

实验三实验报告

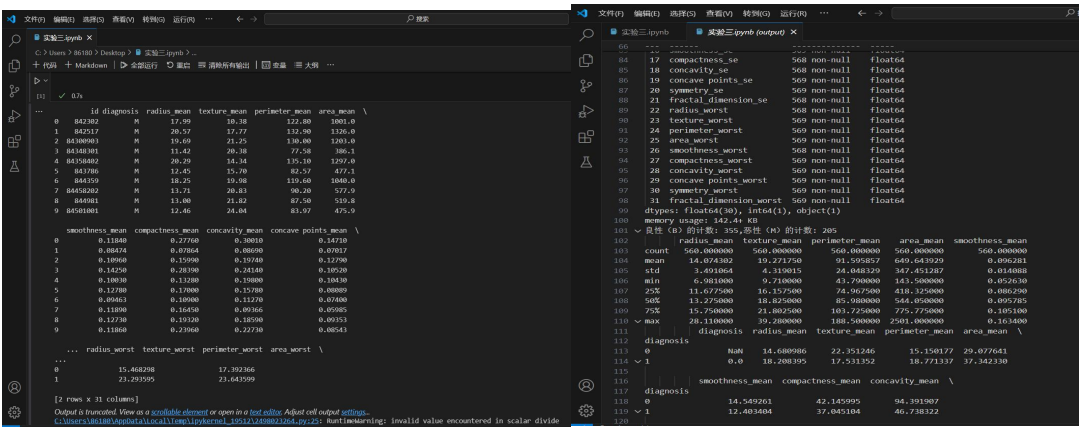
数据分析及实验 刘沛 PB22061259

任务概述

威斯康辛州乳腺癌数据集（Breast Cancer Wisconsin Dataset）由威斯康辛州医院的 Dr. William H. Wolberg 收集 得到涵盖了从 569 个患者收集的乳腺肿瘤特征的测量值及相应肿瘤的诊断标签（良性，恶性）。现欲对该数据集进行探索性分析、可视化展示和简单统计推断，请你按要求编写 Python 代码实现任务列表中的内容。

任务列表

1. （33%）导入 pandas 库，并使用相关方法进行数据集读取、基本信息处理和探索性分析等操作（各子任务请分别使用一行代码完成）。



```
import pandas as pd
df = pd.read_csv('breast-cancer-wisconsin.data')
df.info()
df.describe()
```

The screenshot shows a Jupyter Notebook interface. The left pane displays the code cells, and the right pane shows the output. The code imports pandas as pd and reads the dataset into a DataFrame df. The output shows the DataFrame's structure and summary statistics.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	\
0	842302	M	17.99	10.38	122.80	1001.0	
1	842517	M	20.57	17.77	132.90	1326.0	
2	8430063	M	19.69	21.25	130.00	1301.0	
3	84340301	M	11.42	20.38	77.58	386.1	
4	84340402	M	20.20	14.34	137.10	1297.0	
5	843786	M	12.45	15.70	82.57	477.1	
6	844359	M	18.25	19.98	133.60	1040.0	
7	84456269	M	11.71	20.83	96.20	377.9	
8	844981	M	11.00	21.82	87.50	519.8	
9	84501801	M	12.40	24.04	81.97	479.9	

The output also shows summary statistics for the dataset:

	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	\
0	0.11840	0.27760	0.30810	0.54710	
1	0.08474	0.07864	0.08090	0.07017	
2	0.10908	0.15900	0.19740	0.12790	
3	0.14750	0.29100	0.24140	0.36700	
4	0.10030	0.13200	0.19000	0.10430	
5	0.12700	0.17000	0.15300	0.08000	
6	0.08661	0.10000	0.11270	0.07600	
7	0.11800	0.16050	0.09366	0.05085	
8	0.11730	0.19130	0.18700	0.06351	
9	0.11860	0.23060	0.22730	0.08541	

这个问题中，只需要熟悉 pandas 库，就可以很轻松地完成这些最基本的表格数据分析和处理的基础任务。虽然我之前并不太了解 python 这门语言，对 pandas 库也不知道，但是在网页上简单地了解之后，也可以很自如地使用。

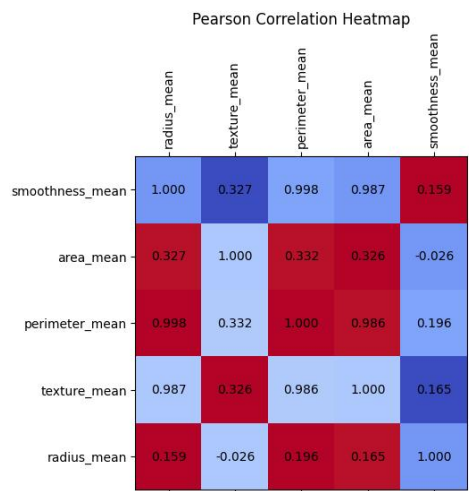
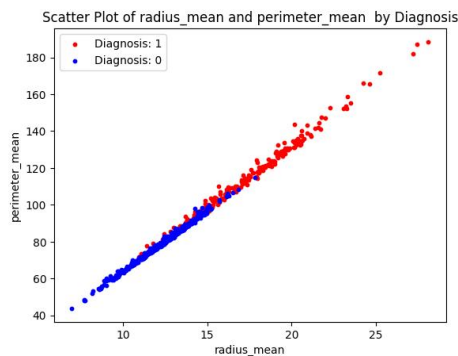
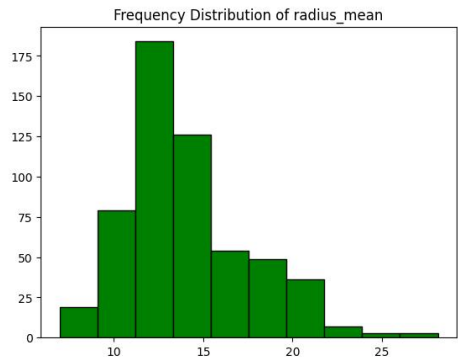
2. （17%）导入 numpy 和 matplotlib 库，对数据集 df 进行一定数据可视化分析。

通过调用 numpy 库可以很容易地完成数组或者矩阵的运算。而通过调用 matplotlib 库可以完成一些基本的数据可视化分析。

这个模块里面的前两个任务都十分容易，但是第三个任务里涉及到热力矩阵图，我之前未曾了解过：于是我在网络上查找了相关定义，并且学会使用库中的函数 matshow 来完成热力矩阵图的完成。

热力矩阵图（Heatmap）是一种使用颜色表示数据值的可视化工具，常用于展示数据之间的关联或比较。在热力矩阵图中，不同的颜色代表不同的数值大小，通过颜色的深浅或色调的变化，可以直观地看出数据之间的差异和趋势。

以下为这个模块的问题里绘制出来的图像：



3. (18%) 线性回归是一类经典的统计建模方法。

Q1

$\hat{\mathbf{w}} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \mathbf{y}$ 按照这个形式解的定义来计算出向量 \mathbf{w} 的值

这里我们首先构造出所需的特征变量的特征矩阵

```
expanded_X= np.column_stack((np.ones(len(X)), X,X*X))
```

然后其实只需要按照给出来的公式，进行简单计算即可
结果如下

模型参数的最小二乘估计为: `[-4.72267699 -0.44070983 3.13966962]`

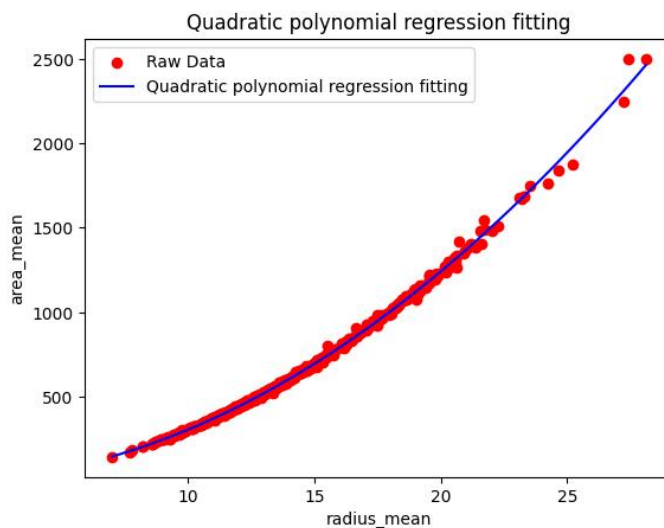
Q2

使用 `polyfit` 得到的参数，粗略来看其实和用最小二乘法给出的拟合差不多，但是我没有弄明白为什么会有小数点后面两到三位的细微差异

使用 `numpy.polyfit()` 得到的模型参数估计为: `[3.14186228 -0.44260792 -4.70867951]`
两种方法得到的参数估计存在差异

Q3

通过调用 `matplotlib` 库里面的 `plot` 函数设置好参数就可以轻松得到下面的曲线



Q4

考虑圆形区域的半径与面积的关系， $A = \pi r^2$ ，因此，从特征的含义上来看，我们预期 `area_mean` 与 `radius_mean` 之间应该存在一个二次关系，而不是线性的。

从散点图上来看，观察到数据点呈现明显的曲线分布（特别是当半径增加时，面积的增加速度越来越快），这进一步表明二者之间可能不是简单的线性关系。

综上所述，线性回归肯定不适用与这个场景,最好是使用二次函数来进行拟合

#4. (18%) 数据降维，即选择性地削减数据集的属性维度，可以在牺牲一小部分信息的情况下大幅增加数据处理的效率。其中主成分分析是一种应用非常广泛的数据降维方法。

以下是 PCA 方法的主要步骤：

1. 标准化处理：首先，对原始数据进行标准化处理，即减去每个特征的均值并除以其标准差，使得每个特征的平均值为 0，方差为 1。这一步是为了消除不同特征之间的量纲差异。
2. 计算协方差矩阵：计算标准化后数据的协方差矩阵。协方差矩阵的每个元素表示两个特征之间的协方差，反映了它们之间的线性相关性。
3. 计算协方差矩阵的特征值和特征向量：对协方差矩阵进行特征分解，得到其特征值和特征向量。特征值表示对应特征向量的重要程度，而特征向量则代表新的主成分方向。
4. 选择主成分：按照特征值的大小，从大到小选择前 k 个特征值对应的特征向量作为

主成分。通常，选择的主成分个数 k 小于原始特征的个数，以达到降维的目的。

5. 转换数据到新空间：将原始数据投影到选定的主成分上，得到降维后的数据。这一步是通过将原始数据乘以选定的主成分矩阵来实现的。

这次实验当中，希望将一个二维数据降成一维。而且其实标准化处理并不是必要的，这点是我通过询问助教获悉。

Q1.

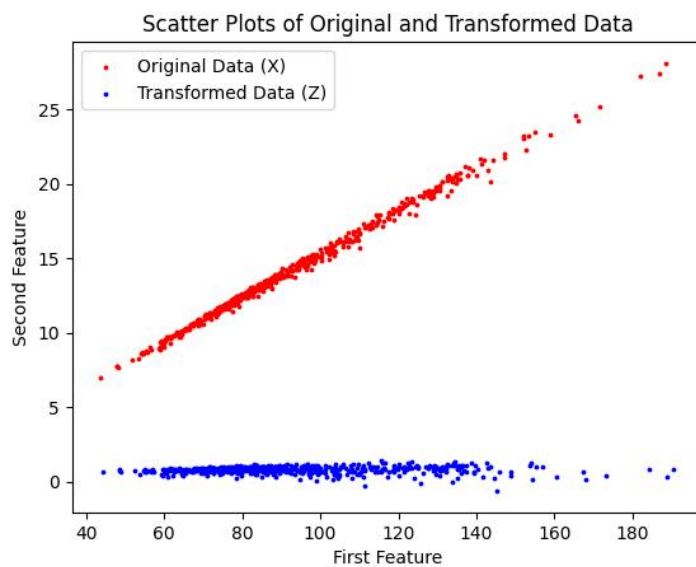
```
协方差矩阵 corX:
[[578.32210982  83.77072122]
 [ 83.77072122  12.18753099]]

排序后的特征值 eigV:
[5.90457503e+02 5.21373729e-02]

排序后的特征向量矩阵 eigMat:
[[ 0.98966947 -0.14336785]
 [ 0.14336785  0.98966947]]

特征向量矩阵的正交性验证:
[[ 1.00000000e+00 -6.44062151e-18]
 [-6.44062151e-18  1.00000000e+00]]
检查是否与单位矩阵接近 True
```

Q2.



Q3.

```
协方差矩阵 corZ:
[[ 5.90457503e+02 -4.51429790e-14]
 [-4.51429790e-14  5.21373729e-02]]

特征值 eigV:
[5.90457503e+02 5.21373729e-02]

corZ的对角线元素 diagonal_corZ:
[5.90457503e+02 5.21373729e-02]
检查对角线元素与排序后的特征值是否接近: True
```

最后，我们删去 Z 的第二列来完成数据降维

```
# 基于以上分析，我们选择保留方差最大的主成分（即对角线上最大的元素对应的成分）  
# 因此，我们将删除Z的第二维数据以完成降维  
Z_reduced = Z[:, :1] # 只保留第一列(特征值大的对应的那个特征向量)，即第一个主成分
```

5. (14%) 假设检验是数理统计学中根据一定假设条件由样本推断总体性质的方法，在统计推断中的地位举足轻重，其中 t 检验是一类非常重要的假设检验方法。

#Q1

成组检验（独立样本 t 检验）适用于两个独立样本之间的比较，即两个样本来自于两个独立的总体，并且没有配对关系。每个样本中的观察值是相互独立的，并且两组样本的大小可以相同也可以不同。

成对检验（配对样本 t 检验）则适用于同一组个体在不同条件或时间下的测量值的比较。在这种情况下，每个观察值都与另一个观察值配对，例如，同一个个体在接受治疗前后的测量值。

我们假设 diagnosis 的值用于将 concavity_worst 特征数据分成两组（良性和恶性）。在这种情况下，两组数据是独立的，没有配对关系。因此，我们应该使用成组检验（独立样本 t 检验）来比较这两组样本间的均值差异。

Q2.

其中，第一组数据为良性，第二组为恶性。

```
第一组数据的平均值: 0.1663615971830986  
第二组数据的平均值: 0.44671356097560977
```

Q3.

原假设: $H_0: \mu_1 \leq \mu_2$

其中， μ_1 和 μ_2 分别是第一组和第二组数据的总体均值。

这个原假设表示第一组数据的均值不大于第二组数据的均值。

Q4.

通过调用函数

```
t_statistic, p_value = stats.ttest_ind(group1, group2, equal_var=False, alternative='greater')
```

可以得到以下结果:

```
t统计量: -19.01722049062515  
p值: 1.0  
接受原假设，没有足够的证据认为第一组数据的均值显著大于第二组数据的均值。
```

实验思路

这次实验，其实就是将课本上学到的一些最基本的数据处理手段应用到给出的数据集，很多时候其实 python 的库函数中都有相关参数的具体计算函数，我们只需要清楚的了解这些参数具有什么作用 and 意义，就可以很好的完成简单的数据预处理。

实验总结

我觉得这次实验里面，最有意思的部分还是数据降维中的主成分分析法。

虽然这次实验里面给出的具体任务十分清晰明了，以至于我甚至可以完全在不了解 PCA 的情况之下，就可以更着要求一步一步的来完成数据降维。但是，其实我是不理解一些操作的意义的，可能上课时间有点久远，加上我也很少回头复习，以至于我连协方差矩阵是什么都忘了。

不过，通过翻看课件和在网络上查找相关资料，我最终还是理解了 PCA 的思想和操作方法。其实，通过这次实验，可以很强烈的感受到数据分析这门课程，光上课听老师讲解相关原理和方法还远远不够，只有自己真正的去面对数据，动手实验来完成具体数据的分析，才能对这些原理和方法有着更深刻的理解。