扫码签到

无监督聚类
K-means

目录

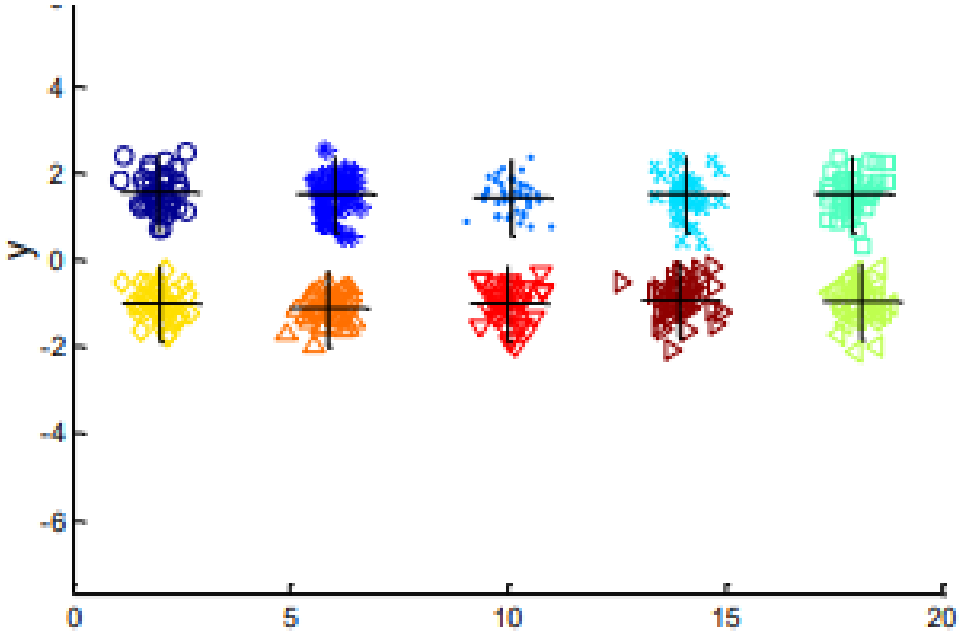
- 1 无监督学习
- 2 K-means回顾
- 3 实验任务

无监督学习

- 在实际场景中，我们获得的数据可能是没有标签信息的
- 对于上节课讲到的KNN，如果训练集中没有标签信息的话，我们也就无法对测试集中的数据进行分类
- 但是我们又想从这些无标签的信息中获取信息，比如说哪些样本更有可能是属于同一类的
- 因此，如何在没有标签信息时提取这些类别信息是一个需要研究的问题

K-means基本思想

- 将数据集划分成K个不相交的块
- 每个块都由一个中心点(centroid)联系起来
- 数据集中每个点都被划分到唯一的最接近的中心点所在的块中



K-means实现细节

- 首先随机初始化K个中心点
- 对于数据集中的N个点，计算其到这K个中心点的距离(度量使用SSE)
- 样本点被划分到距离最近的中心点所在的块
- 数据集中所有样本均划分好块后，对中心点进行更新(一般更新为块中样本点的均值)
- 判断是否收敛

– To get SSE, we square these errors and sum them.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

– x is a data point in cluster C_i and m_i is the representative point for cluster C_i

Ref: https://www.bing.com/th/id/OGC.35df417142a3ebba68aae54d74f998e0?pid=1.7&rurl=https%3a%2f%2fimg-blog.csdnimg.cn%2f20201124124344811.gif%23pic_center&ehk=wz%2b1oVWYIwRF1k4qqvkwSptWPRK8%2fMRMAybY%2bLxGI08%3d

K-means的优缺点

- 优点：
  - 原理简单，实现简单
  - 收敛速度块
  - 当结果块比较密集，且块与块之间区别明显时，效果较好
- 缺点：
  - K值需要事先给定
  - 对中心点选择敏感

K-means++

- 由于K-means对初始质心选择敏感，有人提出了K-means++这一优化算法
- 主要思想：是的初始聚类中心之间的相互距离尽可能远
- 实现过程：
  - 首先随机选择一个数据集中的点作为初始化的聚类中心
  - 对于数据集中的其他样本点x_i，计算其与聚类中心的距离，并记录最短距离d_i
  - 按概率采样一个样本点，距离越大，被采样的概率也就越大
  - 重复上述过程，直到k个聚类中心都被确定
  - 利用K-means进行聚类

$$P(x) = \frac{D(x)^2}{\sum_{x \in X} D(x)^2}$$

注：I表示当前中心点总数，它从1慢慢增长到K

$$D(x) = \min_{i=1,\ldots,I} dist(x, m_i)$$

$$where\ m_k\ is\ the\ representative\ point\ of\ cluster\ C_k$$

实验任务

- 在给定文本数据集完成文本情感聚类训练
- 要求
  - 文本的特征可以使用TF或TF-IDF （可以使用sklearn库提取特征）
  - 利用K-means，K-means++对文本特征进行聚类
  - 计算calinski_harabasz_score（越高越好）

    可调用sklearn.metrics.calinski_harabasz_score函数计算
  - 需提交实验报告+代码
    - 实验报告应包含样本的可视化展示，可以考虑利用TSNE将文本特征投影到直角坐标系中进行展示（可以调用sklearn库的TSNE投影，可视化工具不限）

参考

- TSNE: sklearn.manifold.TSNE — scikit-learn 1.0.2 documentation
- Matplotlib可视化教程: https://www.runoob.com/matplotlib/matplotlib-tutorial.html