Iris 数据集上的对数几率模型实现

实验报告

姓名: 李彪

班级: 计算机 1602

学号: 1610650213

1 问题描述

使用对数几率回归模型, 对 Iris 数据集进行分类并验证模型的效能.

(1). 将数据集的 50%作为训练集, 50%作为测试集, 检验模型在测试集上的分类正确率

(2). 将数据集的 70%作为训练集, 30%作为测试集, 检验模型在测试集上的分类正确率

(3). 将数据集的 90%作为训练集, 10%作为测试集, 检验模型在测试集上的分类正确率

2 数据集描述

Iris 数据集是常用的分类实验数据集, 由 Fisher, 1936 收集整理. Iris 也称鸢尾花卉数据集, 是一类多重变量分析的数据集. 数据集包含 150 个样本, 分为 Setosa(山鸢尾)、Versicolor(杂色鸢尾), 以及 Virginica(维吉尼亚鸢尾)三个种类, 每类 50 个样本, 每个样本包含 4 个属性: Sepal Length(花萼长度), Sepal Width(花萼宽度), Petal Length(花瓣长度), Petal Width(花瓣宽度).

该数据集的前五个样本如下:

表 1 Iris 数据集前五个样本示例

	sepal_length	*sepal_width*	*petal_length*	*petal_width*	*species*
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

3 数据处理及分析

初步处理 Iris 数据集的各个属性, 得到其均值、方差、最值、分位点等信息的数据如下:

表 2　Iris 数据集中的一些信息

	sepal_length	*sepal_width*	*petal_length*	*petal_width*
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

从中可以看出该数据集的数据相对比较完整，没有缺省．下图可以从可视化的角度观察样本数据在各个属性上的分布:

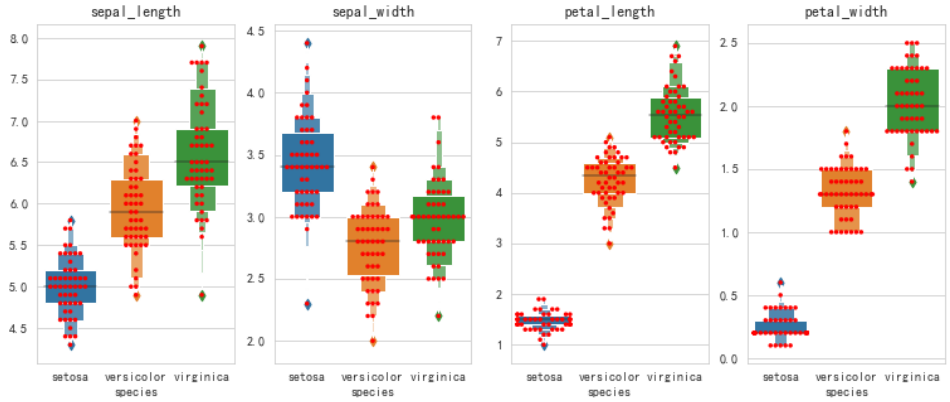

图 1　各属性上的箱线图

考虑到该数据集中的每个样本有 4 个特征，难以直接观察分布情况．这里将任意两个属性单独作为横、纵坐标，绘制多变量图来观察其分布特征，如下图所示:

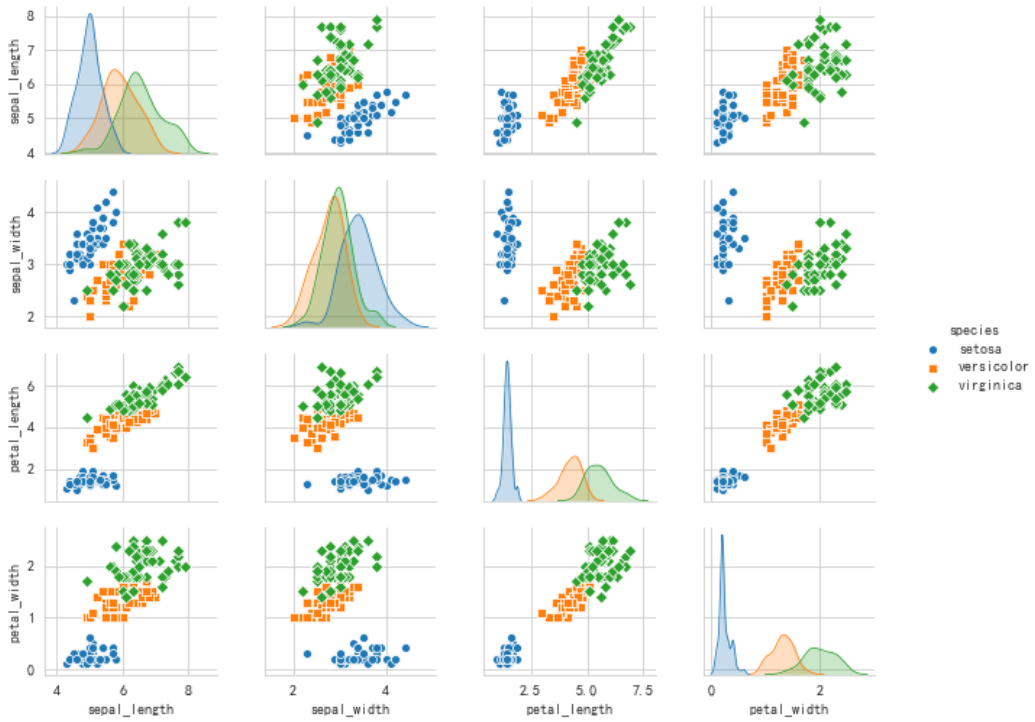

图 2　多变量图

从上图中可以看出：在 Petal Length、Petal Width 两个属性上 Setosa 类型的样本与其他两种的差异较大，易于区分．而就整体而言 Versicolor 和 Virginica 比较难以区分．

4　模型的建立与求解

模型的建立与求解方法均参考周老师的《机器学习》，这里不再赘述．

5　实验结果分析

➢　二分类

将 species 属性进行编码，属于 Setosa 类型的样本标记为 1，否则标记为 0，则问题转化为一个二分类问题．得到预测精度结果如下：

表 3　二分类的预测精度结果

测试集所占比例	训练集预测精度	测试集预测精度
10%	1.000	1.000
30%	1.000	1.000
50%	1.000	1.000

由上表可知：不管在什么样的情况下，训练集和测试集的精度都达到了 100%，这代表在数据标签只有两类的基础上，模型能很好地进行分类．该模型的

损失函数变化曲线如下:

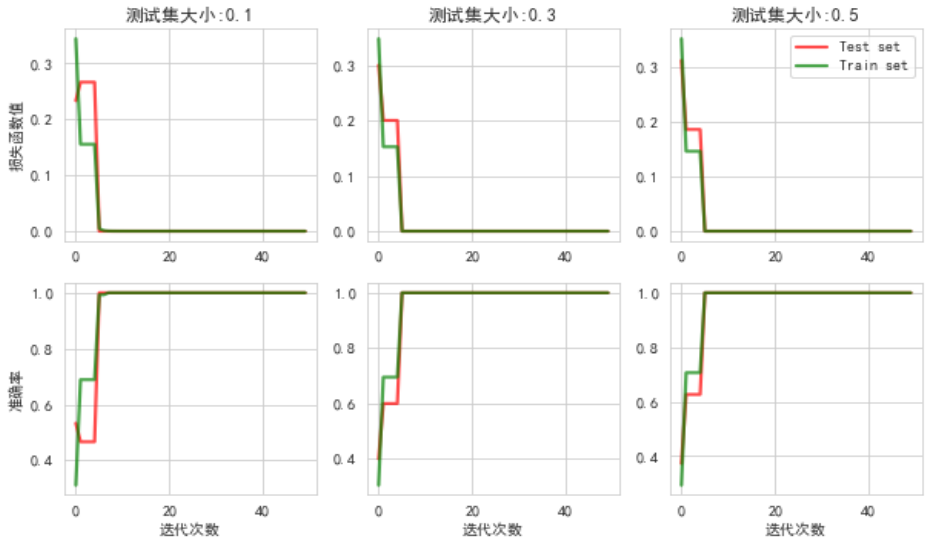

图 3　二分类的损失函数和预测精度变化曲线

可以看到, 在二分类的情况下收敛速度还是相对较快的, 一般在 10 次左右即可达到 100%的精度. 模型的参数如下:

表 4　二分类模型的超参数

选项	数值
学习率	0.100
衰减因子	0.9
迭代次数	50
随机数种子	1234

- **多分类**

将数据集中的 species 属性重新进行编码, 采用 one-hot 编码方式, 并使模型输出三个预测值. 然后取最大值作为预测结果. 则问题转化为一个多分类问题. 得到在不同比例大小的测试集下的预测精度结果如下:

表 5　多分类的预测精度结果

测试集所占比例	训练集预测精度	测试集训练精度
10%	0.963	1.000
30%	0.971	0.978
50%	0.960	0.987

在多分类的情况下模型在测试集和训练集上的精度都有所下降. 回顾之前的**图 2　多变量图**, 不难知道: 这是由于 Versicolor 和 Virginica 两类样本难以区分导致的. 损失函数与模型精度变化曲线如下:

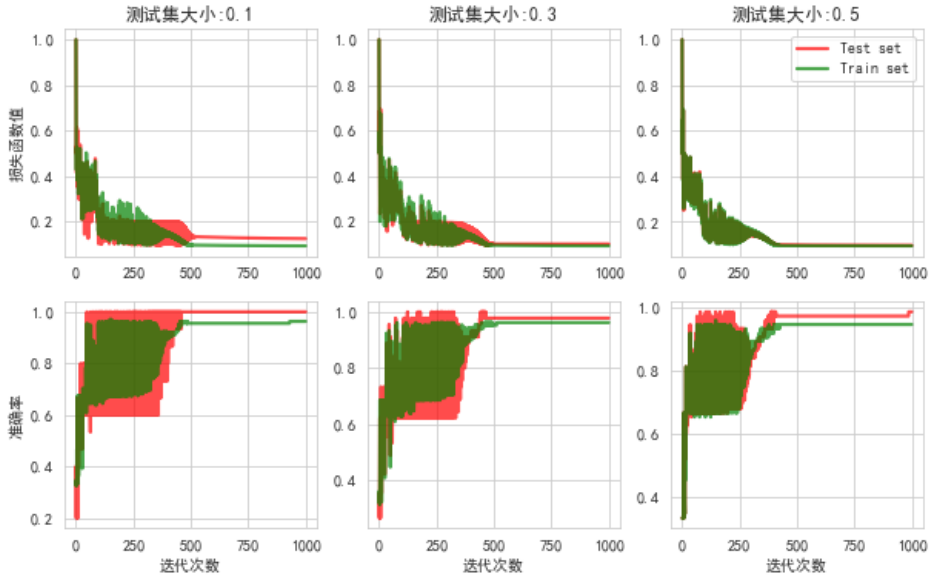

图 4　多分类的损失函数和预测精度变化曲线

由上图可知, 当进行多分类时模型收敛速度变慢, 一般需要 500 个周期以上模型才能到达一个比较稳定的值. 模型的参数如下:

表 6　多分类模型的超参数

选项	数值
学习率	0.100
衰减因子	0.997
迭代次数	1000
随机数种子	1234

6　模型可视化

既然有了模型何不做一下模型的可视化? 由于本数据集的属性较多, 并不能直接进行可视化, 所以考虑选取其中某两个属性进行可视化. 首先需要找到进行可视化的属性, 通过 xgboost 中的方法不难得到特征的重要性:

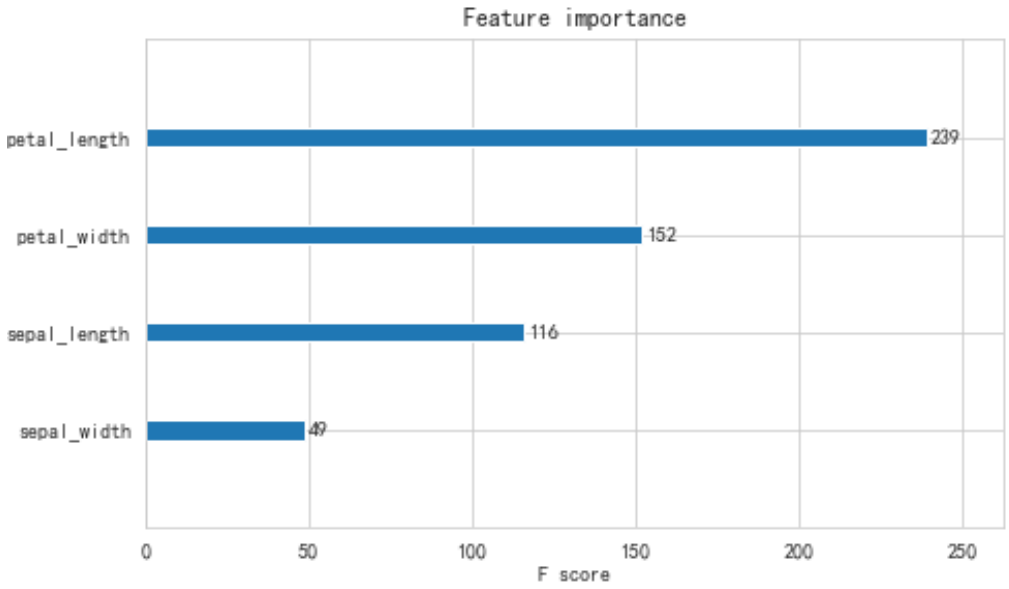

图 5　各属性的重要性图

即 petal length、petal width 属性相对比较重要，进一步进行分析，这里省略复杂的分析过程，得到选取特征个数与模型精度的关系如下:

表 7　各属性与模型精度关系表

选取的特征	模型精度
petal_length	0.933
petal_length、petal_width	0.956
petal_length、petal_width、sepal_length	0.956
petal_length、petal_width、sepal_length、sepal_width	0.956

即模型精度与 petal length、petal width 属性的相关性较大，而 sepal length、sepal width 属性对本模型的影响极小几乎可以忽略. 下一步: 对种类重新进行编码，这里省略复杂的分析过程，最终的编码方式为:

表 8　各种类编码方式表

种类	编码
Setosa	00
Versicolour	10
Virginica	11

至此，对模型进行训练，并将数据绘制处理．以测试集大小为 0.3 时为例，最终得到可视化的模型如下:

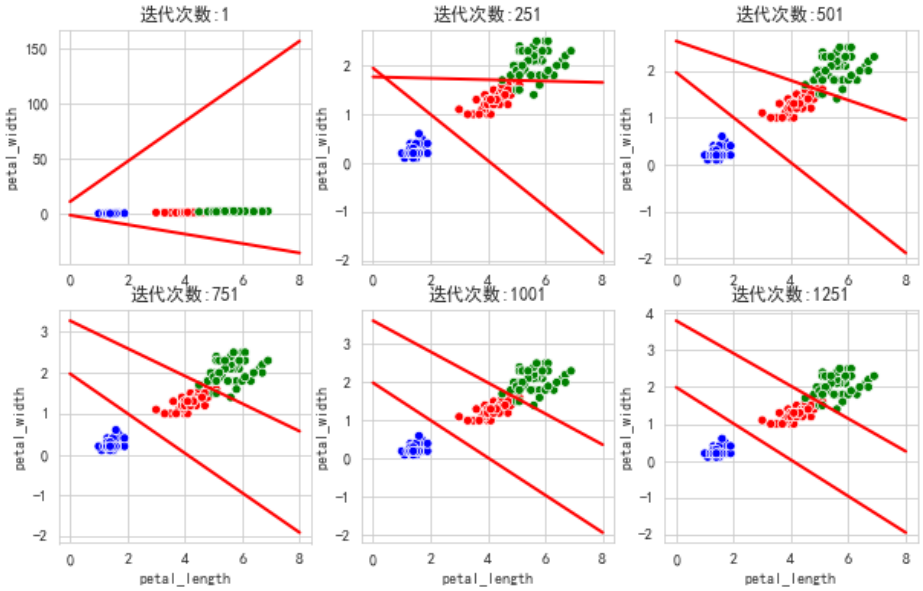

图 6　各迭代次数下的模型可视化

上图中描述了在不同迭代次数下的模型训练情况，其中横坐标为 petal length，纵坐标为 petal width．可以看到: 随着迭代次数的增加，图中的两条线越来越趋向于样本的分割面上(具体动图见附件).

重编码后的数据在模型训练过程中的损失函数和精度变化如下:

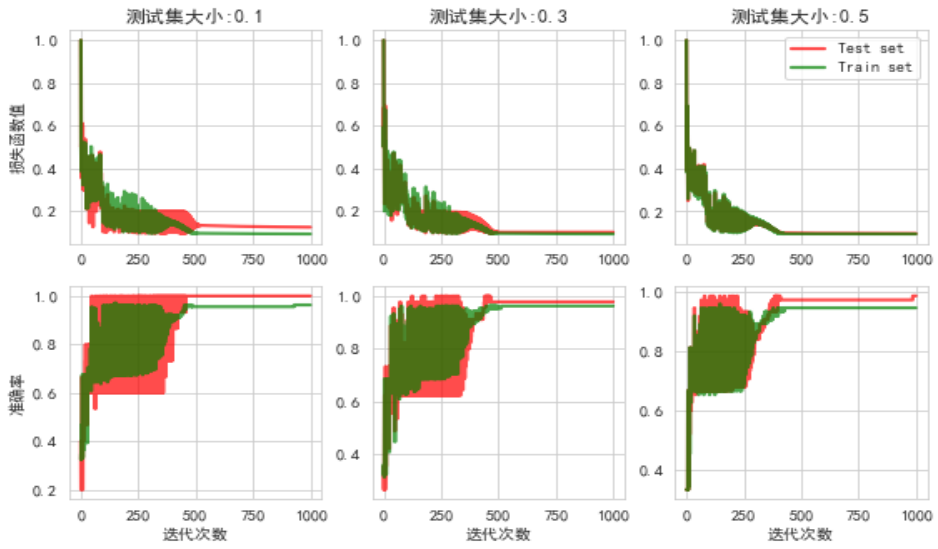

图 7　重编码后的模型的损失函数和预测精度变化曲线

由上图可以看出，仅选用两个属性进行学习的最终模型精度依然很高，这

说明特征的选择比较成功. 同时根据可视化的绘图, 也使模型更加具有可解释性.

7 实验总结

本次实验训练了一个对数几率模型, 发现在二分类与多分类的情况下模型的表现有所差异.

而且本文还依据属性的重要程度对训练数据进行了降维使模型更加简洁, 并做了模型可视化使模型更加具有可解释性.