

K-Means: 根据相似度, 将 n 个数据点分到 k 个类中。

算法描述:

Lloyd's K-Means

1. 随机选定 k 个中心 u_k 的值。
2. 将每个数据点归类到离它最近的那个中心点所代表的 cluster 中。
3. 用公式 $\mu_k = \frac{1}{N_k} \sum_{j \in cluster_k} x_j$ 计算出每个 cluster 新的中心点。
4. 重复第二步, 一直迭代了最大的步数或者前后的 J 值相差小于一个阈值为止。

Hartigan's K-Means

1. 将 n 个数据点随机分配到 k 个 cluster 中。
2. 按某一顺序选取一个数据点 $x(x \in X)$, 将 x 从 X 中移除并且更新 $v_{c(x)}$ 。
3. 选取一个新的 cluster $c^* = \operatorname{argmin}_{c \in C} \Delta D(x, c)$, 更新 v_{c^*} 。
4. 重复第二步, 直至遍历完所有的数据点。如果存在 $c^* \neq c(x)$, 则再次遍历数据。