 

SHANGHAI JIAO TONG UNIVERSITY

研究生课程作业

GRADUATE COURSE PROJECT REPORT

学生姓名: 季华伟

学生学号: 023034910026

任课教师: 陈海宝 副教授

学院(系)： 电子信息与电气工程学院 （微纳电子学系）

开课学期： 2024年 （春季）

# 问题描述

设想你是某门课程的教师，想通过学生平时和期末两次的评定成绩，来决定他们是否通过这门课程。现在你拥有之前学生的可以用于训练的样本集。对于每一个训练样本，你有他们平时和期末评分及最后是通过课程的结果（1为通过，0为不通过）。请利用逻辑回归或支撑向量机方法建立分类器。

首先我们先导入数据集，查看一下所给数据样式，我们发现数据集中共包含100个样本，每个样本包含两个特征，一个标签，对应于问题中的平时和期末成绩，以及最后是否通过结果。显而易见，我们的任务就是建立一个预测分类器，对所给数据进行拟合，找到一个能够正确分割所有样本的超平面。

我们考察了样本不平衡的因素，通过考察标签数据我们发现，正样本有60个，负样本有40个。我们发现直接调用sklearn的划分数据集的方法对数据集进行划分并不能取得最好的效果，因此我们构建了自己的数据集划分器，采用35个正样本，35个负样本作为我们的训练集，剩下的25个正样本和5个负样本作为测试集，最终我们使用degree为2的非线性Logistic Regression取得了100%的准确率！

我们分别使用Logistics Regression和SVM对该问题进行建模，给出了分类边界，并比较了两者的性能及差异。

# 方法描述

方法描述大部分来自sklearn，以及一些相关blog。

首先对于Logistic Regression来说，我们假设我们的标签，其中表示第个样本。给定第个输入后，我们预测的为：

基于上式，我们可以得到Logistic Regression模型的优化目标：

其中，是惩罚因子，用来平衡预测准确率和过拟合之间的关系。是一个element-wise的加权项，因为有时候不同的标签可能会有不同的重要性。是对参数进行正则化的一项，通常情况下，它是或者是。

其次对于SVM，其核心思想是找到一个距不同类别距离最大的超平面。它原问题的优化目标为：

转化为对偶问题为：

其中，为松弛变量，允许模型出现一定错误，提高泛化性能。对应的为支撑向量，即与超平面距离最近的那些点。

# 实验结果

首先在进行分类之前我们需要首先排除异常值和具有多重共线性的特征，但是经过检验，样本中没有异常值，并且两个特征不具有强线性关系。结果如下图所示。

屏幕上有字

描述已自动生成 图片包含 图表

描述已自动生成

（a）箱线图检测异常值 （b）相关系数图

图1 （a）中箱线图显示了不论x1，x2还是标签y，都没有异常值存在。（b）中我们可以看到不同特征之和自己相关，与其他变量无关。

我们首先做了Logistic Regression中线性的情况，包括使用平衡样本和非平衡样本。我们发现平衡训练样本之后，训练准确率提升了7个点，从而说明了样本不平衡的问题确实存在，并且对实验有影响。实验结果如下图所示：

电脑萤幕画面

描述已自动生成 电脑萤幕画面

描述已自动生成

（a）未平衡样本的拟合 （b）平衡样本的拟合

图2 （a）中我们使用了degree为1的Logistic Regression，但未平衡训练样本，分类准确率为0.867。（b）中我们同样使用了degree为1的Logistic Regression，平衡训练样本，分类准确率为0.933.

接着我们又使用了Logistic Regression中非线性的情况，degree设置为2，我们发现准确率达到了100%！如下图所示：

图片包含 图表

描述已自动生成

图3 使用了degree为2的Logistic Regression，平衡训练样本，分类准确率为1.0。

我们接着进行了SVM的实验，其中使用了各种不同的核函数，但是准确率都达不到1。实验结果如下图所示：

图片包含 图形用户界面

描述已自动生成

图4 使用了不同核函数的SVM，平衡训练样本，分类准确率达不到1.0。

# 结论

针对该分类问题，我们进行了不同方式的建模，包括Logistic Regression和SVM，另外我们还考虑了数据集中样本不平衡的问题，通过训练数据的样本平衡以及非线性拟合，我们在测试集上取得了100%的准确率。

# 参考文献：

[1] <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html>

[2] <https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression>

[3] <https://scikit-learn.org/stable/auto_examples/svm/plot_iris_svc.html>

[4] <https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/>

[5] Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods[J]. Advances in large margin classifiers, 1999, 10(3): 61-74.