# 项目开发报告

## 1.1问题分析

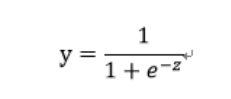
原题目为给定训练集 income.csv，要求根据每个人的属性值来判断此人年收入是否大于 50K。这是一个典型的二分类问题。我们将数据集划分为训练集和测试集，通过训练集，我们试图寻找57个特征与类别（是否大于50K）之间的联系，并用分类模型刻画这种联系，也就是得到logistic回归模型。然后再将logistic回归模型应用到测试集上，通过于测试集上的类别标记进行比较，测试模型的预测效果。对于logistic回归，其核心超参数为学习率alpha，我们还将测试不同alpha下误差曲线和准确率等。

## 1.2设计与分析

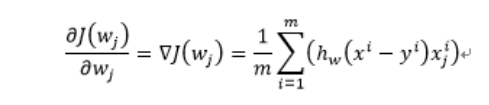
I、我们设计了类LogisticClassifer，其成员变量包括学习率alpha，回归系数theta，迭代结束阈值thresh等，其成员函数包括训练函数train()、预测函数predict()。

其中在train的时候，用到了了梯度下降方法迭代求解回归系数theta。实现了sigmod函数sigmod()，实现了梯度计算函数gradient()，实现分类误差函数cost()。

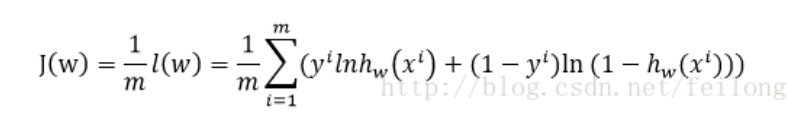
Sigmod函数sigmod()：



梯度计算函数cost()：



分类误差函数：



II、在评估时用到了sklearn的混淆矩阵。sklearn.metrics import classification\_report

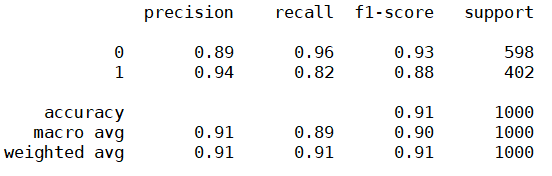
III、通过循环使用不同的学习率，做多次训练、评估、预测流程，得到不同的误差曲线和准确率。

IV 与sklearn提供的logistic回归进行对比。

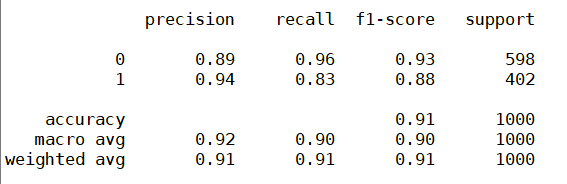
## 1.4结果分析（包括实际运行结果截图，性能结果图、表等）

I、模型准确率评估混淆矩阵。

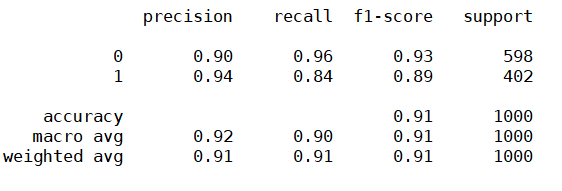
Alpha=0.1



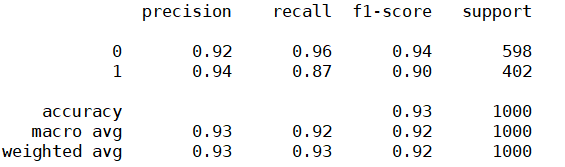
Alpha=0.3



Appha=0.5

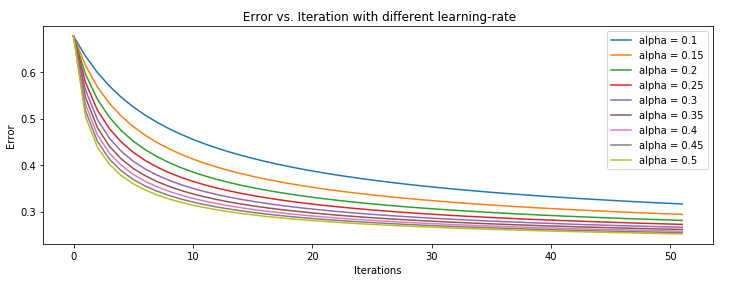


Sklearn logistic分类器



II、误差曲线

迭代次数为50



迭代次数为200

