k-means算法原理

杨航锋

1 k-means算法的损失函数

假设输入空间 $\mathcal{X} \in \mathbb{R}^n$ 为 n 维向量的集合， $\mathcal{X} = \{x^{(1)}, x^{(2)}, \cdots, x^{(m)}\}$ ， $\mathcal{C}$ 为输入空间 $\mathcal{X}$ 的一个划分，不妨令 $\mathcal{C} = \{\mathbb{C}_1, \mathbb{C}_2, \cdots, \mathbb{C}_K\}$ ，因此可以定义 $k\text{-}means$ 算法的损失函数为

$$J(\mathcal{C}) = \sum_{k=1}^{K} \sum_{x^{(i)} \in \mathbb{C}_k} \|x^{(i)} - \mu^{(k)}\|_2^2$$

其中 $\mu^{(k)} = \frac{1}{|\mathbb{C}_k|} \sum\limits_{x^{(i)} \in \mathbb{C}_k} x^{(i)}$ 是簇 $\mathbb{C}_k$ 的聚类中心。

2 优化损失函数

$k\text{-}means$ 算法的损失函数 $J(\mathcal{C})$ 描述了簇类样本围绕簇聚类中心的紧密程度，其值越小，则簇内样本的相似度越高。故 $k\text{-}means$ 算法的优化目标为最小化损失函数

$$\arg\min_{\mathcal{C}} \ J(\mathcal{C}) = \sum_{k=1}^{K} \sum_{x^{(i)} \in \mathbb{C}_k} \|x^{(i)} - \mu^{(k)}\|_2^2$$

如果要优化该损失函数就需要考虑输入空间 $\mathcal{X}$ 的所有划分，这是一个 $NP\text{-}hard$ 问题，实际上是采取贪心的策略通过迭代优化来近似求解，该过程等价于 EM 算法。

1. 首先随机初始化 K 个聚类中心， $\mu^{(1)}, \mu^{(2)}, \cdots, \mu^{(K)}$ ；
2. 然后根据这 K 个聚类中心给出输入空间 $\mathcal{X}$ 的一个划分， $\mathbb{C}_1, \mathbb{C}_2, \cdots, \mathbb{C}_K$ ；
    - 样本离哪个簇的聚类中心最近，则该样本就划归到那个簇

    $$\arg\min_{k} \ \|x^{(i)} - \mu^{(k)}\|_2^2$$
3. 再根据这个划分来更新这 K 个聚类中心

    $$\mu^{(k)} = \frac{1}{|\mathbb{C}_k|} \sum_{x^{(i)} \in \mathbb{C}_k} x^{(i)}$$
4. 重复2、3步骤直至收敛
    - 即 K 个聚类中心不再变化