DSCI 6001P 数据科学基础

作业 1 数据预处理和线性回归

提交截止日期：10.7 号 24 点之前
提交方式：电子版发送至 张刚毅(gangyi@mail.ustc.edu.cn)

1. 现已使用 Pandas 读取数据集 *challenge.csv*

- 请提取该数据集的字段名称，将结果存为 `cols`
- 请获取给数据的字段和样本数量，将结果分别存为 `col_num` 和 `sam_num`
- 请获取该数据集的前五行记录，将最后的 DataFrame 存为 `five_data`

开始答题：

```
import pandas as pd
titanic = pd.read_csv("challenge.csv")

# 获取字段名称
cols =

# 获取字段数量
col_num =

# 获取样本数量
sam_num =

# 获取样本前 5 行样本
five_data=
```

2. 现已使用 Numpy 生成服从均匀分布的一维数据集，样本容量为 100;

- 使用 `scipy` 库中的 `stats` 模块，对生成的数据进行正态性检验，将检验的结果存为 `model`

开始答题：

```
import numpy as np
from scipy.stats import stats
test_data = np.random.random(size=100)

# 验证分布
model  =

print(model)
```

3. 下列属于衡量数据整体散度的是（可多选）：

a. 欧式距离
b. 标准差
c. 分位数
d. 众数

4. 现已使用 Pandas 生成 Series 对象 `example_data`

- 请使用 `isnull()`函数确定 `example_data` 是否含有缺失值，将最后的结果存为 `boolean_array`
- 请使用 `fillna()`函数使用字符串 `missing` 替换缺失值，将替换后的 Series 对象存为 `new_data`

开始答题：

```
import pandas as pd
import numpy as np
example_data = pd.Series([1,2,3,np.nan,4])

# 判断是否含有缺失值
boolean_array =

print(boolean_array)

# 缺失值替换
new_data =

print(new_data)
```

5. 现已使用 Pandas 读取数据集 *birthrate.csv*

- 请对该数据集的 `birth_rates` 特征使用四分位数作为切分点，通过 `qcut()`函数完成等频离散化；将最后的结果存为 `data_qcut`

该数据集详情为：

	country	birth_rates	per_capita_income	proportion_of_population_farming	infant_mortality
0	Venezuela	46.4	392	0.40	68.5
1	Mexico	45.7	110	0.61	87.8
2	Ecuador	45.3	44	0.53	115.8
3	Colombia	38.6	158	0.53	106.8
4	Ceylon	37.2	81	0.53	71.6

开始答题：

```
import pandas  as pd
data = pd.read_csv('birthrate.csv')

#请在下面作答
```

data_qcut =

print(data_qcut)

6. [线性回归] 给定数据：
   X: 0, 0, 1, 1, 2, 2; Y: 0, 1, 0, 1, 0, 1.
(a)拟合模型 Y = a + b X + ε （手算）
(b)拟合模型 Y = b X + ε （手算）

7. [线性回归] 给定数据：
   X: 0, 0, 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9, 10, 10, 11, 11,12, 12
   Y: 42, 44, 51, 48, 51, 54, 57, 54, 57, 63, 61, 69, 70, 70, 70, 72, 74, 83, 84, 81, 84, 85, 91, 86, 91, 95

   写程序拟合模型 Y = a + b X + ε ,并画图显示数据点和拟合曲线。

8. 给定$f(x) = x^3 - 6x^2 + 11x - 6$, 编程实现梯度下降法计算出使 f(x)=0 的解，绘图展示梯度下降法的迭代过程。

9. [自学牛顿方法] 牛顿方法和梯度下降法有什么异同点？请写出牛顿方法的推导过程，编程实现牛顿方法求解上一题，并编程绘图展示迭代计算过程。

10. 数据标准化是将数据按比例缩放到一个特定区间，其主要包括数据同趋化处理和无量纲化处理两个方面。数据标准化的方法有很多种，常用的有最小-最大标准化和 z-score 标准化。
请用户对本题中的变量(不包括变量 ID)进行 z-score 标准化

数据说明：本题数据来自 KEEL，数据集一共包含 1 列 ID，4 列特征变量，共 100 个样本点。

列名	类型	说明	示例
ID	Int	ID 样本号	1
CT	Float	Cement 黏固粉	295.7
FA	Float	FlyAsh 粉煤灰	98.8
WT	Float	Water 水	185.6
SP	Float	SuperPlasticizer 超增塑剂	14.2

预设变量：本题使用的数据变量名、含义及其类型如下：

变量名	含义	类型
data	数据集	DataFrame

答题要求

对 `data` 中的列(不包括变量 ID)进行 z-score 标准化，类型为 DataFrame 对象。

开始答题：

```
import pandas as pd

data =
```