Lab 3: 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. Choose the number of clusters(K) and obtain the data points
2. Place the centroids c_1, c_2, ..... c_k randomly
3. Repeat steps 4 and 5 until convergence or until the end of a fixed number of
    iterations
4. for each data point x_i:
       find the nearest centroid(c_1, c_2 .. c_k)
       assign the point to that cluster
5. for each cluster j = 1..k
       new centroid = mean of all points assigned to that cluster
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
  - 特征提取：①图像或文本的特征可以使用文澜API，指定文本或者图片文件，返回这张图片的文澜编码，详见API README[1]；②图像也可以使用ResNet或者CLIP等模型提取图片特征编码；③特征也可以参考HuggingFace Space中例如 this的demo（可以自行发掘更多），通过底部的「Use via API」调用；
    P.S. 任意两个编码间相似度/距离度量可以用余弦相似度（两个向量内积）衡量
  - 坐标数据生成：sigma * np.random.randn(n, 2) + 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

- 提交一个zip文件，严格注意命名，形如：2023101000+张三+实验3.zip
- 除非必要，zip中应当仅包含一个报告pdf和一个代码文件，命名不作要求
- `https://k.ruc.edu.cn` DDL 2024.10.24 23:59

[1] 需替换原URL为：url = http://bl.mmd.ac.cn:8889/{text,image}_query，该服务支持到10.24 23:59结束