数据挖掘实验实验报告

实验一: 数据预处理

姓 名: 柴 博 文
学 号: 04194012
班 号: 大数据 1901

数据挖掘与机器学习
(秋季, 2021)

西安邮电大学
计算机学院
数据科学与大数据专业

2021 年 10 月 10 日

摘 要

本次实验使用 Julia 语言进行实现.

实验报告采用 LaTeX, 在 overleaf 上进行编写.

通过 DataFrames, CSV, XLSX 读取数据, PyPlots, Plots, StatsPlot 绘制图案.

本次实验代码均可以在github 仓库下找到.

目 录

1 概述

1、掌握数据探索统计特征计算、数据可视化等基本方法

2、掌握数据集缺失值、含噪数据的平滑处理、数据变换、数据集成等预处理方法。

3、掌握 PCA 主成分分析等降维方法

- **数据可视化**对某县广电宽带用户的 5000 条数据（或者自己感兴趣的其他领域的数据）进行探索，通过统计特征可视化进行数据分析，探索发现你感兴趣的知识。
- **数据处理** 2、对北京西安的年薪数据（或者自己感兴趣的其他领域的数据）计算均值，方差等统计特征，绘制据箱体图和小提琴图等图，分析北京西安年薪的差异。
- **数据清洗**用”movie_metadata.csv” 数据集（或者自己感兴趣的其他领域的数据）进行案例分析，这个数据集包含了包括演员、导演、预算、总输入，以及 IMDB 评分和上映时间等信息，进行处理缺失数据，可以是添加默认值，删除不完整的行，异常值处理，重复数据处理，规范化数据类型等等。
- **数据集成**合并两个给定数据集：ReaderRentRecode.csv 和 ReaderInformation.csv（或者自己感兴趣的其他领域的数据），其中两个数据集的共同点是具有相同的 num 属性，最终生成一个综合的数据集。
- **PCA** 使用鸢尾花数据集（或者自己感兴趣的其他领域的数据），这个数据集有 150 个样本，其中每个样本有五个变量，其中四个为特征变量, 分别为萼片长度（Sepal length），萼片宽度（Sepalwidth），花瓣长度（Petallength），花瓣宽度（Petalwidth），还有一个变量是其所属的品种的类别变量（Species），这个鸢尾花内别共有 3 种类别分别是山鸢尾（Iris-setosa）、变色鸢尾（Iris-versicolor）和维吉尼亚鸢尾（Iris-virginica），首先对 4 维的原始数据集实现可视化，可视化一组数据来观察数据分布，然后对数据集进行标准化（归一化），接着利用 PCA 主成分分析将数据降到二维。

2 数据可视化

2.1 实验过程

首先讲旧版 Excel 格式的 xls 文件转换为 CSV 文件github

随后使用 CSV 读取文件内容, 并通过 DataFrame 解析格式以及类型

```
file_path = "file/xian_guangdian.csv";
data = file_path |> CSV.File |> DataFrame
```

图1.

4999×14 DataFrame

Row	计费对象	产品名称	产品到期时间	状态	停机类型	客户编号	用户类型
	String15	String15	String15	String15	String15	String15	String
1	ys0015561	宽带10M产品	6/25/2017	正使用	正常	c10002109695	新用户
2	ys0023214	宽带4M产品	7/24/2017	正使用	正常	c10002109697	新用户
3	ys0082301	宽带10M产品	6/25/2017	正使用	正常	c10002109701	新用户
4	ys0022748	宽带10M产品	2/26/2018	正使用	正常	c10002114185	新用户
5	ys0056409	宽带10M产品	1/3/2018	正使用	正常	c10002114884	新用户
6	ys0083681	宽带10M产品	5/8/2017	正使用	正常	c10002118933	新用户
7	ys0074844	宽带10M产品	10/29/2016	已停用	欠费停机	c10006305966	新用户
8	ys0080014	宽带10M产品	1/5/2018	正使用	正常	c10002120172	新用户
9	ys0070152	宽带4M产品	7/15/2015	已停用	客户报停	c10002120173	新用户
10	ys0040259	宽带10M产品	8/25/2017	正使用	正常	c10002120293	新用户
11	ys0074865	宽带10M产品	11/18/2017	正使用	正常	c10002123835	新用户
12	ys0041355	宽带10M产品	5/27/2017	正使用	正常	c10002125494	新用户
13	ys0057767	宽带10M产品	11/3/2016	已停用	欠费停机	c10007202177	新用户
14	ys0056459	宽带10M产品	3/6/2017	正使用	正常	c10002125848	新用户
15	ys0080035	宽带10M产品	2/5/2018	正使用	正常	c10002125908	新用户
16	ys0046632	宽带10M产品	9/14/2017	正使用	正常	c10002126686	新用户
17	ys0130153	宽带10M产品	1/16/2018	正使用	正常	c10002126687	新用户
18	ys0016491	宽带10M产品	11/29/2017	正使用	正常	c10002203661	新用户
19	ys0045160	宽带10M产品	12/26/2017	正使用	正常	c10002204537	新用户
20	ys0096736	宽带10M产品	11/4/2016	已停用	欠费停机	c10008675505	新用户
21	ys0027721	宽带10M产品	1/22/2018	正使用	正常	c10002204542	新用户
22	ys0031693	宽带10M产品	3/13/2018	正使用	正常	c10002204543	新用户
23	ys0087107	宽带10M产品	5/23/2017	正使用	正常	c10002209922	新用户
24	ys0018042	宽带10M产品	4/16/2017	正使用	正常	c10002209924	新用户
25	ys0046621	宽带10M产品	10/21/2017	正使用	正常	c10002209927	新用户
26	ys0032342	宽带10M产品	8/28/2017	正使用	正常	c10002209929	新用户
27	ys0125430	宽带20M产品	11/30/2017	正使用	正常	c10002209946	新用户
28	ys0012610	宽带10M产品	7/4/2017	正使用	正常	c10002217047	新用户
29	ys0087979	宽带10M产品	10/11/2017	正使用	正常	c10002236312	新用户
30	ys0065899	宽带10M产品	6/11/2017	正使用	正常	c10002236313	新用户
31	ys0043657	宽带10M产品	10/11/2017	正使用	正常	c10002236316	新用户
32	ys0076592	宽带10M产品	3/14/2017	正使用	正常	c10002236322	新用户
33	ys0037844	宽带4M产品	6/29/2017	正使用	正常	c10002236332	新用户
34	ys0085178	宽带10M产品	1/11/2018	正使用	正常	c10002236334	新用户
35	ys0082223	宽带10M产品	6/19/2017	正使用	正常	c10002236335	新用户
36	ys0075274	宽带10M产品	3/21/2018	正使用	正常	c10002236336	新用户
37	ys0044635	宽带4M产品	8/25/2017	正使用	正常	c10002236337	新用户
38	ys0085804	宽带10M产品	10/5/2017	正使用	正常	c10002239189	新用户
39	ys0023467	宽带10M产品	11/30/2017	正使用	正常	c10002241607	新用户
40	ys0127407	宽带10M产品	12/23/2017	正使用	正常	c10002247402	新用户
41	ys0062065	宽带10M产品	1/19/2018	正使用	正常	c10002261387	新用户
42	ys0044083	宽带10M产品	8/25/2017	正使用	正常	c10002262665	新用户
43	ys0046875	宽带10M产品	10/6/2017	正使用	正常	c10002262670	新用户

julia>

图 1: 广电信息图

2.2 实验结果和分析

请在这一节详细说明需要分析的内容。

3 PTX 与 x86 指令的比较

以下请按照上面的说明示例，自行安排章节内容。

附录 A 代码

请在附录A中添加代码。请使用如下 C 或者 C++ 的语法高亮描述方法。

```
using XLSX;
using CSV;
using DataFrames;
using Plots;
using Dates;
using StatsPlots;
using PyPlot;

file_path = "file/xian_guangdian.csv";
data = CSV.File(file_path) |> DataFrame
```

参考文献

[1] P. Erdős, *A selection of problems and results in combinatorics*, Recent trends in combinatorics (Matrahaza, 1995), Cambridge Univ. Press, Cambridge, 2001, pp. 1–6.