

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 12 日

摘 要

本次实验使用 Julia 语言进行实现.

如果需要运行本项目代码, 请安装 python 以及 matplotlib

随后打开终端, 运行 Julia

安装 XLSX, CSV, DataFrames, Plots, Dates, Statistics

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	4
2 数据可视化	5
2.1 实验过程	5
2.2 实验结果和分析	7
3 数据处理	8
3.1 实验过程	8
4 数据预处理	9
4.1 实验过程	9
5 数据合并	10
5.1 实验过程	10
6 PCA	11
6.1 实验过程	11
A 代码	12

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**对北京西安的年薪数据（或者自己感兴趣的其他领域的的数据）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

首先使用 Excel 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
图1.

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	正常使用	正常	c10002109695	新用户
2	ys0023214	宽带4M产品	7/24/2017	正常使用	正常	c10002109697	新用户
3	ys0022301	宽带10M产品	6/25/2017	正常使用	正常	c10002109701	新用户
4	ys0022748	宽带10M产品	2/26/2018	正常使用	正常	c10002114185	新用户
5	ys0056409	宽带10M产品	1/3/2018	正常使用	正常	c10002114084	新用户
6	ys0003481	宽带10M产品	5/8/2017	正常使用	正常	c10002118913	新用户
7	ys0074044	宽带10M产品	10/29/2016	已停用	欠费停机	c10006305966	新用户
8	ys0008014	宽带10M产品	1/5/2018	正常使用	正常	c10002126172	新用户
9	ys0078152	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	正常使用	正常	c10002120293	新用户
11	ys0074865	宽带10M产品	11/18/2017	正常使用	正常	c10002123835	新用户
12	ys0041355	宽带10M产品	5/27/2017	正常使用	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	正常使用	正常	c10002125848	新用户
15	ys0008035	宽带10M产品	2/5/2018	正常使用	正常	c10002125908	新用户
16	ys0046632	宽带10M产品	9/16/2017	正常使用	正常	c10002126686	新用户
17	ys0130153	宽带10M产品	1/16/2018	正常使用	正常	c10002126687	新用户
18	ys0016491	宽带10M产品	11/29/2017	正常使用	正常	c10002205661	新用户
19	ys0045168	宽带10M产品	12/26/2017	正常使用	正常	c10002206537	新用户
20	ys0090756	宽带10M产品	11/4/2016	已停用	欠费停机	c10006379599	新用户
21	ys0027721	宽带10M产品	1/22/2018	正常使用	正常	c10002204543	新用户
22	ys0015493	宽带10M产品	5/13/2018	正常使用	正常	c10002204543	新用户
23	ys0007187	宽带10M产品	5/23/2017	正常使用	正常	c10002209922	新用户
24	ys0018042	宽带10M产品	4/16/2017	正常使用	正常	c10002209924	新用户
25	ys0046621	宽带10M产品	10/21/2017	正常使用	正常	c10002209927	新用户
26	ys0032342	宽带10M产品	8/28/2017	正常使用	正常	c10002209929	新用户
27	ys0125430	宽带10M产品	11/30/2017	正常使用	正常	c10002209946	新用户
28	ys0126180	宽带10M产品	7/4/2017	正常使用	正常	c10002217047	新用户
29	ys0007979	宽带10M产品	10/11/2017	正常使用	正常	c10002236312	新用户
30	ys0065899	宽带10M产品	6/11/2017	正常使用	正常	c10002236313	新用户
31	ys0043657	宽带10M产品	10/11/2017	正常使用	正常	c10002236316	新用户
32	ys0076592	宽带10M产品	3/16/2017	正常使用	正常	c10002236322	新用户
33	ys0037844	宽带4M产品	6/29/2017	正常使用	正常	c10002236332	新用户
34	ys0008378	宽带10M产品	2/11/2018	正常使用	正常	c10002236334	新用户
35	ys0002223	宽带10M产品	6/29/2017	正常使用	正常	c10002236335	新用户
36	ys0075274	宽带10M产品	3/21/2018	正常使用	正常	c10002236356	新用户
37	ys0064635	宽带4M产品	8/25/2017	正常使用	正常	c10002236337	新用户
38	ys0005884	宽带10M产品	10/6/2017	正常使用	正常	c10002239189	新用户
39	ys0023467	宽带10M产品	11/30/2017	正常使用	正常	c10002241687	新用户
40	ys0127487	宽带10M产品	12/23/2017	正常使用	正常	c10002247482	新用户
41	ys0062065	宽带10M产品	1/19/2018	正常使用	正常	c10002261387	新用户
42	ys0044083	宽带10M产品	8/25/2017	正常使用	正常	c10002262665	新用户
43	ys0044875	宽带10M产品	10/6/2017	正常使用	正常	c10002262670	新用户

图 1: 广电信息 CSV

```
quality =  
"lab1/julia/file/xian_guangdian.csv" |>  
CSV.File |>  
DataFrame |>  
data ->  
begin  
  combine(nrow, groupby(select(data, :客户等级), :客户等级)) |>  
  data -> rename(data, :nrow => "用户数量") |> println  
  combine(nrow, groupby(select(data, [:客户等级, :网络类型]),  
                           [:客户等级, :网络类型]))  
  
end |>  
data ->  
  rename(data,  
    :nrow => :quantity,  
    :网络类型 => :net_kind,
```

```

:客户等级 => :user_level)
data = combine(groupby(quality, :net_kind), [:user_level, :quantity])
dict = Dict(
    "5星ABD客户" => "star_5ABD",
    "离线" => "out_link",
    "3星AB客户" => "star_3AB",
    "1星D客户" => "star_1D",
    "1星A客户" => "star_1A",
    "VIP商业个人客户" => "vip",
    "3星AD客户" => "star_3AD",
)
1:(data|>nrow) .|>
i -> begin
    data[i, :net_kind] =
        Dict(
            "农网用户" => "village",
            "城网用户" => "city",
            " " => "unknown"
        )
    data[i, :net_kind] = dict[data[i, :net_kind]]
end
gp = groupby(data, :net_kind)
gp |>
keys .|>
kind -> @df combine(gp[kind], [:user_level, :quantity]) plot(
    :user_level,
    :quantity,
    label = "$kind",
) |> fig -> savefig(fig, "lab1/julia/images/first_$kind")

```

随后将数据根据客户等级进行分组, 总共有 7 组, 见图2.

7×2 DataFrame		
Row	客户等级 String31	用户数量 Int64
1	5星ABD客户	3695
2	离线	498
3	3星AB客户	151
4	1星D客户	403
5	3星AD客户	76
6	VIP商业个人客户	63
7	1星A客户	113

图 2: 分组结果图

再将每组一网络类型进行分组, 图3.

然后将每组画到折线图之上, 图4

7×2 DataFrame

Row	user_level	quantity
	String31	Int64
1	star_5ABD	2192
2	out_link	313
3	star_3AB	88
4	star_1D	204
5	start_3AD	40
6	vip	57
7	star_1A	74

7×2 DataFrame

Row	user_level	quantity
	String31	Int64
1	star_5ABD	1501
2	out_link	185
3	star_3AB	62
4	star_1D	197
5	start_3AD	36
6	vip	6
7	star_1A	39

3×2 DataFrame

Row	user_level	quantity
	String31	Int64
1	star_5ABD	2
2	star_3AB	1
3	star_1D	2

图 3: 分组结果图

2.2 实验结果和分析

通过该次结果可以看出, 在办理了广电业务的客户之中,5 星 ABD 客户数目远远多余其他客户, 而且明显城区用户多余农村用户

但是低级用户和高级用户的数量几乎差不多, 而且最关键的是两个图的趋势是相似的, 说明农村和城市对于网络的需求是很一致的

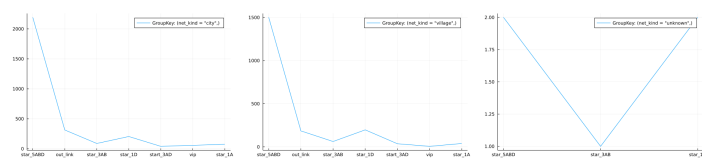


图 4: 城市居民, 农村, 未登记

3 数据处理

3.1 实验过程

使用 XLSX 将文件内容读入, 并使用 DataFrame 对数据进行类型判断并转换位 DataFrame 类型随后使用统计模块中的统计方法求数据的均值, 方差, 标准差, 协方差矩阵, 图5 在使用 Plots 进行绘图, 图6

```
file_path = "lab1/julia/file/xian_beijing_salary.xlsx"
salarys = DataFrame(XLSX.readdata(file_path,
                                   "Sheet1!C3:D14"), :auto) .|> identity
salary = [salarys.x1, salarys.x2]
println("mean:$(salary .|> Statistics.mean)")
println("var:$(salary .|> Statistics.var)")
println("std:$(salary .|> Statistics.std)")
println("cov:$(salary .|> Statistics.cov)")
println("cor:$(salary .|> Statistics.cor)")
violin(["Xi'an"], salarys.x1, label = "Xi'an")
violin!(["Beijing"], salarys.x2, label = "Beijing")
boxplot(["Xi'an"], salarys.x1, label = "Xi'an")
boxplot!(["Beijing"], salarys.x2, label = "Beijing")
```

```
julia> println("mean:$(salary .|> Statistics.mean)")
mean: [77.58333333333333, 97.83333333333333]
julia> println("var:$(salary .|> Statistics.var)")
var: [435.71860606060604, 492.8787878787878]
julia> println("std:$(salary .|> Statistics.std)")
std: [20.8738005049, 22.2458558557]
julia> println("cov:$(salary .|> Statistics.cov)")
cov: [8.0 -48.0 -46.0 -54.0 -52.0 -54.0 -46.0 -52.0 -58.0 -56.0 8.0; -48.0 288.0 238.0 278.0 268.0 278.0 238.0 288.0 268.0 258.0 388.0 0.0; -46.0 238.0 264.5 318.5 299.0 318.5 264.5 238.0 299.0 287.5 345.0 0.0; -54.0 278.0 318.5 364.5 351.0 364.5 318.5 278.0 351.0 359.5 394.5 0.0; -52.0 268.0 299.0 351.0 338.0 268.0 338.0 299.0 351.0 338.0 351.0 0.0; -54.0 278.0 318.5 364.5 351.0 364.5 318.5 278.0 351.0 359.5 394.5 0.0; -46.0 238.0 299.0 318.5 264.5 238.0 299.0 287.5 345.0 0.0; -58.0 258.0 287.5 337.5 325.0 337.5 325.0 258.0 337.5 337.5 337.5 258.0 325.0 325.0 0.0; -56.0 308.0 345.0 405.0 388.0 405.0 388.0 308.0 375.0 420.0 0.0; 8.0 8.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0]
```

图 5: 均值, 方差, 标准差, 协方差

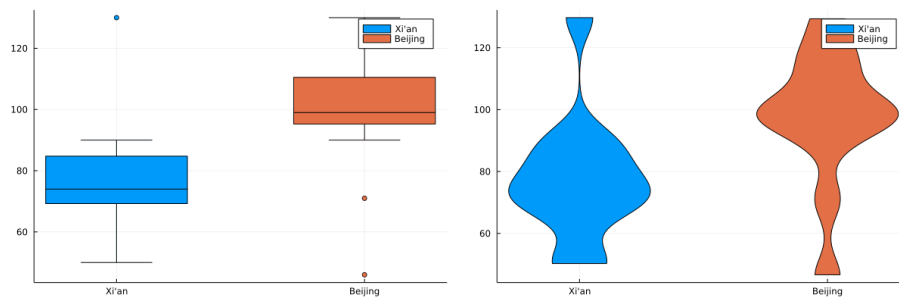


图 6: 箱型图和小提琴图

4 数据预处理

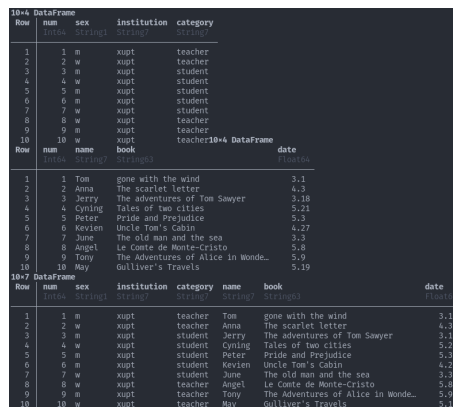
4.1 实验过程

5 数据合并

5.1 实验过程

读取数据表, 通过 join 表上的 num 列对两张表进行合并, 图7

```
["lab1/julia/file/4ReaderInformation.csv",  
 "lab1/julia/file/4ReaderRentRecode.csv"] .|>  
CSV.File .|>  
DataFrame |>  
dates -> begin  
  println(dates...)  
  innerjoin(dates..., on = :num) |>  
  file -> begin  
    file |> println  
    CSV.write("lab1/julia/file/join.csv", file)  
  end  
end
```



Row	num	sex	institution	category
1	1	m	xupt	teacher
2	2	w	xupt	teacher
3	3	m	xupt	student
4	4	w	xupt	student
5	5	m	xupt	student
6	6	m	xupt	student
7	7	w	xupt	student
8	8	w	xupt	teacher
9	9	m	xupt	teacher
10	10	w	xupt	teacher

Row	num	name	book	date
1	1	Tom	gone with the wind	3.1
2	2	Anna	The scarlet letter	4.3
3	3	Jerry	The adventures of Tom Sawyer	3.18
4	4	Cyning	Tales of two cities	5.21
5	5	Peter	Pride and Prejudice	5.3
6	6	Kevin	Uncle Tom's Cabin	4.27
7	7	June	The old man and the sea	3.3
8	8	Angel	Le Conte de Monte-Cristo	5.8
9	9	Tony	The Adventures of Alice in Wonde...	5.9
10	10	May	Gulliver's Travels	5.19

Row	num	sex	institution	category	name	book	date
1	1	m	xupt	teacher	Tom	gone with the wind	3.1
2	2	w	xupt	teacher	Anna	The scarlet letter	4.3
3	3	m	xupt	student	Jerry	The adventures of Tom Sawyer	3.18
4	4	w	xupt	student	Cyning	Tales of two cities	5.21
5	5	m	xupt	student	Peter	Pride and Prejudice	5.3
6	6	m	xupt	student	Kevin	Uncle Tom's Cabin	4.27
7	7	w	xupt	student	June	The old man and the sea	3.3
8	8	w	xupt	teacher	Angel	Le Conte de Monte-Cristo	5.8
9	9	m	xupt	teacher	Tony	The Adventures of Alice in Wonde...	5.9
10	10	w	xupt	teacher	May	Gulliver's Travels	5.19

图 7: 数据合并

6 PCA

6.1 实验过程

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
```