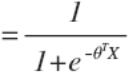
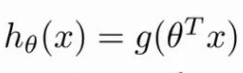
1.模型的建立

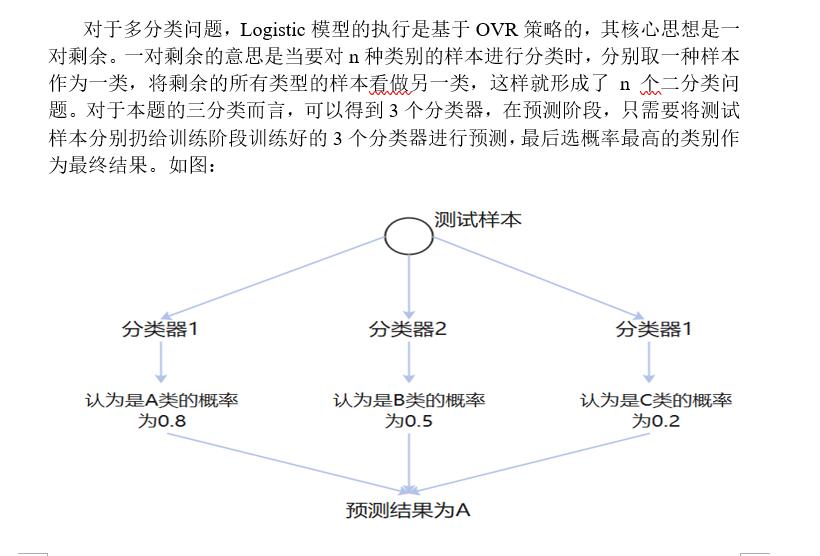
（1）分类函数

本题中鸢尾花有三类，且各个类别不存在互斥现象，对于多分类的有监督学习，可以使用Logistic，为解决LPM内生性，将自变量线性组合作为输入放到一个分布函数中，这里使用sigmad函数，具体分类模型如下



其中theta表示自变量的系数

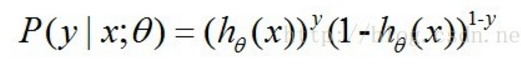
（2）推广至多分类



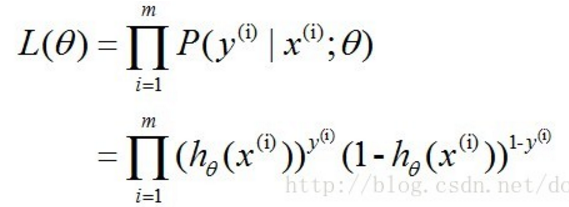
2.模型的求解

（1）求解系数

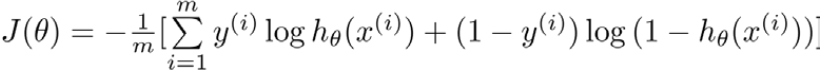
为求解分类函数中的系数，这里使用最大似然估计引入代价函数J



对应的似然函数为：

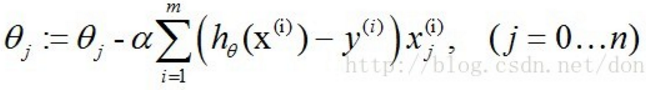


最大似然估计就是要求得使*l(θ*)取最大值时的**θ**，将似然函数取对数后求解的结果乘以-1/M，这样即可找到一个theat使得代价函数最小。



（2）梯度下降

这里插一个图



其中alpha是学习速率，该公式表示theat的更新过程

（3）分类结果

将全部数据作为训练数据，放入logistic回归模型进行训练，使用梯度下降法对theat进行求解，得出theat = ？？这里画一个分类图

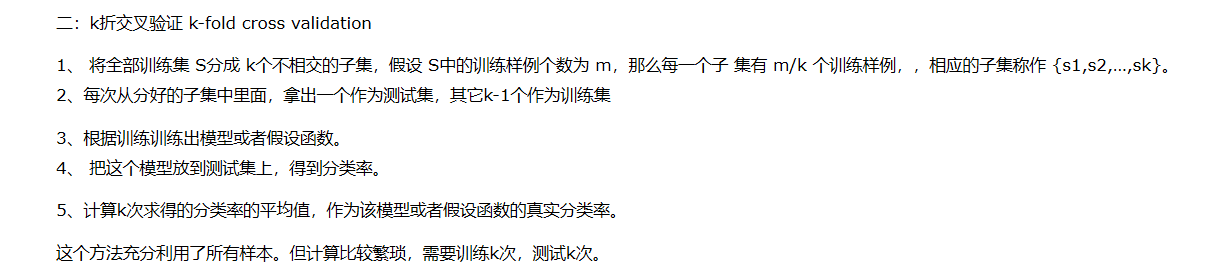
将所要分类的花的各项数据代入模型中，最终结果如下表所示

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

3.模型的评价

（1）K折交叉验证

有监督的多分类问题可以用probit、逻辑回归或SVM，只是需要多个二分类来组成多分类，也可以使用softmax多分类器。使用K折交叉验证不同方法的平均预测正确率。



由于数据量比较小，将样本按三比七划分测试集和训练集，具体结果如下表所示

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

通过K折交叉验证给出的结果，发现使用Logistic回归在测试集上表现更好，它是一种具体的、行之有效的、实用价值很高的分类方法，因此本题选用logistic。

（2）缺点

对模型中自变量多重共线性较为敏感，例如两个高度相关自变量同时放入模型，可能导致较弱的一个自变量回归符号不符合预期，符号被扭转。​需要利用因子分析或者变量聚类分析等手段来选择代表性的自变量，以减少候选变量之间的相关性；

预测结果呈“S”型，因此从log(odds)向概率转化的过程是非线性的，在两端随着​log(odds)值的变化，概率变化很小，边际值太小，slope太小，而中间概率的变化很大，很敏感。 导致很多区间的变量变化对目标概率的影响没有区分度，无法确定阀值。

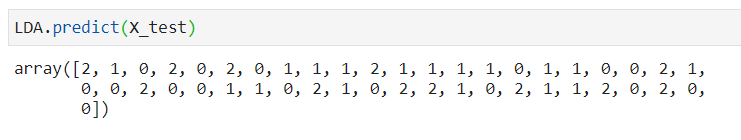
模型的建立与求解

Fisher判别分析模型建立

LDA是一种经典的线性判别方法，又称Fisher判别分析。该方法思想是，设法将样例投影到一维的直线上，使得同类样例的投影点尽可能接近和密集，异类投影点尽可能远离。

LDA模型的求解

将样本分为训练集和测试集，利用LDA分类器在训练集上进行拟合，之后在测试集上进行预测，预测结果如下：



利用拟合后的LDA分类器可以在测试集上得出预测精度，

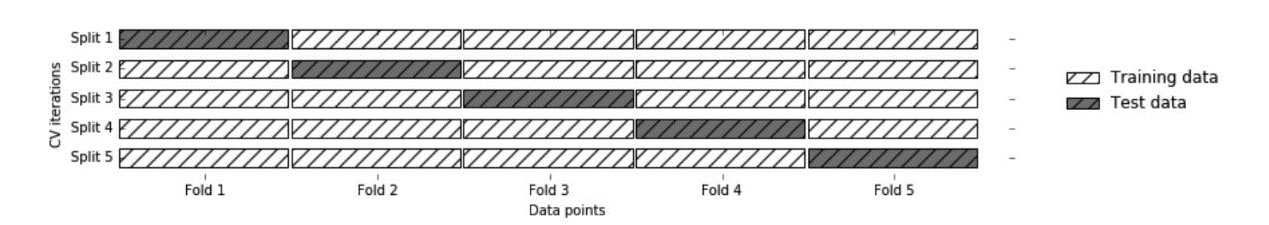


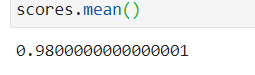
可以看出LDA分类器在鸢尾花三分类问题中有着出色的表现

模型的改进

叙述改为流程图

为了减小模型的随机性而带来的误差，采取交叉验证的方式对数据进行训练，可以使模型更加稳定、全面。最常用的交叉验证是k折交叉验证，k通常取5或10。在执行 5 折交叉验证时，首先将数据划分为（大致）相等的 5 部分，每一部分叫作折（fold）。接下来训练一系列模型。使用第 1 折作为测试集、其他折（2~5）作为训练集来训练第一个模型。利用 2~5 折中的数据来构建模型，然后在 1 折上评估精度。之后构建另一个模型，这次使用 2 折作为测试集， 1、 3、 4、 5 折中的数据作为训练集。利用 3、 4、 5 折作为测试集继续重复这一过程。对于将数据划分为训练集和测试集的这 5 次划分，每一次都要计算精度。最后我们得到了 5 个精度值。整个过程如图 所示



最终的得分为，可以看到精度有所提升