**使用python实现对数几率回归**

计算机1601 1611640110 李俊杰

* 问题描述

选取四个属性之中其中两个：sepal length、sepal width。且只分为了两类，一类为Iris Setosa，另一类为Iris Versicolour。

问题目标利用对数几率回归实现二分类

H(x) =

Cost function = y \*ln(h(x)) + (1-y)\*(ln(1-h(x))（凸函数）

* 数据集描述

Data Set Information:

This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda & Hart, for example.) The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.   
  
Predicted attribute: class of iris plant.   
  
This is an exceedingly simple domain.   
  
This data differs from the data presented in Fishers article (identified by Steve Chadwick, spchadwick **'@'** espeedaz.net ). The 35th sample should be: 4.9,3.1,1.5,0.2,"Iris-setosa" where the error is in the fourth feature. The 38th sample: 4.9,3.6,1.4,0.1,"Iris-setosa" where the errors are in the second and third features.

Arrtibute Information:

1. sepal length in cm   
2. sepal width in cm   
3. petal length in cm   
4. petal width in cm   
5. class:   
-- Iris Setosa   
-- Iris Versicolour   
-- Iris Virginica

来源：archive.ics.uci.edu/ml/datasets/Iris

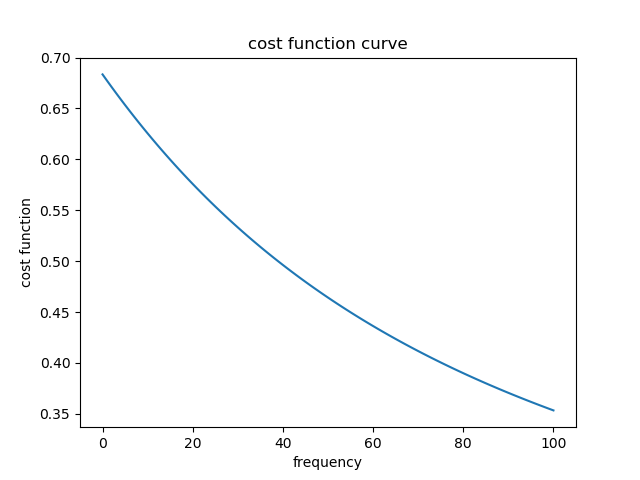
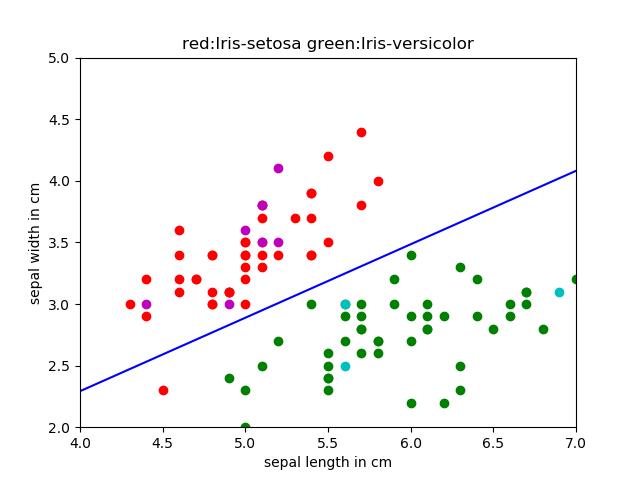
* 实验结果图

（注：红色（训练集）和品红（测试集）代表iriis-setosa 绿色（训练集）和青色（测试集）代表iris-versicolor）

一、90%训练 10%测试

cost\_function:0.353333

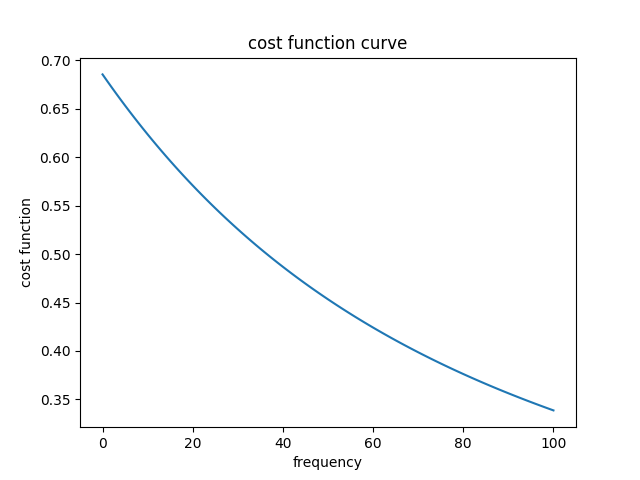
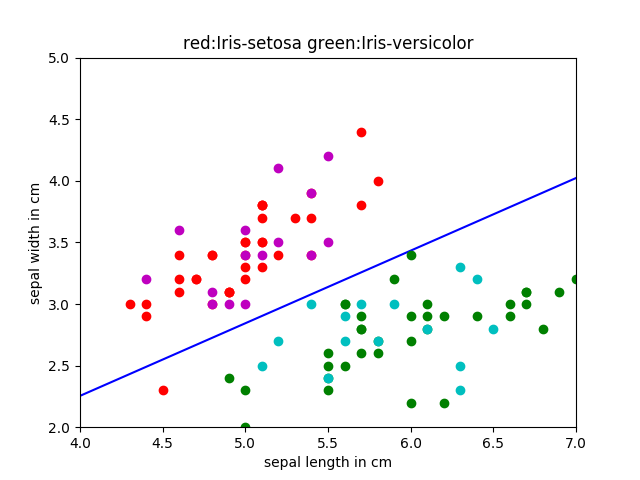
w1=-0.924655,w2=1.550722,b=0.144003

正确率为：100.000000%

二、70%训练 30%测试

cost\_function:0.330073

w1=-0.924291,w2=1.623614,b=0.126019

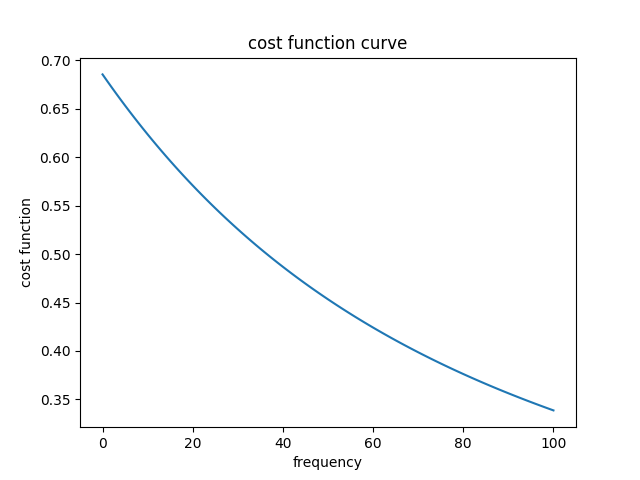
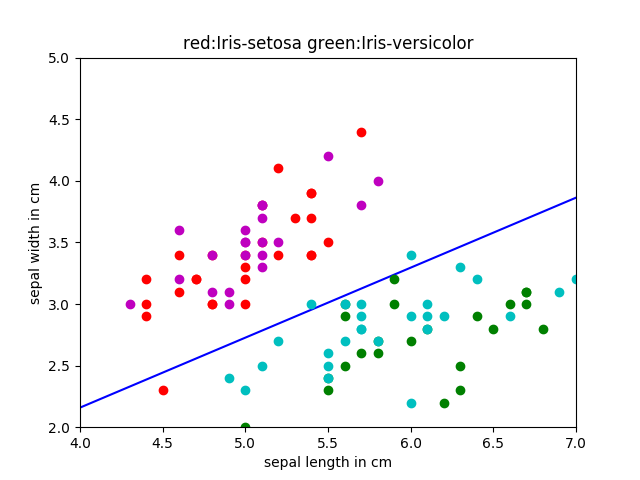
正确率为：93.333333%

三、50%训练50%测试

cost\_function:0.342479

w1=-0.908425,w2=1.581385,b=0.162187

正确率为：95.918367%



* 实验结果分析

Iris数据集前两类花的特征区别比较明显，即使经过较少次数的训练也能够比较好的使两类数据线性可分

随着训练次数增加，分割会更加明显

1-9训练结果经过测试具有较高的准确率