本科生《机器学习》第三次实验作业

聚类分析

截止时间：2024.12.08 23：59

1 作业一：K-均值聚类

1.1 手动实现 K-均值聚类

（1） 总体要求

使用 K-means 算法手动实现聚类，不要直接调用库函数；在任意数据集上评估聚类结果；比较不同参数（如簇的个数、初始聚类中心、迭代次数等）对模型性能的影响。

（2） 实验说明

a. 自选一个真实数据集（dataset/01-True）或两个合成数据集（dataset/02-Synthetic）；

b. 手动实现 K-means 聚类算法；

c. 若采用真实数据集，至少计算一个聚类评估指标，并比较不同参数对模型性能的影响；若采用合成数据集，分析两个合成数据集的数据分布特性，可视化比较不同参数对模型性能的影响。

d. Notebook 中需要包含关键代码注释、结果分析等内容。

1.2 （选做）图像分割

（1） 总体要求

自选图片，使用 K-means 算法进行基于颜色特征的图像分割实验，观察颜色空间（RGB 和 HSV）和不同聚类数对分割效果的影响。可调用手动实现的 K-means 函数，也可使用 sklearn 等库中封装好的函数。

（2） 实验说明

a. 自选图片：尽量选择颜色对比鲜明、主体清晰的照片；将图片上传至（dataset/03-MyData）

b. 读取图片：导入 PIL.Image 库，使用

Image.open(image_path).convert('RGB')/
Image.open(image_path).convert('HSV')按指定色彩空间读取照片；

c. 将图像展平后（2 维*3 通道->1 维*3 通道），使用 K-means 算法聚类；

d. 可视化像素点聚类结果，即图像分割结果；

e. 观察颜色空间（RGB 和 HSV）和不同聚类数对分割效果的影响。

2 作业二：手动实现高斯混合聚类

（1） 总体要求

使用 GMM 算法手动实现聚类，不要直接调用库函数；在任意数据集上评估聚类结果；比较不同参数对模型性能的影响。

（2） 实验说明

a. 自选一个真实数据集（dataset/01-True）或两个合成数据集（dataset/02-Synthetic）；

b. 手动实现 GMM 聚类算法；

c. 若采用真实数据集，至少计算一个聚类评估指标，并比较不同参数对模型性能的影响；若采用合成数据集，分析两个合成数据集的数据分布特性，可视化比较不同参数对模型性能的影响。

d. Notebook 中需要包含关键代码注释、结果分析等内容。

3 作业三：密度聚类：实现 DBSCAN 聚类算法

（1） 总体要求

使用 DBSCAN 算法手动实现聚类，不要直接调用库函数；在合成数据集上评估聚类结果；比较不同参数对模型性能的影响。

（2） 实验说明

a. 自选至少一个合成数据集（dataset/02- Synthetic）；

b. 手动实现 DBSCAN 聚类算法；

c. 分析合成数据集的数据分布特性，可视化比较 ϵ-邻域大小及聚类数对模型性能的影响。

d. Notebook 中需要包含关键代码注释、结果分析等内容。

4 作业四：层次聚类

（3） 总体要求

dataset/01-True/country.csv 是摘自《世界竞争力报告--1997》关于 20 个国家和地区的信息基础设施发展状况数据，根据该数据对这些国家和地区进行分层聚类分析，比较不同距离定义下的聚类结果。

（4） 实验说明

a. 数据集说明：

I.Call—每千人拥有电话线数

II.movecall—每千房居民蜂窝移动电话数

III.fee—高峰时期每三分钟国际电话的成本

IV.Computer—每千人拥有的计算机数

V.mips—每千人中计算机功率《每秒百万指令》

VI.net—每千人互联网络户主数

结合实际，可以选择把这 20 个国家分为两类的结果，其中巴西、智利、墨西哥、俄罗斯、波兰、匈牙利、马来西亚、泰国、印度为一类，它们基本上都是当时的转型国家和亚洲、拉美的发展中国家，属于信息基础设施比较落后的国家；而其它 11 个国家和地区为一类，包括美、日、欧洲发达国家和新兴工业化国家和地区。

b. 手动实现分层聚类；

c. 比较不同距离定义下的聚类结果。

d. Notebook 中需要包含关键代码注释、结果分析等内容。

5 （选做）模型比较

在同一个数据集上，比较不同聚类方法的性能，通过可视化、评价指标，文字叙述等方式呈现。