中国科学技术大学

University of Science and Technology of China

Lab 3

姓名：高茂航

学号：PB22061161

日期：2024.4.28

Report 3

1 Task1

1.1 Algorithm Description

本节主要是调用了 pandas 库，对数据进行一系列处理，包括读取数据、查看数据的前 10 行、查看数据的信息、删除缺失值、重置索引、删除 id 列、查看 diagnosis 列的值的个数、将 diagnosis 列的值转换为 0 和 1、查看 2-7 列的统计信息、查看不同 diagnosis 值的各组数据所有特征的变异系数，详情在代码文件中显示。

1.2 Results

	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	84300903	M	19.69	21.25	130.00	1203.0	
3	84348301	M	11.42	20.38	77.58	386.1	
4	84358402	M	20.29	14.34	135.10	1297.0	
5	843786	M	12.45	15.70	82.57	477.1	
6	844359	M	18.25	19.98	119.60	1040.0	
7	84458202	M	13.71	20.83	90.20	577.9	
8	844981	M	13.00	21.82	87.50	519.8	
9	84501001	M	12.46	24.04	83.97	475.9	

	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	\
0	0.11840	0.27760	0.30010	0.14710	
1	0.08474	0.07864	0.08690	0.07017	
2	0.10960	0.15990	0.19740	0.12790	
3	0.14250	0.28390	0.24140	0.10520	
4	0.10030	0.13280	0.19800	0.10430	
5	0.12780	0.17000	0.15780	0.08089	
6	0.09463	0.10900	0.11270	0.07400	
7	0.11890	0.16450	0.09366	0.05985	
8	0.12730	0.19320	0.18590	0.09353	
9	0.11860	0.23960	0.22730	0.08543	

	...	radius_worst	texture_worst	perimeter_worst	area_worst	\
...						
0		0.481809	0.154683		0.173924	
1		0.250869	0.232936		0.236436	

[2 rows x 30 columns]

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

限于篇幅，没将 Task1 的全部结果显示出来。

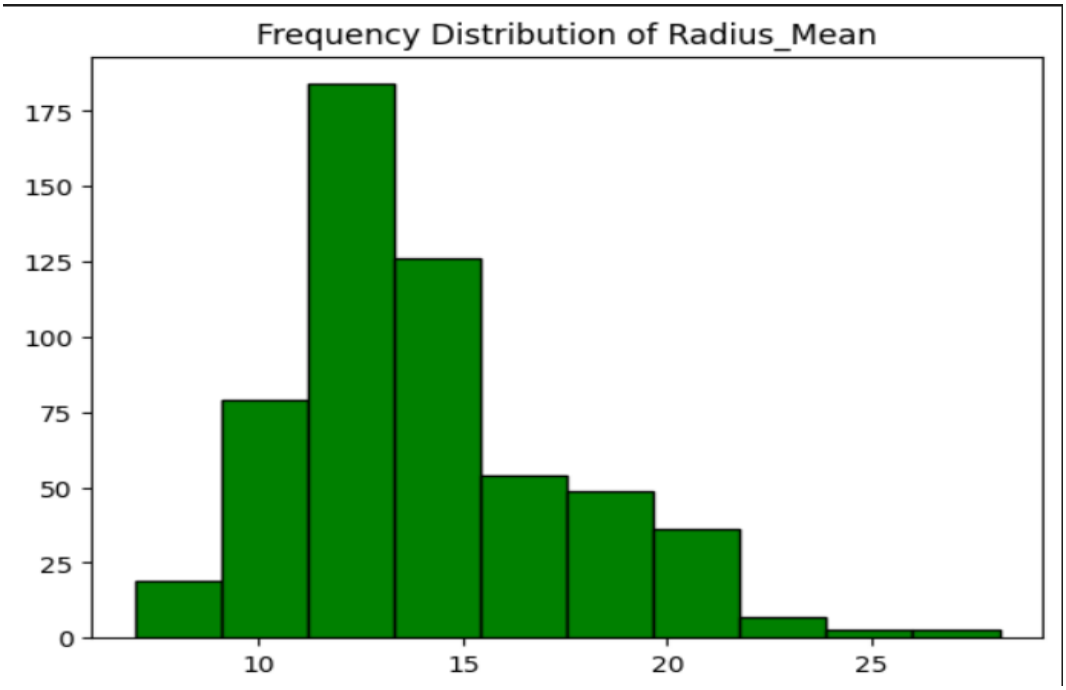
2 Task2

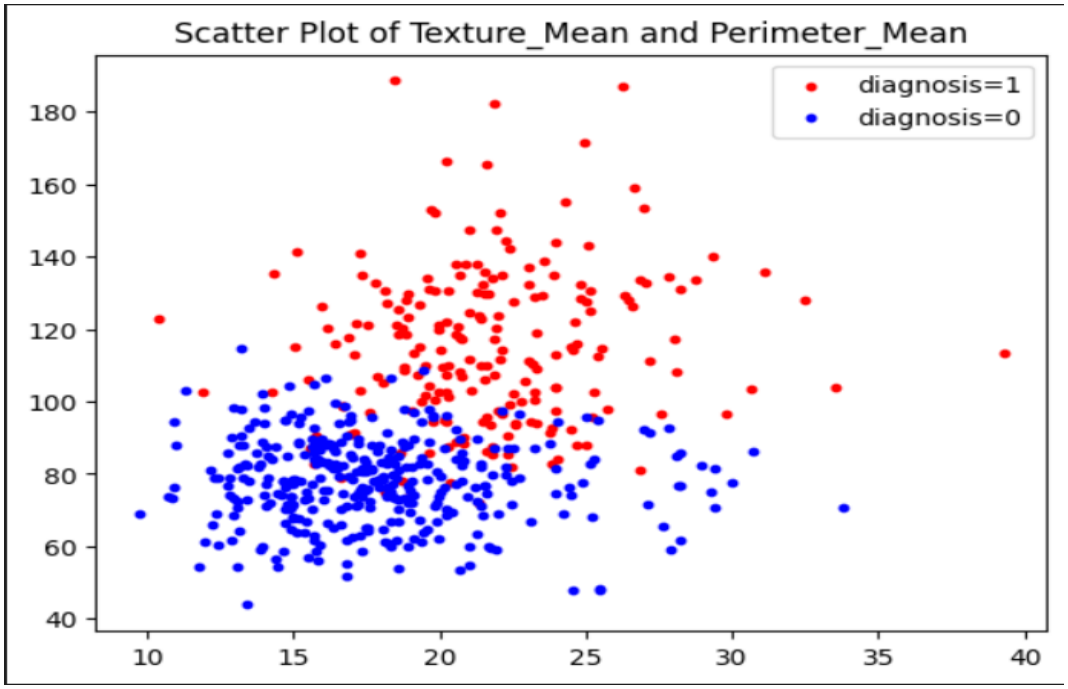
2.1 Algorithm Description

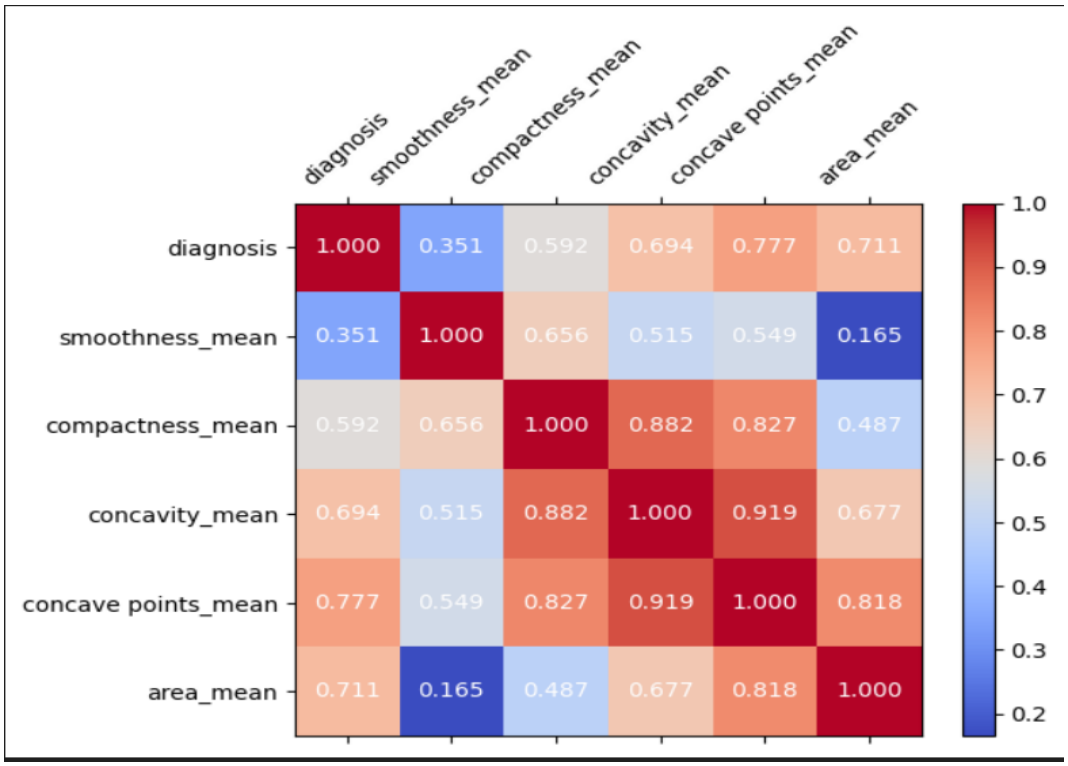
本节调用了 numpy 和 matplotlib 库，对数据进行一系列处理，如绘制频率分布直方图、绘制分布散点图、corr() 求 Pearson 相关系数矩阵、用 matshow(corr, cmap='coolwarm') 绘

制相应的相关系数热力矩阵图等，由于都是固定的操作，故不再详述代码细节。

2.2 Results







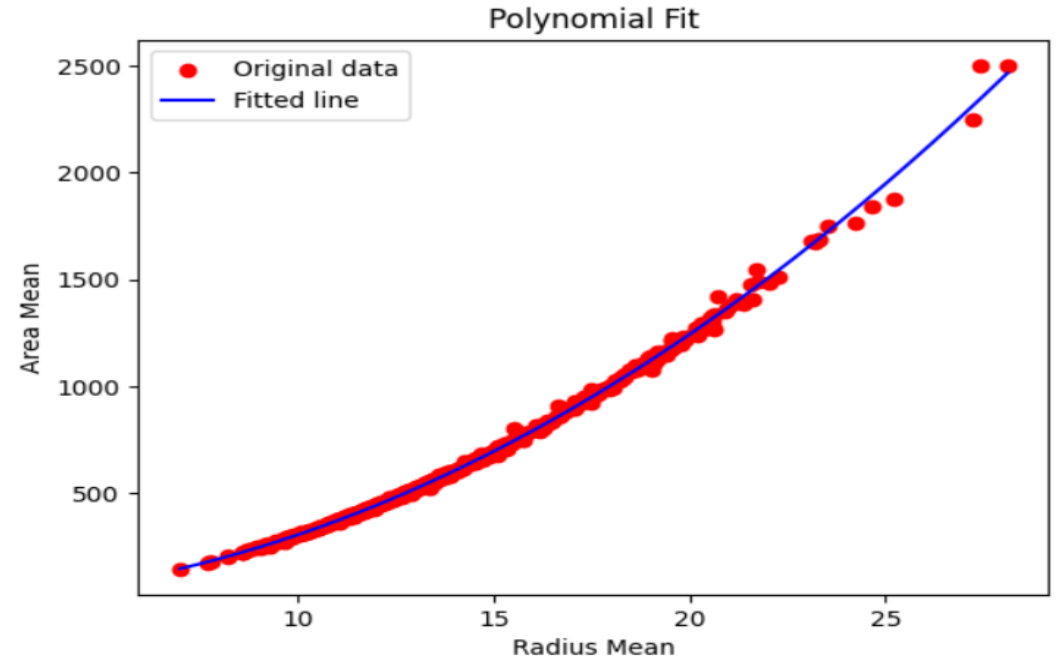
3 Task3

3.1 Algorithm Description

先用 np.column_stack() 将三个一维数组堆叠成一个二维数组，每个一维数组成为新二维数组的一列，进而代入公式求最小二乘估计，再使用 numpy.polyfit() 方法做二次多项式拟合，并绘制拟合曲线，可以发现数据适合二次函数拟合，而不适合线性拟合。

3.2 Results

```
Q1 Model parameters:  [ 3.14186228 -0.44260792 -4.70867951]
Q2 Model parameters:  [ 3.14186228 -0.44260792 -4.70867951]
omega - coeffs =  [-2.98872038e-13  7.63683561e-12 -4.56710225e-11]
```



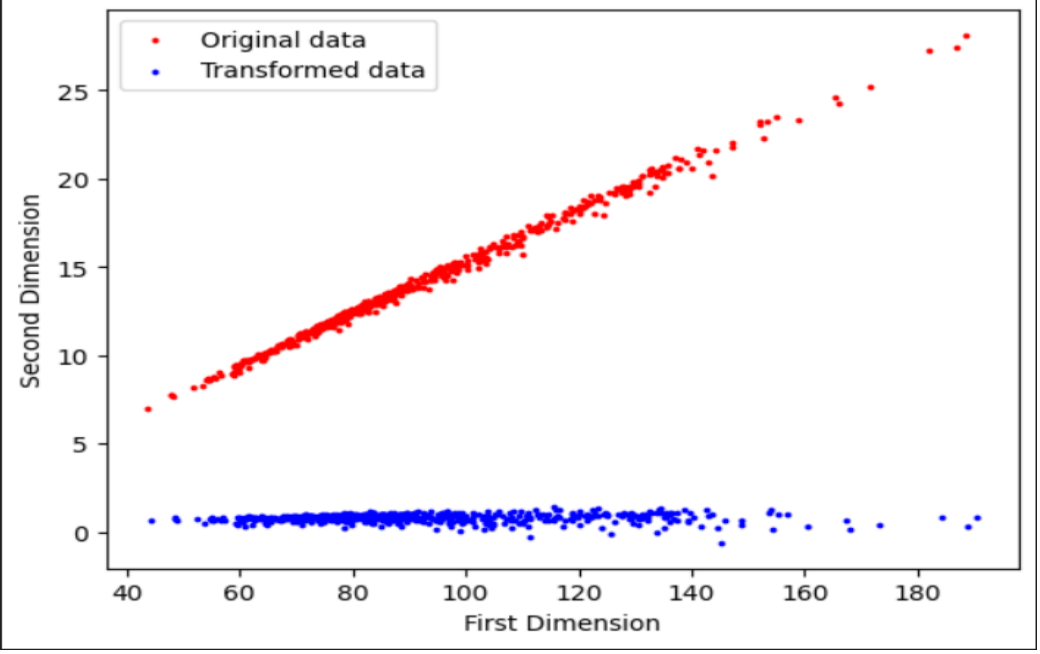
4 Task4

4.1 Algorithm Description

利用 numpy.cov() 和 numpy.linalg.eig() 求 X 的协方差矩阵 corX，进而求 corX 的特征值 eigV 与特征向量矩阵 eigMat，通过计算验证 eigMat 的正交性，接着对 X 进行 PCA，得到 Z，可以发现 Z 的数据的纵坐标基本在 0 附近，即 PCA 将 X 的数据映射到接近一维空间上。也能发现 cov(Z) 的对角元与 eigV 的值相等，且非对角元十分接近于 0。删除 Z 的一维数据，即可完成主成分分析降维过程。

4.2 Results

```
Eigenvalues:  [5.90457503e+02 5.21373729e-02]
Eigenvectors:
 [[ 0.98966947 -0.14336785]
 [ 0.14336785  0.98966947]]
Orthogonality check:  True
```




```
Covariance of Z:
 [[5.90457503e+02 1.97591893e-14]
 [1.97591893e-14 5.21373729e-02]]
Z_reduced :  [[124.11059852]
 [134.47614923]
 [131.47994405]
 [ 78.41581832]
 [136.61327906]
 [ 83.50193786]
 [120.98093186]
 [ 91.23375941]
 [ 88.45986066]
 [ 84.8889088 ]
 [103.93580752]
 [133.78059949]
 [104.90110445]
 [ 94.60150296]
 [ 97.81529636]
 [ 95.86592562]
 [109.29579312]
 [ 88.49769253]
```

限于篇幅，没将 Z_reduced 的全部结果显示出来。

5 Task5

5.1 Algorithm Description

本情景应该进行成组检验，因为两组数据相互独立，单侧检验原假设为 Mean of group1 = Mean of group2。因为 p 值为 5.469799049160595e-56 远小于 0.05，因此很有把握否认原假设，接受备择假设，即 Mean of group1 <= Mean of group2。

5.2 Results

```
Mean of group1:  0.1663615971830986
Mean of group2:  0.44671356097560977
t-statistic:  -19.01722049062515
p-value for one-tailed test:  5.469799049160595e-56
```