[EEG3235A]Python和数据科学

**Answer to Assignment 1**

1. 当数据中存在缺失值（Missing Value）时，有哪些处理方法？（查阅文献，给出参考文献）

如果数据集中存在标签（Label，监督学习），而某条数据的标签缺失，则删除该条数据。

否则，我们可以[1]：

1. 忽略（删除）有缺失值的样本；
2. 使用特殊值来填补缺失值，比如-1；
3. 使用可用特征的均值来填补缺失值；
4. 使用相似样本的均值填补缺失值；
5. 使用另外的机器学习算法预测缺失值。

参考文献：

[1] Peter Harrington著, 李锐、李鹏、曲亚东、王斌译. 机器学习实战. 人民邮电出版社, 2017.

1. 编写Python代码，统计Adult Data Set中的缺失值情况。比如，一共有多少个缺失值？有哪些特征('age', 'wc', 'fnlwgt', 'edu', 'edu-num', 'ms', 'occ', 'rs', 'race', 'sex', 'cap-g', 'cap-l', 'hpw', 'native'等)存在缺失值？有多少个样本存在缺失值？有多少个样本只有一个特征存在缺失值？有多少个样本有两个特征同时存在缺失值？
2. 答案：数值计算（实验）的方法、过程（参数的选择）、结论等。
3. Python代码（文本、不要图片剪贴、不需要运行结果）。

经过对Adult数据集中的缺失值进行统计分析，我们得到Adult数据集中：

[1] 共存在4262个缺失值；

[2] 有三个特征存在缺失值，即wc，occ和native；

[3] 有2399个样本存在缺失值；

[4] 有563个样本仅有一个特征存在缺失值；

[5] 有1809个样本同时有两个特征存在缺失值；

[6] 有27个样本同时有三个特征存在缺失值。

Python代码如下：

import pandas as pd

import numpy as np

"""

读取已经处理过的的数据

"""

path = "D:\\Work at STU\\Data\\adult.csv"

data = pd.read\_csv(path)

D = data.isnull().sum(axis=1)

E = data.isnull().sum(axis=0)

print('Name of features containing missing values:',

E[E > 0].index) #存在缺失值的特征

print('Total number of missing values:',

sum(E)) #缺失值的总数

num = len(D[D > 0])

num1 = len(D[D == 1])

num2 = len(D[D == 2])

num3 = len(D[D == 3])

print('Total number of samples containing missing values:',

num) #存在缺失值的样本数目

print('Total number of samples containing one missing value:',

num1) #只有一个特征存在缺失值的样本数目

print('Total number of samples containing two missing values simultaneously:',

num2) #两个特征同时存在缺失值的样本数目

print('Total number of samples containing three missing values simultaneously:', num3) #三个特征同时存在缺失值的样本数目

"""

非常感谢赖丽琴同学。以上代码受赖丽琴同学的作业的代码启发。因为她应用了对象data.isnull()的属性和方法。

以下黑体字代码为 赖丽琴 同学的代码；

之后，还有许多同学递交的作业的代码也是基于对象data.isnull()的属性和方法，在此不一一叙述。

"""

**'''**

**统计Adult Data Set中的缺失值情况**

**'''**

**#统计一共有多少个缺失值**

**A=data.isnull().sum().sum()**

**print('缺失值总数:%s' %(A))**

**#统计有哪些特征存在缺失值**

**B=data.isnull().any()**

**print('各特征存在缺失值情况如下\n')**

**print(B)**

**#统计有多少个样本存在缺失值**

**C=data.isnull().any(axis=1).sum()**

**print('存在缺失值的样本个数：%s' %(C))**

**#统计有多少个样本只有一个特征存在缺失值**

**D=data.isnull().sum(axis=1);D1=len(D.index[D.values==1])**

**print('只有一个特征存在缺失值的样本个数：%s' %(D1))**

**#统计有多少个样本有两个特征同时存在缺失值**

**D2=len(D.index[D.values==2])**

**print('两个特征同时存在缺失值的样本个数：%s' %(D2))**

"""

以上黑体字代码为 赖丽琴 同学的代码

"""

经统计分析，我们可以得到：

1. 整个数据集的样本数（32561）足够大；

2. 包含缺失值样本的比例（2399/32561 = 7.37%）较小；

3. 整个数据集中正负样本（高、低收入样本）的比例是：24.08%:75.92%； 去掉含缺失值样本的数据集中的正负样本比例是：24.89%:75.11%。也就是说，删除包含缺失值的样本不会明显影响剩余数据集中正负样本的比例。

所以，我们选择删除包含缺失值的样本。

逻辑回归（Logistic Regression）中用到sigmoid函数，即。由于，即它对结果的预测不具有任何倾向性。因此，如果把缺失值用0替代，既可以保留现有数据，也不会对误差造成影响。所以，可以选择用0来替换所有缺失值 [1]。

参考文献：

[1] Peter Harrington著, 李锐、李鹏、曲亚东、王斌译. 机器学习实战. 人民邮电出版社, 2017.

1. 阅读关于《How to lie with statistics》的文章，谈谈你的看法。