# 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:

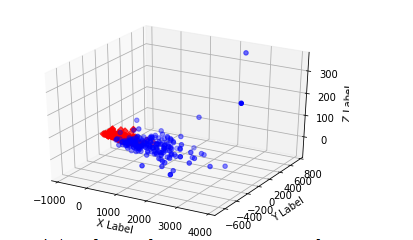
1. 对检查数据进⾏处理并使⽤降维⽅法（如PCA）进⾏降维（2维或3维）。通过可视化观察降维结果，并推测该数据是否适合进⾏分类学习。

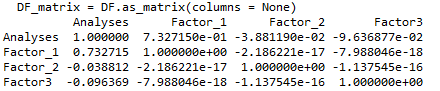
2. 使⽤分类⽅法（如logistic regression）对上述问题进⾏分类学习，并与你的推测结果进⾏对⽐和思考。（实验过程中请注意评价指标、训练误差、泛化误差、测试数据划分等内容，并记录在实验报告中。）

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：

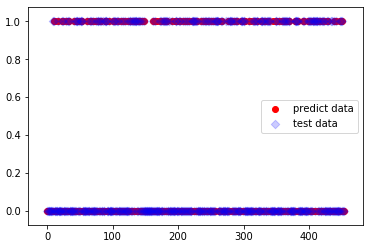
**问题1：**

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：**

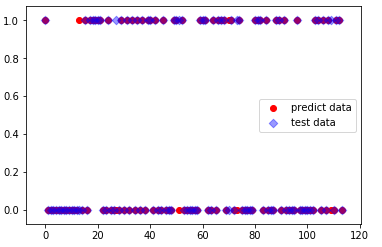
利用了skilearn工具对该数据进行了多元线性回归分析。

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。