KDD Lab 作业 3

1. 作业上交时间：

4/16 14:00~ 4/30 22:00 通过 blackboard 上传作业文件，本次作业不支持补交或缓交。逾期不收！逾期不收！逾期不收！

2. 作业题目：

第三次作业共有 **2 个小题**，主要是让同学们学会熟练的使用 python 对给定的无标签数据集寻找合适的聚类算法，进行聚类。

题目 1： 给定数据集 HW3_1_data.csv，要求使用两种不同的聚类算法对给出的二维数据进行聚类，聚类算法的重要参数需要学生自行设计选取，且需注释说明参数选择的过程和目的，代码总运行时间 < 5min。

题目 2： 使用的无标签数据为 HW3_2_data.csv，该数据描述了顾客的信用卡消费记录，要求基于该数据，使用 1 种聚类算法将顾客进行聚类，数据描述如下：

- CUSTID: Identification of Credit Card holder (Categorical)
- BALANCE: Balance amount left in their account to make purchases
- BALANCEFREQUENCY: How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)
- PURCHASES: Amount of purchases made from account
- ONEOFFPURCHASES: Maximum purchase amount done in one-go
- INSTALLMENTSPURCHASES: Amount of purchase done in installment
- CASHADVANCE: Cash in advance given by the user
- PURCHASESFREQUENCY: How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)
- ONEOFFPURCHASESFREQUENCY: How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)
- PURCHASESINSTALLMENTSFREQUENCY: How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)
- CASHADVANCEFREQUENCY: How frequently the cash in advance being paid
- CASHADVANCETRX: Number of Transactions made with "Cash in Advanced"

- PURCHASESTRX: Numbe of purchase transactions made
- CREDITLIMIT: Limit of Credit Card for user
- PAYMENTS: Amount of Payment done by user
- MINIMUM_PAYMENTS: Minimum amount of payments made by user
- PRCFULLPAYMENT: Percent of full payment paid by user
- TENURE: Tenure of credit card service for user

要求学生将数据预处理和聚类算法的参数选择过程都进行注释，方便作业批改。代码运行时间 < 5min。

3. 作业要求：

按照上述 2 小题要求完成代码，在作业的上交期限之前，上交 ipynb 文件至 blackboard，作业命名规则：HW3_学生学号.ipynb

备注：

（1）为方便作业批改，尽量提交 ipynb 格式文件，提交前先清除所有输出，减少 ipynb 文件的大小

（2）为方便作业批改，请读取在同一个目录下的 HW3_1_data.csv 和 HW3_2_data.csv（比如在 read csv 时使用相对路径）

（3）请勿写入源文件 HW3_1_data.csv 和 HW3_2_data.csv ！！

4. 作业计分：

题目 1：6 分（两个聚类算法各 3 分）； 题目 2：4 分；总分 10 分扣完即止。

（1）内容扣分点：

题目 1：

聚类算法 1 无法实现： -2 分

聚类算法 2 无法实现： -2 分

没有聚类算法 1 的参数选取过程或者选取过程存在逻辑错误：-1 分

聚类算法 1 没有注释参数的选择过程：-1 分

没有聚类算法 2 的参数选取过程或者选取过程存在逻辑错误：-1 分

聚类算法 2 没有注释参数的选择过程：-1 分

没有打印最终模型的评价指标： -1 分

运行时间不满足要求： -1 分

满分 6 分，扣完即止。

题目 2:

无法实现聚类算法：-2 分

聚类算法运行时间不满足要求：-1 分

没有数据预处理： -1 分

没有将数据预处理过程进行注释：-1 分

没有算法参数选取过程或者选取过程存在逻辑错误：-1 分

没有将参数选取过程进行注释：-1 分

没有打印模型的评价指标: -1 分

满分 4 分，扣完即止。

(2) 其他扣分点:

迟交：-10 分

不按照命名规则命名：-1 分

完全不注意备注信息：-1 分

题目 1 程序无法执行：-2 分

题目 2 程序无法执行：-2 分