KMeans 推导

Soul Walker

2020 年 7 月 22 日

输入：训练数据集

$$T = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} = \{\mathbf{x}_i\}_{i=1}^N$$

其中，$\mathbf{x}_i \in \chi \in \mathbb{R}^N$ 为实例的特征向量，$\mathbf{x}_i = (\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, ..., \mathbf{x}_i^{(n)})^T$ ，$i = 1, 2, \cdots, N$ ，实例特征向量 $\mathbf{x}$，$\mathbf{X}$ 为训练样本集，形状为 (N, p)。

输出：簇的划分。

KMeans 算法用来解决无监督问题的分类问题。基本思想是对于给定的样本集，按照样本之间的距离大小（这里采用欧氏距离），将样本集划分为 K 个簇，并且让簇内的点尽量紧密排布，而让簇与簇之间的距离尽量的大。设给定要分的类别数为 K

距离度量（欧氏距离的平方）

$$distance(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{x}_j)^2$$

那么我们可以定义样本与其所属的类中心之间的距离总和为损失函数

$$J = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} d(\mathbf{x} - \mu_i)$$

其中

$$\mu_i = \frac{1}{C_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}$$

那么我们面临的优化问题就是

$$C^* = \arg\min_{C_i} \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} d(\mathbf{x} - \frac{1}{C_i} \sum_{x \in C_i} \mathbf{x})$$

那我们就可以通过迭代来达到优化问题的求解。