统计学习：第十四章

1 试写出分裂聚类算法，自上而下地对数据进行聚类，并给出其算法复杂度。

答：输入：几个样本组成的样本集合及样本之间的距离；

输出：对样本集合的一个层次化聚类

1. 构造一个类，包含所有点

2. 计算每个类的类内方差

3. 挑选方差最大的类，应用 k-means 将其分为两个类

4，若类的个数为 n, 终止计算，否则回到步 2

算法复杂度为：$o(n^3m)$

2 证明类或簇的四个定义中，第一个定义可推出其他三个定义.

答:

1. 若定义 14.5 成立，可知任取一个样本 $x_j \in G$ 有 $d_ij \leq T$

2. 假设 G 为满足定义 14.5 的一个类，$\sum_(x_j \in G)d_{ij}/(n_G-1) \leq (n_G-1)*T/(n_G-1) = T$

3. 假设 G 为满足定义 14.5 的一个类，

$$\frac{1}{n_G(n_g-1)}\sum_{x_i\in G}\sum_{x_j\in G}d_{ij} \leq \frac{1}{n_G(n_g-1)}\sum_{x_i\in G}\sum_{x_j\in G}T = \frac{n_G(n_g-1)}{n_G(n_g-1)}T = T$$

3 证明式 (14.21) 成立，即 k 均值的可能解的个数是指数级的。

答：运用容斥原理，不考虑类为空时，共有 k^n 种分法，再减去至少一个类为空的情况，再加上至少两个类为空的情况，以此类推得到

$$S(n,k) = \frac{1}{k!}\sum_{l=1}^{k}(-1)^{(k-l)}C_k^l l^n$$

4 比较 k 均值聚类与高斯混合模型 EM 算法的异同。

答：二者都假设数据有多个中心生成，使用迭代优化，需要预先确定 k 的取值，二者都基于聚类的均值去估计中心位置，二者都依据决定类的归属或者概率

k 均值聚类使用硬聚类与高斯混合模型使用软聚类；k 均值聚类没有概率模型，高斯混合模型使用概率假设。