## **广州大学学生实验报告**

**开课学院及实验室：**计算机科学与工程实验室 **2019年12月18日**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **学院** | **计算机科学与网络工程学院** | **年级/专业/班** | **软件181** | **姓名** | 连梓煜 | **学号** | 1806300105 |
| **实验课程名称** | **人工智能导论实验** | | | | | **成绩** |  |
| **实验项目名称** |  | | | | | **指导老师** |  |

**一、实验内容**

1. 读取数据，并做预处理。
2. 至少实现两种聚类算法，选择评价方法比较结果并分析原因
3. 选择适当可视化方法显示结果。
4. \*扩展选做题：分析考虑数据的特性并和具体聚类方法的匹配。

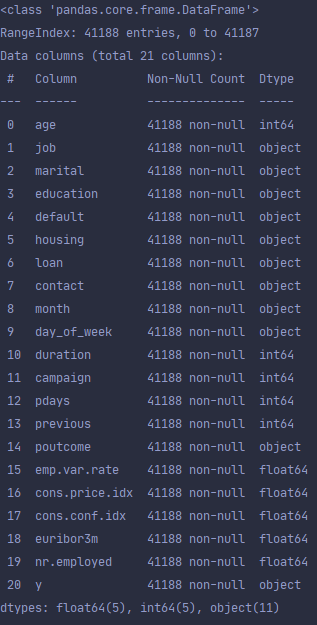
**二、实验设备**

1. 实验设备：计算机
2. 平台：Windows操作系统，Python Anaconda

**三、实验步骤**

1. 读取数据，并做预处理。
   1. 观察数据

*import* pandas *as* pd  
  
*if* \_\_name\_\_ == '\_\_main\_\_':  
 data\_path = "../data/bank-additional-full.csv"  
 data = pd.read\_csv(data\_path, sep=";")  
 *print*(data.info())



观察数据可以看到，总共有21列，其中数值类型有10项，字符串类型有11项。

根据数据描述文档可以知道，字符串类型是固定的几个类型，因此，可以对字符串类型转为固定的数值类型，方便进行机器学习的输入。

查看unknown的数量

*for* i *in* data.columns:  
 *if type*(data[i][0]) *is str*:  
 *print*("未知的列为：" + i + ": " + *str*(data[data[i] == "unknown"]["y"].count()))



观察数据可以看到，字符串类型的11项中，其中job未知为330，marital未知为80，education未知为1731，default未知为8597，housing未知为990，loan未知为990

对于数量较小的为未知的数据项，可以选择删除，在这里，选择删除martial,job为空的项

* 1. Df

1. 至少实现两种聚类算法，选择评价方法比较结果并分析原因
2. 选择适当可视化方法显示结果。
3. \*扩展选做题：分析考虑数据的特性并和具体聚类方法的匹配。

**四、分析说明（包括核心代码及解释）**

**五、总结心得**