课程：人工智能 **作业 4**

北航软件学院

学期：2023，春季 提交截止时间：2022年5月15日，11：59 PM

提醒注意：

- 本次作业发布于2023年4月25日，截止于2023年5月15日。
- 作业一分为三部分：问答题、实训题、以及实训题报告
  - 问答题答案可以手写并扫描，或者用latex（或word）手打，最终以QA.pdf文件命名。
  - 实训题按照项目共享链接内要求和基础代码进行作答。
  - 报告部分同样可以手写或者手打，以Report.pdf文件命名。
  - 作业提交格式：$<studentID>_<name>$_A4.zip。比如ZY1921102_田嘉怡_A4.zip
  - 提交的zip文件要求（仅）包括：
    - 实训题文件：包括 main.ipynb, pred.npy, 以及work 目录下的 model.pkl, pca.pkl, scaler.pkl
    - 问答题答案：QA.pdf
    - 报告：Report.pdf
- 作业压缩包需要在spoc平台上提交。
- 每迟交1天（不满1天按1天计算），本次作业扣除10%分数。
- 不按作业要求和格式提交，视情况扣分。不得抄袭。

第一部分：问答题

Q 1
图 1 给出了K-均值聚类算法的初始状态和后续迭代结果。在图1左图给出的初始状态中，圆形代表被聚类数据点，方形代表聚类质心，给出后续迭代中聚类之心在对应图中的结果。

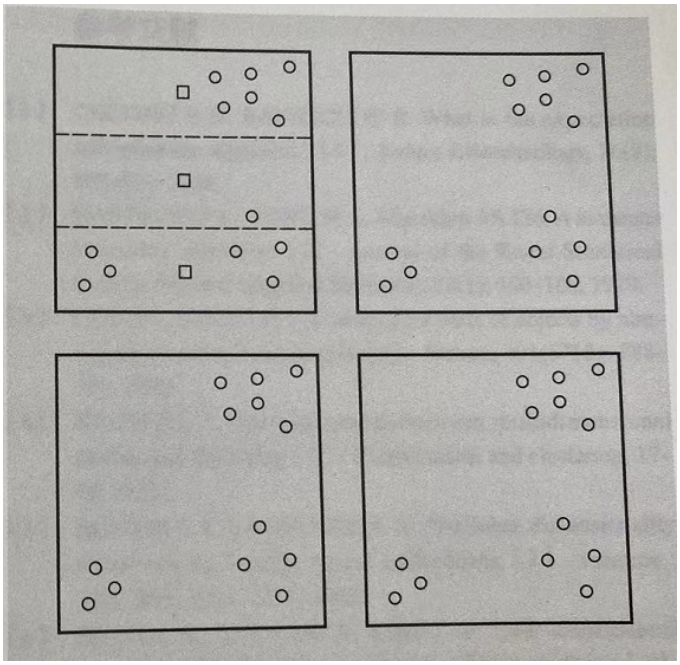

图 1: K-均值聚类结果初始状态及后续迭代聚类结果示意图

Q 2
说明在K-均值聚类算法执行过程中，其目标函数 $\sum_{i=1}^{K}\sum_{x\in G_i}||x-c_i||^2$ 是严格递减的，并解释为什么K-均值聚类算法可以确保在有限步内收敛。

Q 3 特征值分解
令 X 表示一个 $m\times n$ 的矩阵, 其奇异值分解为

$$X = U\Sigma V^T,$$

其中 $\Sigma = diag(\sigma_1, \sigma_2, ..., \sigma_{min(m,n)})^T$ 由 X 的奇异值所组成.

(a) XX^T 的特征值和特征向量是什么?

(b) X^TX 的特征值和特征向量是什么?

(c) XX^T 与X^TX各自的特征值之间存在什么样的联系?

(d) X的奇异值与XX^T (X^TX)的特征值之间存在什么样的联系?

(e) 如果 $m = 2$、$n = 100000$, 你会如何计算 X^TX 的特征值?

Q 4 主成分分析
主成分分析是一种典型的无监督线性特征降维（特征提取）方法。其目标函数有两种形式，分别对两种目标函数进行优化，并对优化结果进行对比分析。
假设原始数据组成的矩阵为$X = [x_1, ..., x_i, ..., x_N] \in \mathbb{R}^{N\times D}$， 其中$x_i \in \mathbb{R}^D$ 为一个D维向量的数据样本。经过线性变换矩阵$W \in \mathbb{R}^{D\times d}$后，得到降维后的特征为$Z = [z_1, ..., z_i, ..., z_N] \in \mathbb{R}^{N\times d}$，其中$z_i \in \mathbb{R}^d$，且$d < D$。

(a) 当考虑第一个投影方向w时，降维后数据为$z_i = x_i^Tw$，其方差为

$$Var(x^Tw) = \frac{1}{N}\sum_{i=1}^{N}(x_i^Tw - \bar{x}^Tw)^2$$

其中$\bar{x} = \frac{1}{N}\sum_{i=1}^{N}x_i$。 因此，最大化方差目标函数定义为

$$\max_w \frac{1}{N}\sum_{i=1}^{N}(x_i^Tw - \bar{x}^Tw)^2, \quad \text{subject to} \quad \|w\| = 1$$

请根据此目标函数计算出优化后的w，并给出推导过程。

(b) 从最小化重构误差角度出发，其目标函数为

$$\min_{w,z} \frac{1}{N}\sum_{i=1}^{N}\|x_i - (\bar{x} + z_iw)\|^2$$

其中$z = (z_1, z_2, ..., z_N)^T$为投影到第一个投影方向后的数据。请根据此目标函数计算出优化后的w，并给出推导过程。

(c) 对比两种目标函数优化结果并讨论约束条件$\|w\| = 1$存在的必要性。

第二部分：实训题（共7分）

实训题要求:

- 本次作业包括1个实训题，作业要求以及基础代码以Aistudio项目的形式发布。
- 发布项目链接有效期3天，请在作业发布3天内fork这个项目，生成“我的项目”，并在自己fork的项目下进行作答，生成答案后按要求保存提交。

Q 1 聚类问题-KMeans 实现异常点检测

异常值检测（outlier detection ）是一种数据挖掘过程，用于发现数据集中的异常值并确定异常值的详细信息。 当前数据容量大、数据类型多样、获取数据速度快；但是数据也比较复杂，数据的质量有待商榷；而数据容量大意味着手动标记异常值成本高、效率低下；因此能够自动检测异常值至关重要。 自动异常检测具有广泛的应用，例如信用卡欺诈检测、系统健康监测、故障检测以及传感器网络中的事件检测系统等。

本实验要求使用Pandas, Numpy, Sklearn 等库进行相关特征处理，自行编写KMeans算法，完成异常点检测。

实验介绍详情和参考基础代码请参见Aistudio中的共享项目“人工智能23-作业四-KMeans实现异常点检测”。

第三部分：实训题实验报告（共3分）

- 请按照实验报告模板完成实验报告。
- 实验报告模板是通用模板，可根据每个作业要求的差别，自由进行微调。