型，而GMM是概率模型。

> 具体来讲就是，k-means算法基于欧氏距离的度量方式来将样本划分到与它距离最小的簇类，而GMM则是计算由各个高斯分布生成样本的概率，将样本划分到取得最大概率的高斯分布中。

② 两者需要计算的参数不同。

> k-means计算的是簇类的均值，GMM计算的是高斯分布的参数（即均值、方差和高斯混合系数）

③ k-means是硬聚类，要么属于这一类要么属于那一类；而GMM算法是软聚类，给出的是属于某些类别的概率。

④ GMM每一步迭代的计算量比k-means要大。

联系：

① 都是聚类算法

② 都需要指定K值，且都受初始值的影响。k-means初始化k个聚类中心，GMM初始化k个高斯分布。

③ 都是通过迭代的方式求解，而且都是局部最优解。k-means的求解过程其实也可以用EM算法的E步和M步来理解。

参考资料

李航--《统计学习方法》

K-means算法和高斯混合模型的异同 https://blog.csdn.net/qq_38150441/article/details/80498590