9.4.1 k均值算法

$D=\{x_1,\cdots,x_m\}$ $C=\{C_1,\cdots,C_k\}$ 最小化

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}\|x-\mu_i\|_2^2 \qquad \mu_i=\frac{1}{|C_i|}\sum_{x\in C_i}x$$

输入：样本集 $D=\{x_1,x_2,\cdots,x_m\}$

聚类簇数 k.

过程：

从D中随机选择k个样本作为初始化均值向量 $\{\mu_1,\mu_2,\cdots,\mu_k\}$.

repeat:

令 $C_i=\emptyset$ $(1\le i\le k)$

for $j=1,2,\cdots,m$ do

计算样本 x_j 与各均值向量 μ_i $(1\le i\le k)$ 距离：$d_{ji}=\|x_j-\mu_i\|_2$

根据距离最近的均值向量确定 x_j 的簇标记：$\lambda_j=\underset{i\in\{1,2,\cdots,k\}}{\arg\min} d_{ji}$.

将样本 x_j 划入相应簇：$C_{\lambda_j}=C_{\lambda_j}\cup\{x_j\}$

end for

for $i=1,2,\cdots,k$ do

计算新均值向量：$\mu_i'=\frac{1}{|C_i|}\sum_{x\in C_i}x$;

if $\mu_i'\neq\mu_i$ then

$\mu_i=\mu_i'$

until 当前均值向量均未更新.

输出：簇划分 $C=\{C_1,C_2,\cdots,C_k\}$.