使用Python实现对数几率回归模型

天津工业大学 余元强 计算机1601

## 问题描述

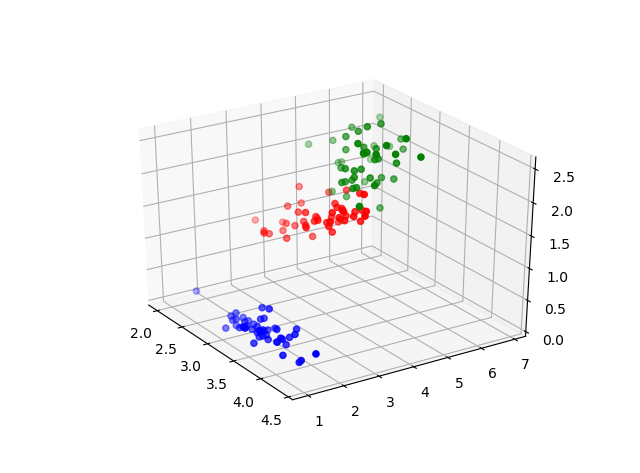
该实验要求编程实现对数几率回归模型，并对Iris数据集进行分类以验证模型的效能。另外，题目要求我们完成以下任务：

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。

## 数据集描述

Iris数据集包含150个数据集，分为3类，每类50个数据，每个数据包含4个属性。其中3个种类分别为Iris Setosa（山鸢尾）、Iris Versicolour（杂色鸢尾），以及Iris Virginica（维吉尼亚鸢尾），4个属性分别为Sepal.Length（花萼长度），Sepal.Width（花萼宽度）， Petal.Length（花瓣长度），Petal.Width（花瓣宽度），单位均为厘米。

抽取其中三个特征绘制三维散点图如下所示。



其中前五行的特征如下表所示。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 索引 | SL | SW | PL | PW |
| 0 | 5.1 | 3.5 | 1.4 | 0.2 |
| 1 | 4.9 | 3.0 | 1.4 | 0.2 |
| 2 | 4.7 | 3.2 | 1.3 | 0.2 |
| 3 | 4.6 | 3.1 | 1.5 | 0.2 |
| 4 | 5.0 | 3.6 | 1.4 | 0.2 |

## 问题的建模与求解

### 3.1 激活函数定义

激活函数（Activation functions）对于人工神经网络模型去学习、理解非常复杂和非线性的函数来说具有十分重要的作用。对于此Logistic回归问题，本文采用的激活函数为Sigmoid函数，即对于任何输入向量X，其属于正例的概率为



为了求解模型中的两个参数，我们接下来定义损失函数。

## 3.2 损失函数定义

针对Logistic模型，某样本属于类别y的概率可以表示为



而参数W与b可采用极大似然法对其进行估计。假设训练数据集有m个训练样本，则其似然函数可以表示成

为了计算上的方便，本文采用Log似然函数，且希望损失函数值越小越好，故还需对似然函数进行取反运算，则损失函数为

## 3.3 梯度下降法

为了求出损失函数的最小值，这里采用梯度下降法进行求解，梯度表达式经过推导可得

故由梯度下降法，得到如下的更新公式：

## 3.4 模型的求解

由于山鸢尾和杂色鸢尾的特征区分过于明显，导致较少的训练集依然可以由很高的正确率，如下图所示。

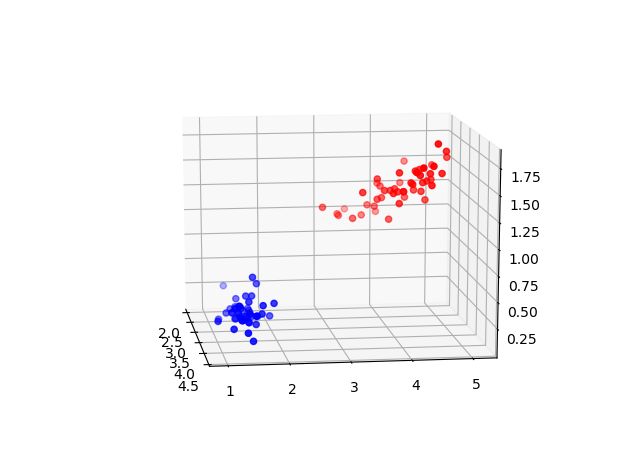


图 两例的散点图

故为了测试模型的稳健性，这里对杂色鸢尾和维吉尼亚鸢尾的花进行分类。本文将4个特征均拿来作为训练数据，散点图如下所示，其中红色表示杂色鸢尾的三个特征，蓝色表示维吉尼亚鸢尾的三个特征。

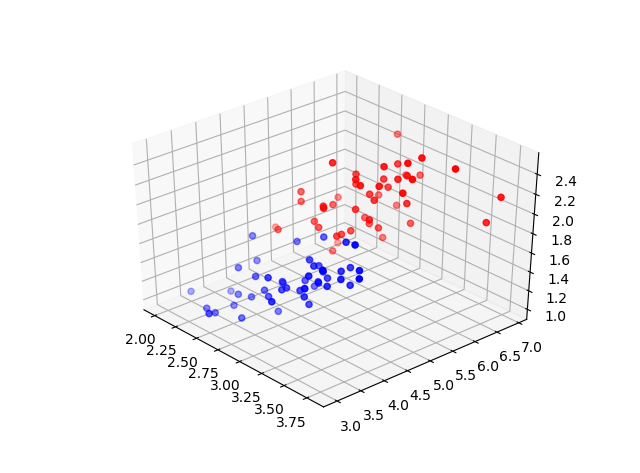


图 两例的散点图

将数据集的 50%作为训练集，50%作为测试集时，学习率为0.001，最大迭代次数为1000，训练后正确率为66%，模型损失函数变化曲线如图1所示。

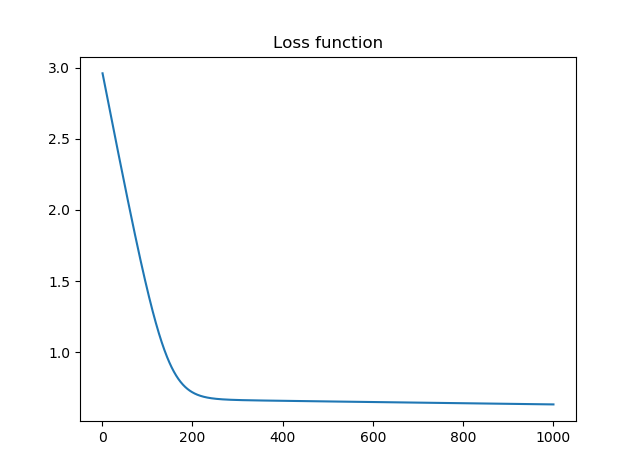


图1 50%训练集时损失函数曲线图

将数据集的 70%作为训练集，30%作为测试集时，学习率为0.005，最大迭代次数为1000，训练后正确率为93.3%，模型损失函数变化曲线如图2所示。

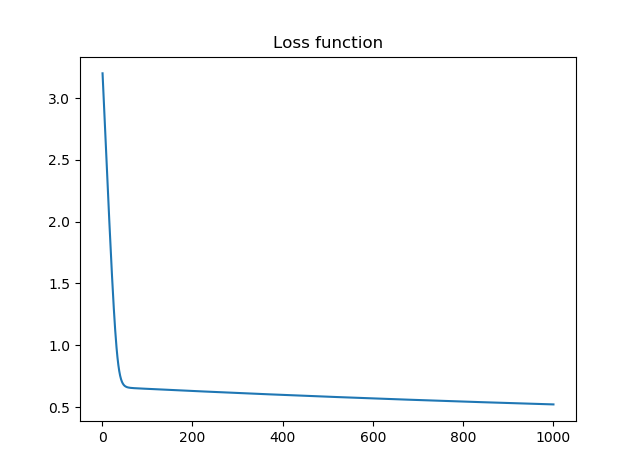


图2 70%训练集时损失函数曲线图

将数据集的 90%作为训练集，10%作为测试集时，学习率为0.005，最大迭代次数为1000，训练后正确率为100%，模型损失函数变化曲线如图3所示。

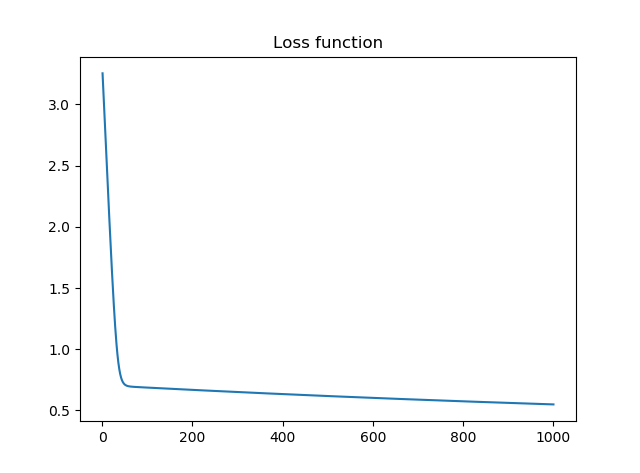


图3 90%训练集时损失函数曲线图

## 实验结果分析

将实验结果总结为下表。

|  |  |  |
| --- | --- | --- |
| 训练集占比 | 测试集占比 | 分类正确率 |
| 50% | 50% | **66%** |
| 70% | 30% | **93.3%** |
| 90% | 10% | **100%** |

由实验结果可以直观看到，当训练集在所有样本中所占比例越来越大时，其模型训练出来的准确率越来越高，且占比为90%时准确率为1，该模型灵敏度与鲁棒性较高，具有一定的推广能力。

本文仅能区分第二类和第三类，对于三分类的问题，我们也可以采用Andrew NG提出的方法构造3个二分类器，如下所示。

对于第1个二分类器，将山鸢尾的标签记为1，杂色鸢尾和维吉尼亚鸢尾记为0，进行训练；

对于第2个二分类器，将杂色鸢尾的标签记为1，山鸢尾和维吉尼亚鸢尾记为0，进行训练；

对于第3个二分类器，将维吉尼亚鸢尾的标签记为1，山鸢尾和杂色鸢尾记为0，进行训练。

在对新的样本进行预测时，将该样本所有特征放入三个Logistic分类器模型中，得出3个输出值，比较将其归为1的概率中哪个分类器输出值大，则将其标记为对应分类器的鸢尾花。事实证明，该方法的确有一定的效果，例如，将第一个样本代入3个分类器中进行预测，标定为正例的概率分别为[0.56465853, 0.34038794, 0.39631655]，故该样本属于第一类。