班级：计算机1603

姓名：张家源

学号：1611640321

# Python 逻辑回归（Logistic Regression）

# 分析鸢尾花数据

## 问题描述：

本次实验要解决的问题是通过python实现对数几率回归模型，使用鸢尾花数据对模型进行训练，进而实现对鸢尾花实现分类的功能。

这个数据集来源于科学家在一岛上找到一种花的三种不同亚类别，但是这三个种类并不是很好分辩，从花萼长度，花萼宽度，花瓣长度，花瓣宽度这四个角度测量不同的种类用于定量分析。

鸢尾花数据集共有四个特征，查阅相关资料可知，花瓣的长度和宽度像个属性的相关性较高，所以本次实验选择其中的花萼长度和宽度作为特征进行训练。

## 数据集描述：

### 数据集概述：

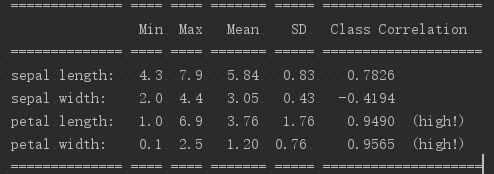
(1)鸢尾花数据集共有150个数据，其中1-50数据为“Iris-Setosa”

51-100数据为“Iris-Versicolour”，101-150数据为“Iris-Virginica”

(2)本次使用的数据集为.csv格式的文件，其中第一列数据为编号，第二列数据为花瓣的长度，第三列数据为花瓣的宽度，第四列数据为花萼的长度，第五列数据为花萼的宽度，以上数据的单位为厘米，第六列数据为花的种类。

### 数据集分析：

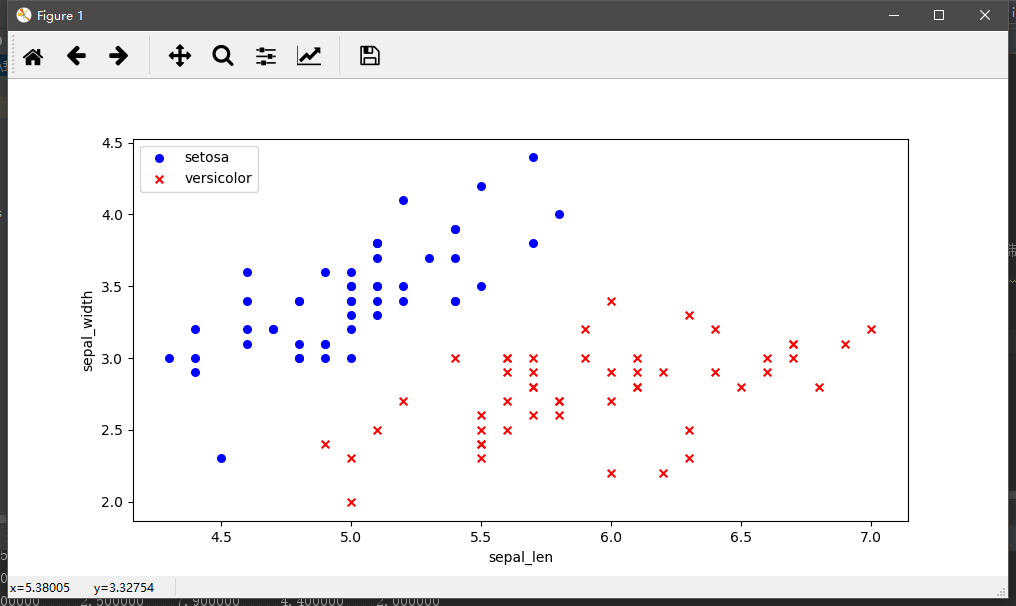
通过使用print(iris['DESCR']) 来打印数据集的描述信息：



由上图可知，petal length和petal width两个属性的相关性较高，故选择这两个属性作为判别属性。

此次实验选择前两种花作为分类目标。Iris-Setosa，Iris-Versicolour，Virginica。

### 打印初始数据分布图：



## 实验过程及结果：

### 数据读取及处理：

1. 用pandas读取数据，并将数据的最后一列字符串替换成0，1和2，分别对应三种花的种类：Iris-Setosa，Iris-Versicolour和Virginica



1. 在第三列插入数字1，与参数向量的维度保持一致。



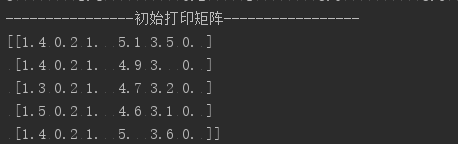
(3) 构造一个矩阵



（4）保留数据的前一百列:

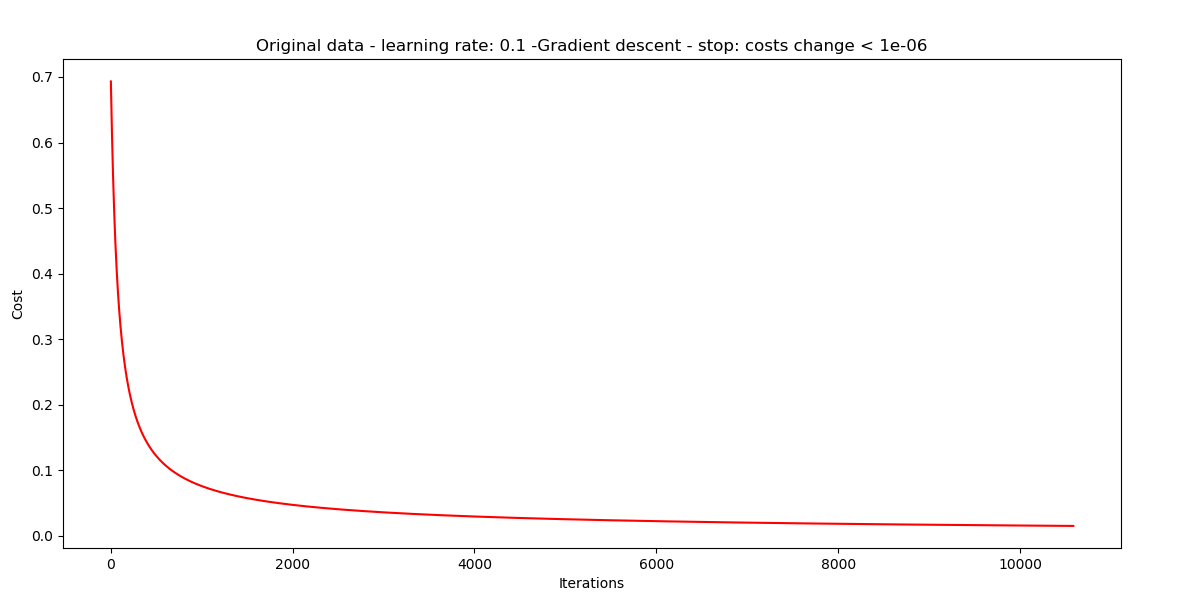


(5) 处理完的数据为一个100行6列的矩阵。



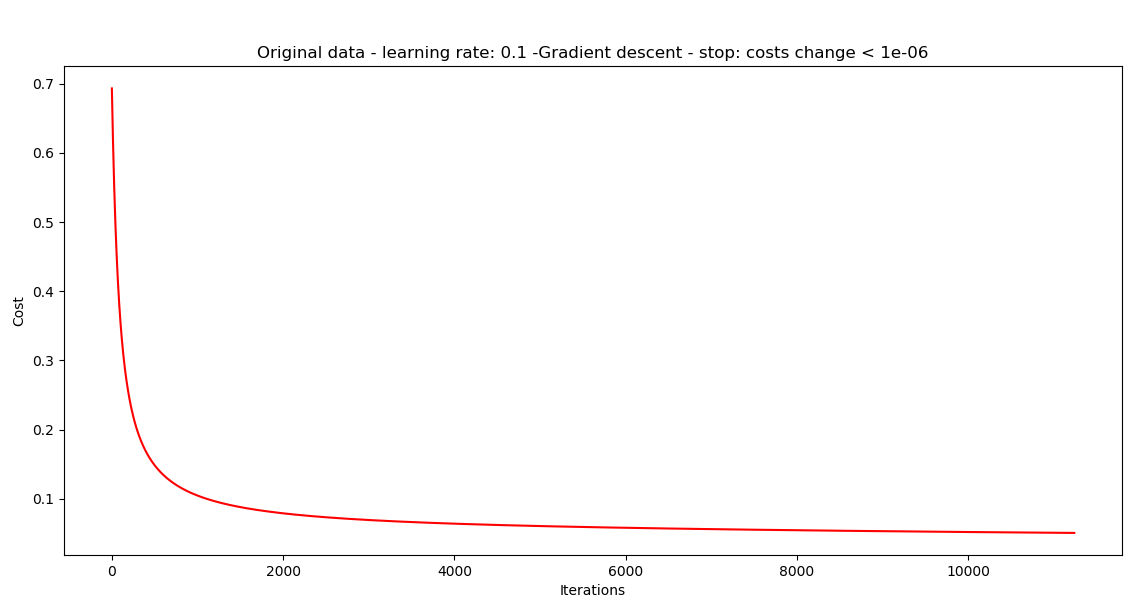
### 实验过程及结果：

停止策略：当损失值误差小于1e-6时停止

（1）50%的数据作为训练集，50%的数据作为测试集

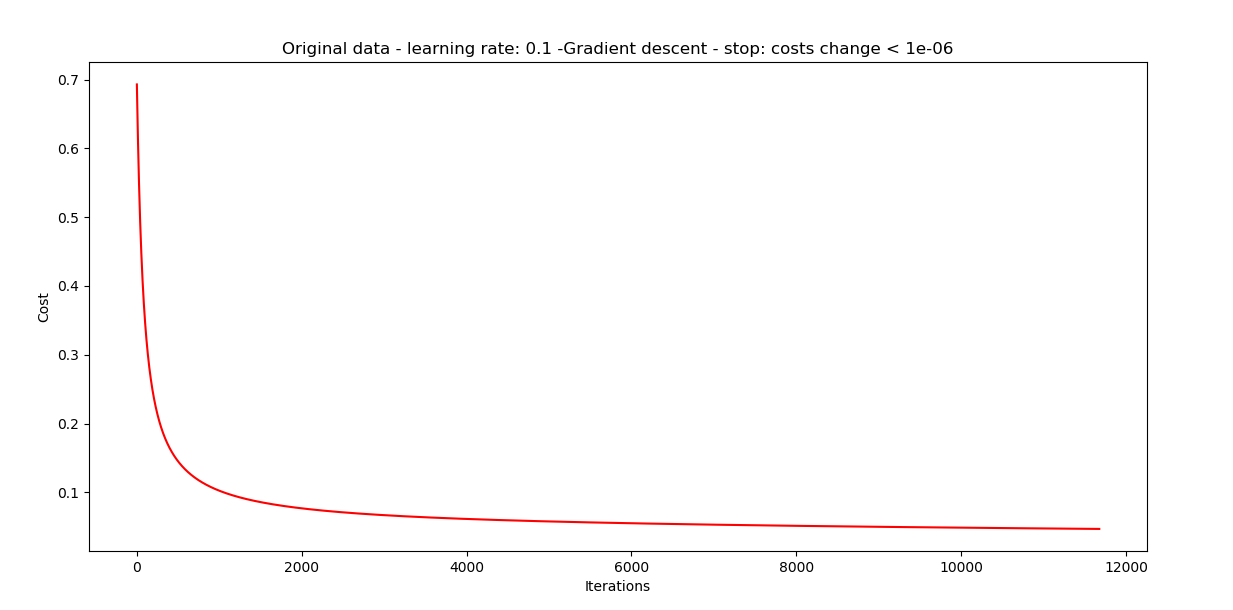
损失值为：0.015355358595476369，正确率：98.0%

Duration: 4.09s，迭代次数: 10586次

（2）70%数据作为训练集，30%数据作为测试集

损失值：0.050552654867945856 ，正确率 = 96.7%

Duration: 5.16s ， 迭代次数：11239

（3）90%数据作为训练集，10%的数据作为测试集：

损失值：0.04668555275233854，正确率：100%

Duration: 5.64s，迭代次数：11674

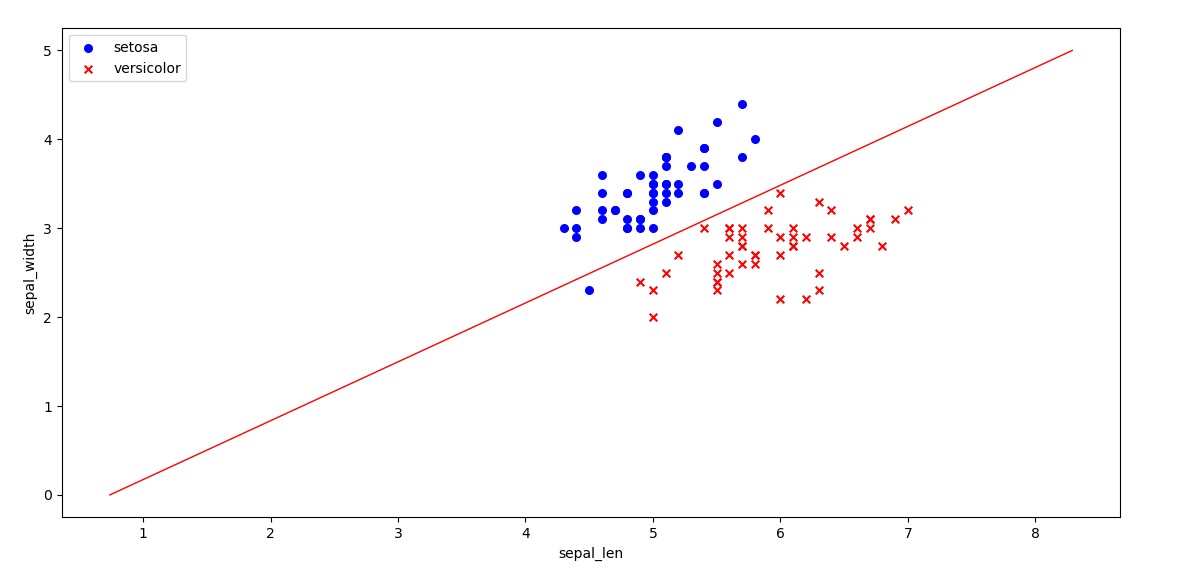
### 实验结果分析：

1. 三次实验的损失函数的值分别为0.01，0.05，0.04。由结果可知，当数据的测试比列为30%，训练比例为70%的时候，损失值最大，误差也最大。测试的准确率也最低，说明当训练数据较多时，损失值会有所上升。
2. 当数据的测试比例为50%，训练比例为50%的时候，损失函数值最小，为0.01，但是正确率为98%，损失值较小，正确率较低，说明训练数据较小，训练出来的模型准确性不够，虽然损失函数的值较小。
3. 当数据的测试比例为10%，训练数据比例为90%的时候，损失函数的值和数据比例分别为30%，70%的值要小，说明当训练数据较多时，损失函数的值会减小，测试结果表明正确率为100%，可以推断出当训练数据较多时，模型的正确率会上升。但是正确率过高也有可能是用于测试的数据过少导致的。

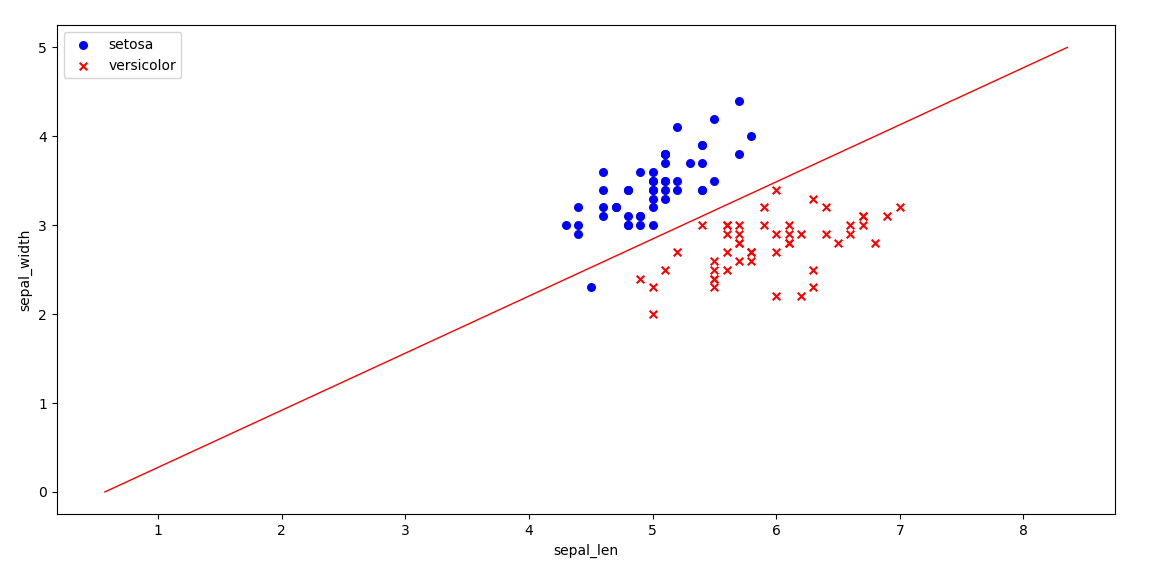
总结：根据以上的结果分析，可以推断出当训练的数据增加时，损失函数的值会增长，迭代次数增加，训练时间变长，当测试的数据减小时，预测结果的准确率会上升。

### 划分结果图：

50%作为训练集，50%作为测试集



70%作为训练集，30%作为测试集



90%作为训练集，10%作为测试集