实验 8_任务 2_二分类

问题：

任务描述: 该数据为乳腺癌检查的医疗检测数据，每行对应一个案例(以逗号分隔每个数据)，其中第 一列为案例编号，第二列为诊断结果(M，malignant; B，benign)，其后十列为各项检查数据。分类 任务是通过后十项的检查数据来预测诊断结果。请完成下列工作:

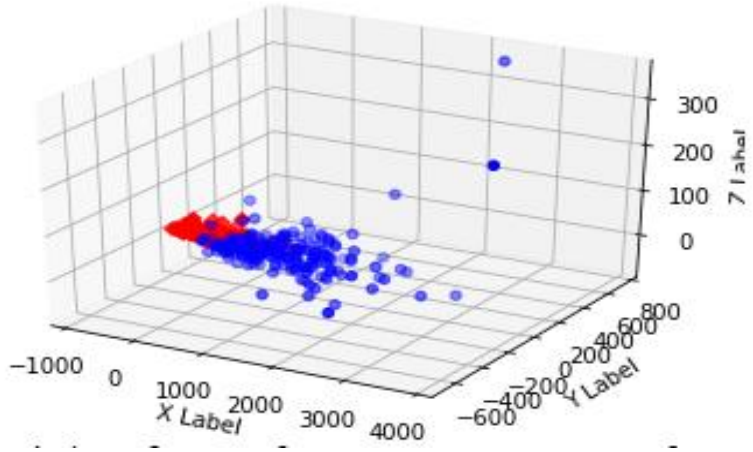
1. 对检查数据进行处理并使用降维方法（如 PCA）进行降维（2 维或 3 维）。通过可视化观察降维结果，并推测该数据是否适合进行分类学习。

2. 使用分类方法（如 logistic regression）对上述问题进行分类学习，并与你的推测结果进行对比和思考。（实验过程中请注意评价指标、训练误差、泛化误差、测试数据划分等内容，并记录在实验报告中。）

3. （附加题）尝试使用降维前后的数据表示分别进行分类，并比较分类的结果，思考降维对该分类

解答及实验过程记录：

1.数据可视化分析：



皮尔逊相关系数：

```
DF_matrix = DF.as_matrix(columns = None)
```

	Analyses	Factor_1	Factor_2	Factor3
Analyses	1.000000	7.327150e-01	-3.881190e-02	-9.636877e-02
Factor_1	0.732715	1.000000e+00	-2.186221e-17	-7.988046e-18
Factor_2	-0.038812	-2.186221e-17	1.000000e+00	-1.137545e-16
Factor3	-0.096369	-7.988046e-18	-1.137545e-16	1.000000e+00

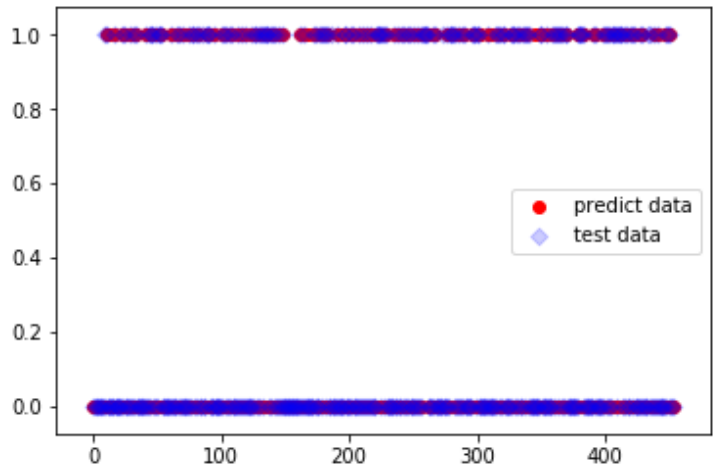
问题 1：

为了方便后面的数据处理，我首先把数据集中的诊断结果‘M’和‘B’分别用 1 和 0 代替。之后，利用 PCA 把数据降至 3 维，并且把诊断结果添加到每一条数据中，形成一个矩阵。把这个把诊断结果为 1 和 0 的数据分别用降维后的三个数据作为 x-y-z 坐标并绘制散点图，其中诊断结果为’M‘的用蓝点表示，诊断结果为’B‘的用红点表示，发现在 3 维坐标中，蓝点和红点分别聚集在一起，因此，我认为此数据集适合做分类学习。

问题 2：

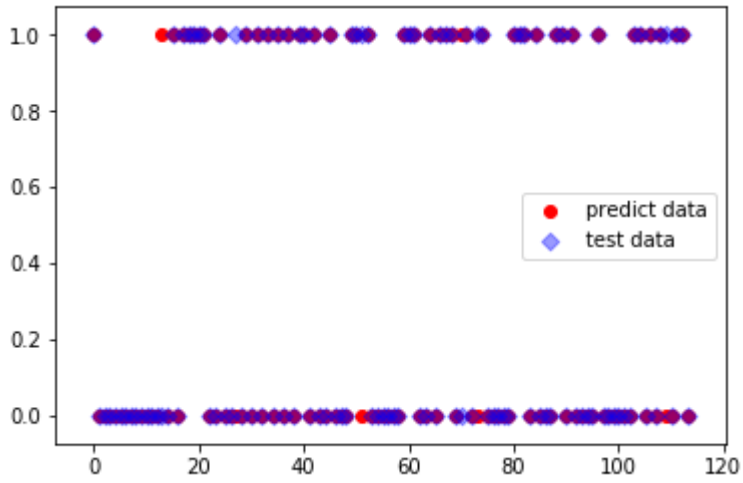
利用了 skilearn 工具对该数据进行了多元线性回归分析。

1. 划分数据集：80%作为训练集，20%作为测试集。
2. 评估方法：错误率，将预测结果与训练集比较，算出判断错误的样例占总样例的比重。
3. 训练误差：



Error Rate：0.07048458149779736

4.泛化误差：



Error Rate：0.05263157894736842

通过 matplotli 作图并且将测试点的透明度设为（0.3），可以发现大部分预测点与真实点重合，预测准确的比较高，与之前推测的相同，该数据适合分类学习。