K-means算法原理推导

给定mxn维度大小的样本集合 $X=\{x_1,x_2\cdots x_m\}$，K均值聚类是将m个样本划分到K个类别区域，通常K<m。其学习策略是通过最小化损失函数来选取最优划分。

假设使用欧式距离作为K均值聚类算法的距离度量方式，样本间的距离 d_{ij} 可定义为：

$$d_{ij}=\left(\sum_{i=1}^{n}(x_{ki}-x_{kj})^2\right)^{\frac{1}{2}}=(\|x_i-x_j\|^2)^{\frac{1}{2}}$$

定义样本与其所属类中心之间的距离总和为最终损失函数：

$$L(C)=\sum_{i=1}^{K}\sum_{C(i)=l}(\|x_i-\bar{x}_l\|^2)^{\frac{1}{2}}$$

其中，$\bar{x}_l=(\bar{x}_{1l},\bar{x}_{2l}\cdots\bar{x}_{ml})$ 为第l个类的质心（centroid），即类的中心点。。K均值聚类可以被约为一个优化问题来进行求解：

$$C^*=\arg\min_C L(C)$$

$$=\arg\min_C\sum_{l=1}^{K}\sum_{C(i)=l}(\|x_i-x_j\|^2)^{\frac{1}{2}}$$

该问题是一个NP难的组合优化问题，实际求解时我们采用迭代的方法。

(1) 初始化质心。即在第0次迭代时随机选择K个样本点作为初始化聚类质心。

(2) 按照样本与质心距离对样本进行聚类。

(3) 计算上一步聚类结果的新的质心。

(4) 如果迭代收敛或者满足迭代停止条件，则输出最后的聚类结果。