
Lab 4: K-means

1 Intro

K-means 聚类算法是一类重要的无监督学习方法。 k 均值聚类算法 (K-means clustering algorithm) 是一种迭代求解的聚类分析算法, 其步骤是, 预将数据分为 K 组, 随机选取 K 个对象作为初始的聚类中心, 然后计算每个对象与各个种子聚类中心之间的距离, 把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。每分配一次样本, 聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有 (或最小数目) 对象被重新分配给不同的聚类, 没有 (或最小数目) 聚类中心再发生变化, 误差平方和局部最小。

Pseudo-code

```
1 1. Choose the number of clusters(K) and obtain the data points
2 2. Place the centroids  $c_1, c_2, \dots, c_k$  randomly
3 3. Repeat steps 4 and 5 until convergence or until the end of a fixed number of
   iterations
4 4. for each data point  $x_i$ :
5     find the nearest centroid( $c_1, c_2 \dots c_k$ )
6     assign the point to that cluster
7 5. for each cluster  $j = 1..k$ 
8     new centroid = mean of all points assigned to that cluster
9 6. End
```

2 Task

TODO

- **coding:**实现**K-means**聚类算法, 算法输入 n 个数据、分类数 k , 输出 n 个数据的聚类类别、 k 个类别中心
- **evaluation:**可视化聚类结果, 作为实验报告

Note

- 聚类实验数据及预处理: 不作限制, 可以自行收集/生成文本、图像、离散坐标等数据, 可以自行选择是否需要数据预处理 (提取特征、降维); 一些资源供参考:
 - 图像数据集: AI Cat and Dog Images by DALL·E Mini, 54 cat images + 54 dog images, 由DALL·E Mini生成
 - 图像特征提取: 文澜API, 可以上传图片文件, 返回这张图片的文澜编码, 详见API Readme, P.S. 任意两个文澜编码间相似度/距离度量采用的是余弦相似度 (两个编码向量内积)
 - 坐标数据生成: $\text{sigma} * \text{np.random.randn}(n, 2) + \text{mu}$, 可以生成 n 个来自均值为 mu 、方差为 sigma 的二维高斯分布的采样, 可以随机生成 k 簇这样的采样, 来测试类别数为 k 的K-means

自行选取实验数据时尽量保证数据有明显的类别信息

- 可视化: 处理数据是二维/三维可以直接展示坐标点、不同类别采取不同颜色; 处理数据是高维可以采用降维到二维/三维的方式可视化 (降维工具: `sklearn.decomposition.PCA`等)

Submit

- 2021xxxxxx_xiaoming_lab4.zip (./code ./report.pdf)
- Email xihuaw@ruc.edu.cn, DDL 2022.10.21 20:00