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号晚上24点之前

提交方式：电子版发送至 潘航(hungpaan@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=

1. 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)

1. 下列属于衡量数据整体散度的是（可多选）：
2. 欧式距离
3. 标准差
4. 分位数
5. 众数
6. 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

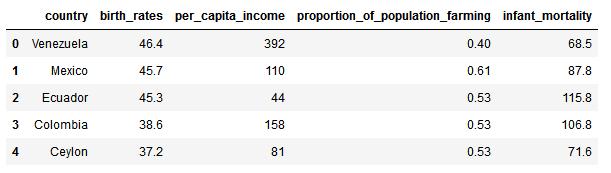
new\_data =

print(new\_data)

1. 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：



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)

1. [线性回归] 给定数据：

X: 0, 0, 1, 1, 2, 2; Y: 0, 1, 0, 1, 0, 1.

1. 拟合模型 Y = a + b X + ε（手算）
2. 拟合模型 Y = b X + ε （手算）
3. [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。

1. 给定, 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。
2. [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。
3. 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 =