数据科学导论实验指南

实验 2

Python 数据应用实验

（版本号：2021 年 2 月 21 日版本）

目录

实验 2：Python 数据应用实验

1.实验目的

为后续上机实验做准备，熟悉常用的 python 基本数据函数分析库：matplotlib 和 scipy；掌握常用的 python 数据分析方法：数据探索；数据清理；数据变换和标准化；数据相关性分析

2.实验平台

工具：anconda/spyder/pycharm
语言：python

3.实验内容和要求

● Python 基本函数分析库

（一）熟悉常用的 Scipy 科学计算操作

(1) 求解非线性方程组 $5x_1 - x_2^2 = 1, x_1^2 - x_2 = 6$

1) 在 scipy.optimize 中导入求解方程组的函数 import fsolve

2) 定义一个要求求解的非线性方程组 f(x)

3) 输入初值[1, 1]并求解

4) 输出结果

(2) 线性插值计算

1) 导入线性插值函数 interpolate

2) 随机创建自变量 x,

3) 自定义创建因变量 y1 和 y2

4) 通过一维线性插值函数 scipy.interpolate.interpld(x, y, kind='linear', axis=-1, copy=True, bounds_error=None, fill_value=nan, assume_sorted=False)拟合(x,y1)和（x,y2）两个线性插值

5) 输出 x 为 0.1，5，9 时，线性插值 y1 和 y2 的值。

（二）熟悉常用的 matplotlib 绘图操作

(1) 绘制简单的 $y_1 = \sin x, y_2 = \cos x$ 在一张图上，图中要求：

1）设置图像大小为（12，5）；

2）设置标题“sinx 与 cosx”；

3）设置标签为 y1=sinx,y2=cosx、线条颜色分别为红色，蓝色、线条大小都为 2；为 y1=sinx,y2=cosx；

4）设置 x 轴，y 轴名称分别为 x 值，y 值；

5）输出结果。

● Python 基础数据分析

（一） 探索数据描述

（在执行以下每个操作之后，使用 head（）、tail（）来显示数据。）

(1) 从 kaggle 下载 fifa19 完整球员数据集(https://www.kaggle.com/karangadiya/fifa19)在数据.csv 文件包含许多列，包括每个玩家的行号、ID、name、age....。读取 fifa19 数据集。

(2) 将数据读入 Pandas 数据框。

(3) 筛选年龄在 25 岁以下的年轻球员。

(4) 根据 Jumping 分数对数据进行排序。

(5) 使用 describe（）方法显示列 Volleys 和 Dribbling 的 count、mean、std、min 和分位数数据。

（二） 数据清理

（在执行以下每个操作之后，使用 head（）、tail（）来显示数据。）

(1) 从数据中通过 drop 删除以下三个属性 Photo, Flag, Club Logo，显示结果；

(2) 用 isnull（）找出所有缺少值的条目；

(3) 用 0 填补空值；

(4) 将“Real Face”属性的特征值‘Yes’替换为 1，‘No’变为 0；“Preferred Foot”属性的特征值‘Right’变为 1，‘Left’变为 0, 并输出结果。

（三） 数据变换操作

(1) 定义最大-最小规范化函数，并对列 Age 进行最大-最小化处理，输出结果，并用直方图显示；

$$x' = \frac{x - \min}{\max - \min}$$

(2) 定义 z-score 规范化函数，并对列'Age','Crossing','Finishing'进行 z-score 规范化处理，输出结果，并用箱图显示

$$x' = \frac{x - \bar{x}}{\sigma}$$

（四） 数据相关性分析

(1) 制作数据中两个属性 Dribbling 和 BallControl 的 Pearson 相关系数，并判断这两个属性是正相关还是负相关，或者没有相关性？

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

其中，r_{xy} 为变量 x 和 y 之间线性关系的相关系数。

(2) 使用 pandas 中包含内置的求解 pearson 系数方法函数,求两个属性 Dribbling 和 BallControl 的 Pearson 相关系数，判断这两个属性是正相关还是负相关，或者没有相关性？

(3) 判断前两种方法算出结果是否一致？并用散点图绘制上述两个属性的相关性分析图

（五） 扩展实验

(4) 假设你负责一个数据分析项目，研究 FIFA 19 的数据集，你认为你可以从

给定的数据集得出什么样的结论。写出至少 3 个问题，找出可以用来回答你问题的数据。例如，一个问题是每个年龄组有多少球员（假设你每 5 年指定一个年龄组，比如 16-20，21-25，26-30…？读取 FIFA 19 数据集并自主绘制该数据集至少三个不同的图表？等）

(5) 参考绘制图官网：https://matplotlib.org/stable/index.html

4. 实验报告

《数据科学导论》课程实验报告				
题目：		姓名		日期
实验环境：				
实验内容与完成情况：				
出现的问题：				
解决方案（列出遇到的问题和解决办法，列出没有解决的问题）：				