k-means

Mia Feng

2018 年 3 月 4 日

1 概述

k-means：无监督，确定性模型。用于聚类，即将没有类标的数据集分组为一些异质性的子群（cohesive "clusters"）[1]。无监督问题的难点在于没有正确的类标给与分类提示，所以算法对初始值敏感。

求解目标：聚类，每个样本点被标以类标。

$$c^{(i)} = \left\{ j | j \in \left[1, 2, \cdots, k \right] \right\} \tag{1}$$

求解思路：最小化误差平方和, 取距离样本点最近的类标作为样本的 label。Concretely, 通过迭代寻找 k 个聚类，使得这 k 个聚类的均值所代表相应各类样本时所得的总体误差最小。记 c 为类标，μ 为 cluster 中心，代价函数为 [2]：

$$\boldsymbol{J}\left(c, \mu\right) = \sum_{i=1}^{k} \left\| x^{(i)} - \mu_{c^{(i)}} \right\|^2 \tag{2}$$

求解方法：误差平方和最小化。

1.1 推导

聚类中心 cluster centroids，是 cluster 内样本点的均值。有 k 个，在初始化时候指定。更新时，聚类中心是求取当前 cluster 的均值。对于类 j，其类中心 μ_j（对于聚类中心的当前猜值）为

$$\mu_j := \frac{\sum\limits_{i=1}^{m} \mathbb{1}\left\{ c^{(i)} = j \right\} x^{(i)}}{\sum\limits_{i=1}^{m} \mathbb{1}\left\{ c^{(i)} = j \right\}} \tag{3}$$

类标

$$c^{(i)} := \arg\min_j \left\|x^{(i)} - \mu_j\right\|^2 \tag{4}$$

2 算法实现

见 CS229[1]

1. 随机初始化 cluster centroids $\mu_1, \mu_2, \cdots, \mu_k \in \mathbb{R}^n$
2. 迭代直至收敛 {

   对于每一个样例 i，计算类标

   $$c^{(i)} := \arg\min_j \left\|x^{(i)} - \mu_j\right\|^2 \tag{5}$$

   对于每一个类 j，更新 cluster centroids：

   $$\mu_j := \frac{\sum_{i=1}^{m} \mathbb{1}\left\{c^{(i)} = j\right\} x^{(i)}}{\sum_{i=1}^{m} \mathbb{1}\left\{c^{(i)} = j\right\}} \tag{6}$$

   }

3 Implementation

聚类测试：数据在 data.csv

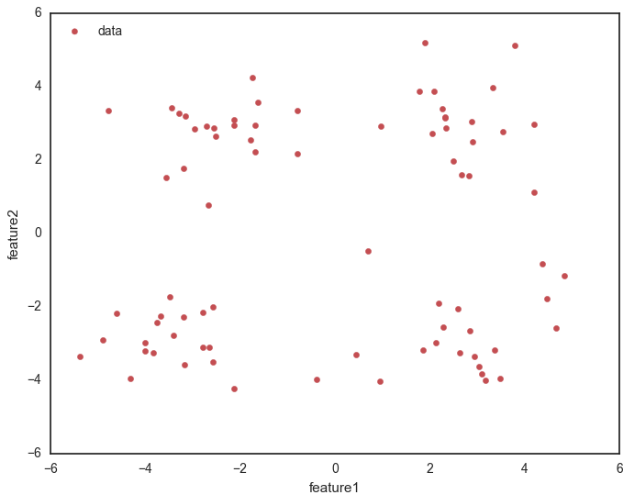

图 1: 训练数据

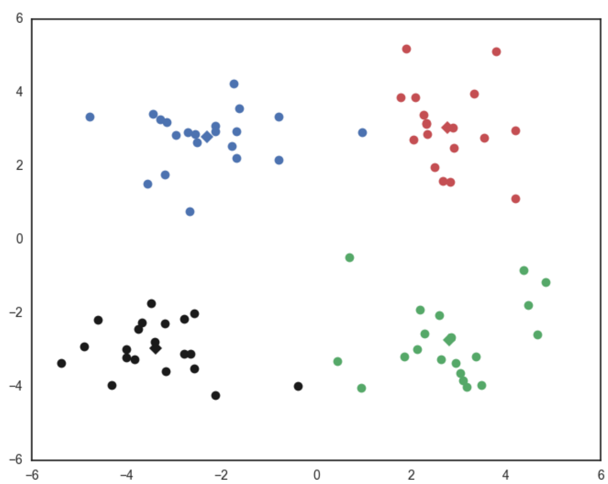

图 2: kmeans 运行结果，iter=1，k=4。菱形标记聚类中心，点标记数据

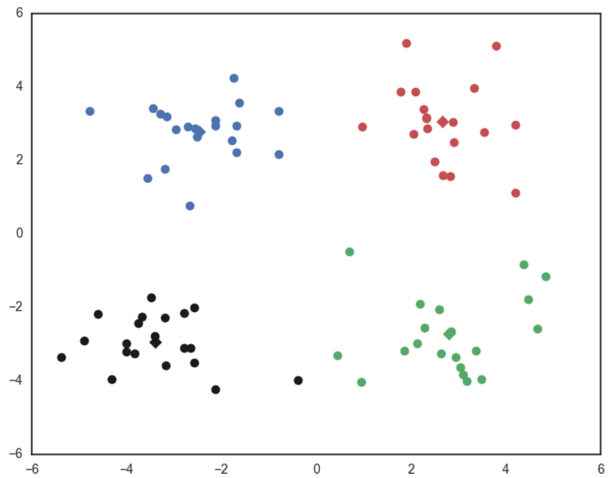

图 3: kmeans 运行结果，iter=2，k=4。菱形标记聚类中心，点标记数据

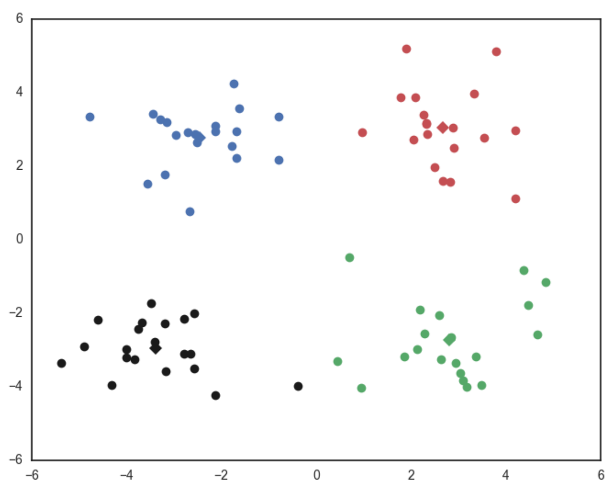

图 4: kmeans 运行结果，iter=3，k=4。菱形标记聚类中心，点标记数据

参考文献

[1] Andrew Ng. cs229-notes7a. http://cs229.stanford.edu/notes/cs229-notes7a.pdf, 2012. Stanford-CS229, 2012.

[2] zouxy09. 机器学习算法与 python 实践之（五）k 均值聚类（k-means）. http://blog.csdn.net/zouxy09/article/details/17589329, 2013. Blog, 2013.