数据挖掘应用实验 1

1. 数据集（10 分）

- （6 分）从 UCI dataset repository 中下载一个数据集，形成数据集 D_1，满足以下要求：
  - 包含至少 4 列以上连续的数值型数据，包含一列符号型数据，作为样本的类别标签；
  - 至少包含 100 个以上的样本；

  下载以后，仔细阅读数据集的使用说明，理解其用途及每一列数据的含义。
- （4 分）选定 D_1 中的两列数值型数据，对其中的每一项数据添加大小不一的噪声，使其中出现离群点。提取这两列数据，将其存入文件，形成数据集 D_2。

2. 实验任务（70 分）

编写程序，完成以下任务：

1. 认识数据（10 分）：对下载的数据进行分析，计算每一列数值型数据的均值、方差，画出该列数据的盒图；
2. 数据标准化（20 分）：分别用 z_score、min-max、十进制小数定标和 logistic 方法对数据集 D_1 进行标准化处理，使所有列的数据处于同一规模，处理后的数据集记为 D_1-zscore、D_1-minmax、D_1-float、D_1-log；
3. 数据离散化（20 分）：对数据集 D_1 的前 4 列分别使用等距离散化、信息增益离散化、卡方离散化、CAIM 离散化方法进行离散化，处理后的数据集记为 D_1-discrete；
4. 离群点检测（20 分）：用 LOF 方法检测数据集 D_2 中的离群点。

对于上述每一个任务，编程过程中可以使用 numpy、pandas、scikit-learn、scorecard-bundle 及 matplotlib 包中的相关功能。

3. 实验报告（20 分）

实验报告应包含实验目的、相关算法介绍（不能照抄、照搬文献、网页中的文字，需要大家对相关方法进行提炼、总结，并用书面语言重新进行描述）、实验过程论述、数据集介绍、实验结果及分析、结论与思考等内容，实验报告的最后应该列出引用的文献。

- 对数据集进行介绍，应至少说明使用了什么样的数据集，从哪里获取了那个数据集（针对 D_1），对其进行了怎么样的处理得到了新的数据集（针对 D_2）；
- 对实验结果的呈现，必须以文字形式进行阐述、解释或者说明，不能只是简单地展示结果的图，否则会减分；调整图的大小，使之清晰美观，否则会减分；
- 应以正规的书面语言进行客观的阐述，切勿使用口语化的表达方式或使用随意的网络用语；
- 插图应使用矢量图，如果用 Wps/Word 书写，则插图应该转成.emf 或者.wmf 格式，如果用 LaTeX 书写，则应使用 pdf 或 eps 格式的矢量图（不能从屏幕截图以后另存为.emf、.wmf、.pdf 或.eps 格式，屏幕截图是位图，不是矢量图），使其在缩放时不失真；图、表要添加编号与标题，并在正文中引用其编号；
- 报告中对使用的算法应引用其出处的参考文献，引用格式为用方括号括起来的上标数字形式，按引用次序顺序编号，并在报告末尾添加“参考文献”一节；每一条文献条目中至少应包括作者名，文章标题，期刊名，期号，卷号，出版年月，pp：页码范围，DOI 号或官网的 URL。

4. 必须提交的材料

- 下载的数据集 D_1 及添加噪声后形成的包含离群点的数据集 D_2 及各个任务处理后得到的结果数据集：各个数据集各自存入一个文件中，文件名为程序中使用该数据集时的名称；
- python 的源程序：用 python 语言（建议使用 python 3.6 以后的版本）实现各个任务的源程序，每一任务的源程序各自存入一个文件（不建议使用 jupyter notebook），文件名能体现其作用；
- pdf 版本的实验报告；
- 以上三部分压缩成一个压缩包，以学号 + 姓名对压缩包进行命名。