CSDM-NJUSME

南京大学工程管理学院2020年春

研究生数据挖掘课程作业1

一、作业说明

以Credit.xls为数据，用R或者python完成数据的预处理工作，包括

1. 缺失值和异常值处理；
2. 分类属性预处理，使得分类属性取值数目不超过5个；
3. 数值属性预处理，对数值属性进行离散化，离散化后取值数目不超过5个；

用Python 3实现，IDE是PyCharm，因为第一次作业涉及的数据清洗过程需要不断查看数据，所以使用PyCharm内嵌的Jupyter Notebook，方便查看处理效果。

二、文件说明：

1. 文件结构

```
$ tree
.
├── README.md（说明文档）
├── README.pdf（说明文档的pdf，方便查看）
├── data
│   ├── Credit.csv（原数据）
│   └── preprocess.csv（清洗后的数据）
├── preprocess
│   └── preprocess.ipynb（预处理的源代码）
└── preprocess.pdf（预处理的源代码的pdf版本，方便查看）
```

2. 清洗过程

   `preprocess.ipynb`：数据清洗有很多原则和方法，效果只能根据之后的模型来检验。所以作业里是根据老师ppt介绍的适合信用评分的这类数据的方法进行的数据清洗，包括odds值合并取值、k-means给取值聚类等。下面介绍具体实现过程，但建议查看`preprocess.pdf`，结合代码运行结果来看。

- 1 分类变量预处理 首先查看分类型变量的描述性统计信息，有多少种取值、具体是什么值取值，每个取值占比多少，最大值、最小值、众数等。
  - 1.1 分类变量的缺失值和异常值处理
    - 1.1.1 分类变量的缺失值处理

4个分类属性，且只有MARITAL_STATUS中有缺失值Unknown，占比10.269%，我们不删除该属性，将Unknown单独作为一类。

- 1.1.2 分类变量的异常值处理

  无异常值

- 1.2 分类变量取值个数处理 目的：使得每个分类属性的取值都不超过5个
  - 一是检查删除
    - 所有分类属性不存在某个取值占了90%以上，所以没因此删除属性
    - 所有分类属性不存在缺失值或者异常值占了50%以上，所以没因此删除属性
    - 分类属性只有2种取值，且这两种取值的odds又相等，则对违约分类无区分作用，所以要删除这种属性。下面来检查这种情况，性别、贷款类型、还款类型都是只有2种取值，经过检查，发现两个取值的odds都不相等，所以不用删除。
  - 二是检查合并 分类属性中某个取值占比小于5%时，要与odds相近的取值合并成一类，直至每个取值占比大于5%。
    - 数据集里只有婚姻状态的Divorce取值低于5%，所以接下来找出Married、Single、Unknown这三个取值中与Divorce的odds最近的那个。下面的结果表示Unknown与Divorce的odds最接近，所以将二者并成一类，新类Unknown_Divorce的占比超过了5%，故停止合并。
    - 继续检查合并，若属性中取值个数大于2，且其中某两个取值的odds相等，则需要将相等odds的取值归为一类。 我们这里就只有属性婚姻状态取值个数大于2，那么就检查它的三个取值的odds有没有相等的情况，有就进行合并。从下面的输出结果可看出没有取值的odds相等，所以不合并。
    - 最后，把所有字符串取值都变成数字形式方便输入模型

- 2 数值变量预处理
  - 2.1 数值变量的缺失值和异常值处理
    - 2.1.1 数值变量的缺失值处理

      数据集中有6个数值变量，其中只有MONTHLY_INCOME_WHITHOUT_TAX统计出的数目比总样本个数少，说明有缺失值，我们进行查看，发现缺失值占0.7%，于是我们决定将缺失值单独归类为-1，并最终要将该属性进行离散化。

    - 2.1.2 数值变量的异常值处理 异常的定义需要结合专业知识和常识
      - 我们觉得每个客户的GAGE_TOTLE_PRICE都应该大于APPLY_AMOUNT，如果不是，说明其中登记有误。 下面进行检查，发现所有客户都符合抵押价值大于贷款值，故认为GAGE_TOTLE_PRICE和APPLY_AMOUNT都无误。
      - 年龄和贷款时间长度、利率也都正常，最大值和最小值都符合常识。
      - 税前月收入，除了1.2.1中已经提到有缺失值，我们还发现最大值上亿，最小值为0。最小值0我们暂时理解为收集数据时该客户没有月收入，但是最大值我们认为是异常的，要进行观察。 我们发现top 25%的客户税前月收入超过15000，所以我们先观察税前月收入超过15000的客户，有多少异常的。 我们发现异常的主要是月收入超过100万的，如5833333.33300、12500000.00000、8333333.33300、28129588.00000、833333333.30000、83789551.00000、6666666.66700、5000000.00000。我们认为要结合其抵押价值和贷款金额来综合查看该数据是否异常。我们发现很多月入“千万”、“百万”级别的客户，抵押品价值只有几十万，贷款也是几十万甚至几万，与“身家”不符，所以我们认为这些客户的月收入填写异常。不删除这部分客户，另存为一类-2。

- 2.2 数值属性预处理 目的：数值属性离散化，离散化后取值不超过5个
  - 对MONTHLY_INCOME_WHITHOUT_TAX以外的数值属性的2~5的聚类个数进行尝试，并取calinski_harabasz_score得分最高的k值，还绘制最佳k对应的聚类结果。最终k除了APPLY_INTEREST_RATE取4，其他都取5，并且都没有跳跃现象，我们认为聚类效果可以接受，接下来就要更改数值成分类结果了。
  - GAGE_TOTLE_PRICE、APPLY_AMOUNT存在类取值占比过小的问题，下面考虑合并，**合并目的是最终每个类大于5%，并且不能违反连续的原则**。
    - GAGE_TOTLE_PRICE中数目小的类1、4、2分别是取值最大、取值第二大、取值第三大的类，所以可以合并
    - APPLY_AMOUNT中数目小的类2、4、1分别是取值最大、取值第二大、取值第三大的类，所以可以合并

3. 清洗结果

   `preprocess.csv`：分类属性取值数目不超过5个，数值属性离散化后取值数目不超过5个。

三、文件执行说明：

1. 输入下列指令创建环境`ml`，同时也安装好了依赖包

```
conda create --name ml python=3 scikit-learn pandas numpy matplotlib
```

2. 进入环境`ml`

```
conda activate ml
```

3. 从根目录进入`/preprocess`文件夹，执行`preprocess.ipynb`文件，要从上到下逐个单元格执行，即可复现已有的运行结果。

   如果不需要复现，可直接查看根目录下的`preprocess.pdf`，里面有代码运行结果和辅助的分析。