机器学习第八章聚类作业

162050127 颜劭铭

2022 年 6 月 13 日

1 基础题

1.1 给定样本集 $\mathcal{D} = \{x_1, x_2, \cdots, x_n\}$,k-means 聚类算法希望得到簇划分 $\mathcal{C} = \{C_1, C_2, \cdots, C_k\}$，使得最小化欧式距离

$$J(\gamma, \mu_1, \ldots, \mu_k) = \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2 \tag{1}$$

其中 $\mu_1, \cdots, \mu_k$ 为第 k 个簇的中心 (means)，$\gamma \in \mathbb{R}^{n \times k}$ 为指示矩阵 (indicator matrix) 定义如下：若 x_i 属于第 j 个簇，则 $\gamma_{ij} = 1$，否则为 0，则最经典的算法流程如算法 1 所示

1.1.1 试证明，在算法 1 中，Step 和 Step2 都会使目标函数 J 的值降低

Step1:

$$\gamma_{ij} = \begin{cases} 1, & \|\mathbf{x}_i - \mu_j\|^2 \leq \|\mathbf{x}_i - \mu_{j'}\|^2, \forall j' \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

由于 γ 是一个 $n \times k$ 的矩阵，它代表的是第 i 个向量与第 j 个簇的关系，当它属于第 j 个簇时，$\gamma_{ij} = 1$。

因此首先非常容易可以证明，当簇中心 μ_j 比簇中心 μ'_j 到 x_i 的距离还要近的使用，令 $\gamma_{ij} = 1, \gamma_{ij'} = 0$，此时计算的 x_i 到簇中心 μ_j 的距离最短，目标函数 $\mathcal{J}$ 的值减小。

同样非常容易证明当 x_i 选择离他最近的簇中心的时候，每个 $\|x_i - \mu_j\|^2$ 此时取到最小值，因此目标函数 $\mathcal{J}$ 的值会下降。

Step2:

$$\mu_j = \frac{\sum_{i=1}^{n} \gamma_{ij} \mathbf{x}_i}{\sum_{i=1}^{n} \gamma_{ij}} \tag{3}$$

这个式子的意思是选取每个属于类 j 的 x_i，将他们求和后除于类 j 里面的向量的个数，得到新的簇中心。

因此我们需要证明当簇中心变换之后，目标函数的值会下降。即证明：设 $x_1, x_2, \cdots, x_n$ 是欧式空间的 n 个向量，则 $\sum_{i=1}^{n}\sum_{j=1}^{k} \gamma_{ij}\|x_i - \mu_j\|^2$ 取到最小值时当且仅当 μ_j 是这 n 个 $\gamma_{ij} = 1$ 的向量的中心位置，即 $\mu_j = \frac{\sum_{i=1}^{n} \gamma_{ij} \mathbf{x}_i}{\sum_{i=1}^{n} \gamma_{ij}}$.

$$\frac{\partial \sum_{i=1}^{n}\sum_{j=1}^{k} \gamma_{ij}\|x_i - \mu_j\|^2}{\partial \mu_j} = -2\left(\frac{\sum_{i=1}^{n} \gamma_{ij} \mathbf{x}_i}{\sum_{i=1}^{n} \gamma_{ij}} - \mu_j\right) = 0$$

因此可得：$\mu_j = \frac{\sum_{i=1}^{n} \gamma_{ij} \mathbf{x}_i}{\sum_{i=1}^{n} \gamma_{ij}}$ 是该函数的一个驻点，又因为该目标函数显然是严格凸的，因此 $\mu_j = \frac{\sum_{i=1}^{n} \gamma_{ij} \mathbf{x}_i}{\sum_{i=1}^{n} \gamma_{ij}}$ 是目标函数的唯一最小值点，因此只要 Step2 中有某个中心位置发生了变化，那么目标函数的值就会减小。

1.1.2 试证明，算法 1 会在有限步内停止

我们将目标函数记为 J(T), 其中 T 是对于数据集的一种划分方式，例如划分 T_1 是将数据集划分成 $C_1, C_2, \cdots, C_k$ 这 k 个互不相交的集合，因此：

$$J(T) = \sum_{i=1}^{n}\sum_{j=1}^{k} \gamma_{ij} \left\|\mathbf{x}_i - \mu_j\right\|^2$$

显然对于任何一种划分方式，我们可以知道：$J(T) \geqslant 0$, 因此对于任意选择的初始点开始不断进行迭代，我们可以得到对于数据集的不同的划分：$T_1, T_2, T_3 \cdots$，并且对应的目标函数值为 $J(T_1), J(T_2), J(T_3), \cdots$，在第一问中我们已经证明了 $J(T)$ 一定是单调递减的，并且我们可以做个极端的猜想，当 n 个向量被分成 n 个簇时，他们的簇中心就是他们自身，此时目标函数应该可以取到最小值 0，因此我们可以得到目标函数 $J(T)$ 有下界 0.

因此由单调有界数列的收敛定理可得：

$$\lim_{n\to\infty} J(T)\text{存在}$$

因此，算法 1 一定会收敛，在有限步内停止。

2 附加题

2.1 在公式 1 中, 我们使用 ℓ_2 - 范数来度量距离 (即欧式距离), 下面我们考虑使用 ℓ_1-范数来度量距离

$$J'(\gamma, \mu_1, \ldots, \mu_k) = \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} \left\| \mathbf{x}_i - \mu_j \right\|_1 \tag{4}$$

2.1.1 请仿效算法 1 , 给出新的算法 (命名为 $k-$ means- ℓ_1 算法) 以优化公式 4 中的目标函数 J'.

Algorithm 1: $k-$ means- ℓ_1

初始化 $\mu_1, \mu_2, \cdots, \mu_k$

while 目标函数 J 改变 **do**

Step1：计算各个点到簇中心的距离，将每个点划分到离它最近的簇中心，形成 K 个簇，

$$\gamma_{ij} = \begin{cases} 1, & \left\| \mathbf{x}_i - \mu_j \right\|_1 \leq \left\| \mathbf{x}_i - \mu_{j'} \right\|_1, \forall j' \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Step2: 用更新后的指示矩阵 γ 重新计算 k 个簇的簇中心 μ_j

$$\min \sum_{k=1}^{n} \sum_{i=1}^{n} \gamma_{ij} \| x_i - \gamma_{kj} x_k \|_1 \tag{6}$$

$$\mu_j = \text{使上式最小的} x_i$$

也就是说计算簇内所有样本点到其中一个样本点的曼哈顿距离和，并选出使簇的曼哈顿距离和最小的样本点作为质心

end

2.1.2 当样本集中存在少量异常点 (outliers) 时, 上述的 $k-$ means- ℓ_2 和 $k-$ means- ℓ_1 算法, 我们应该采用哪种算法？即哪个算法具有更好的鲁棒性？请说明理由。

应该采用算法 $k-$ means- ℓ_1。

从算法的运行步骤我们可以知道一件事情：$k-$ means- ℓ_2 算法的簇中心是各个样本点的平均，可能是样本点中不存在的点，而 $k-$ means- ℓ_1 的簇中心一定是某个样本点的值。

我们可以举一个例子来描述这件事情：

当一个簇的样本集只有少数几个点，如 (1,1),(2,2),(100,100)。我们可以知道 (100,100) 是异常点。如果按照 $k-$ means- ℓ_2 算法执行的话簇中心大致会处在 (1,1),(100,100) 中间，大概是在 (50,50) 左右的位置，这其实是一个非常偏离其他样本点的簇中心，这显然不是我们想要的。

这时 $k-$ means- ℓ_1 会在 (1,1),(2,2),(100,100) 中选出一个样本点使簇的 ℓ_1 范数和最小，因此此时一定会在前两个点中选取。

从这个例子我们可以看出算法 $k-$ means- ℓ_1 在遇到异常点也就是噪声的时候有较好的鲁棒性。