

**《机器学习》课程实验报告**

**学 院 软件学院**

**专 业 软件工程**

**组 员**  **袁澍民**

**学 号 20153061352**

**邮 箱 695228799@qq.com**

**指导教师**  **吴庆耀**

**提交日期** **2017年 12月 7日**

## 实验题目:

## 线性回归、线性分类及梯度下降

## 2. 实验时间：

## 2017年 12月 2日

## 报告人:

## 袁澍民

## 4. 实验目的:

1. 进一步理解线性回归和梯度下降的原理。
2. 在小规模数据集上实践。
3. 体会优化和调参的过程。

## 数据集以及数据分析：

线性回归使用的是LIBSVM Data中的Housing数据，包含506个样本， 每个样本有13个属性。Housing数据集每行共有14列，第一列为y值， 是房价值，其他13个为13个x属性值。

线性分类使用的是LIBSVM Data中的australian数据，包含690个样本， 每个样本有14 个属性。australia数据集每行共有15列，第一列为分类 （正类与负类），其他14个为14个x属性值。

## 实验步骤:

线性回归和梯度下降

1.读取实验数据

2.将数据集切分为训练集和验证集

3.线性模型参数初始化

4.选择Loss函数及对其求导

5.求得所有样本对Loss函数的梯度

6.取梯度的负方向

7.更新模型参数

8.在训练集上测试并得到Loss函数值，在验证集上测试并得到Loss函数 值

9.重复步骤5-8若干次，画出和随迭代次数的变化图

线性分类和梯度下降

1.读取实验数据，使用sklearn库的load\_svmlight\_file函数读取数据。

2.将数据集切分为训练集和验证集

3.支持向量机模型参数初始化

4.选择Loss函数及对其求导

5.求得所有样本对Loss函数的梯度。

6.取梯度的负方向

7.更新模型参数

8.选择合适的阈值

9.重复步骤5-8若干次，画出和随迭代次数的变化图

## 代码内容:

线性回归：

-----------------------------------------------------------------------------------------------

梯度下降

sumT=0

LT=[0 for i in range(70)]

LV=[0 for i in range(70)]

k=0

for i in range(80):

L=np.array((X2train\*W-Ytrain))

for j in range(404):

sumT=sumT+L[j]\*\*2

sumT=sumT/404

W=W-0.0004\*(2\*X2train.T\*(X2train\*W-Ytrain))

if(i>=10):

LT[k]=sumT

LV[k],a=validation(X2vali,Yvali,W)

k+=1

-----------------------------------------------------------------------------------------------

线性分类

-----------------------------------------------------------------------------------------------

C=0.1

alpha=0.0015

LT=[0 for i in range(150)]

LV=[0 for i in range(150)]

k=0

b=0

for i in range(150):

dw,Lt=Gw(w,Xtrain,Ytrain,552,b)

db=Gb(w,Xtrain,Ytrain,552,b)

dw=w+C\*dw

db=C\*db

w=w-alpha\*dw

b=b-alpha\*db

LT[k]=Lt

LV[k]=validation(w,Xvali,Yvali,138,b)

k+=1

-------------------------------------------------------------------------------------------------

## 选择的评估方法（留出法，交叉验证，k折交叉验证等）:

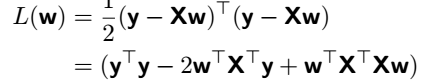
线性回归、线性分类：留出法

## 模型参数的初始化方法:

线性回归、线性分类：全0初始化

## 选择的loss函数及其导数:

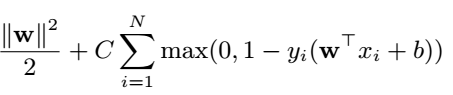
线性回归loss函数：



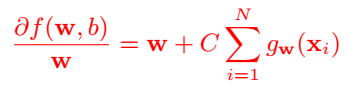
导数：

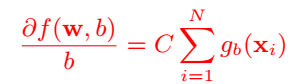


线性分类loss函数：



导数：





## 11.实验结果和曲线图:

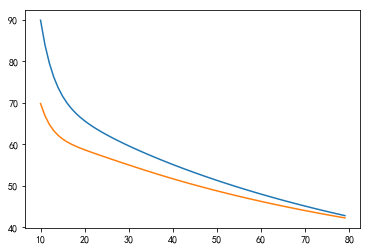
## ·超参数选择（η,epoch等）：

线性回归：η=0.0004

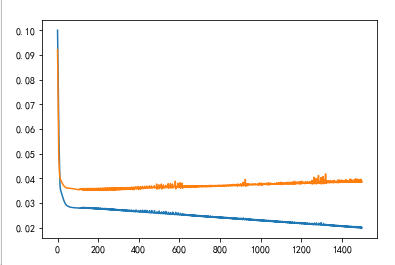
线性分类：η=0.0015 C=0.1

## ·评估结果（根据选择的评估方法）：

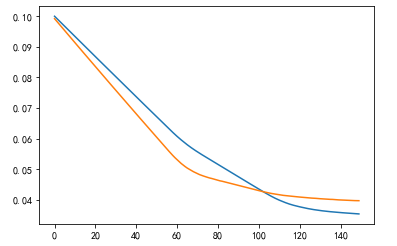
线性回归在学习率很小的时候没有达到最优解



线性分类在迭代次数过多的情况下出现过拟合



线性分类在学习率很小的时候没有达到最优解



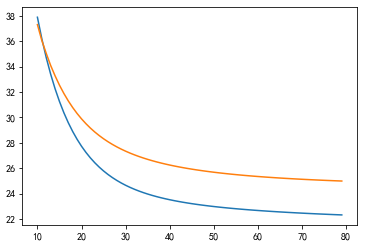
## ·预测结果（最佳结果）：

线性回归Loss为26

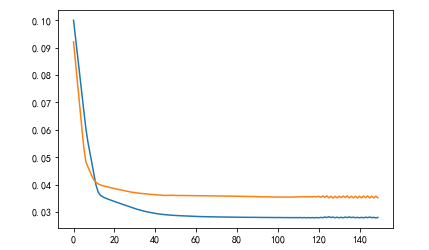
线性分类Loss为0.04

## ·loss曲线图：（蓝色为训练集，橙色为验证集）

线性回归：



线性分类：（这里为了能够更好的与测试集进行对比，就将 loss除以了样本总数）



## 实验结果分析:

这里我的训练次数还不是很多，不过我也测试了增加训练次数的结果，无论是线性回归还是线性分类都出现了过拟合现象，即训练集Loss减少但是验证集Loss反而增加。我这里取了一个相对合适的训练次数，拟合效果也还可以，loss总和还可以接受。

## 对比线性回归和线性分类的异同点：

线性回归与线性分类都是用线性模型去模拟一组特征数据，不过线性回归是去拟合数据，而线性分类是去区分数据。正是由于这一点，线性分类可能存在多条符合要求的分类的（超平面），所以还要综合考虑其容忍扰动最大的超平面。

## 实验总结：

这次实验做的还算成功，不过刚入手的时候还是比较生疏，所以花的时间也不短。通过这次实验可以说是再一次、更深的理解了机器学习这一概念，对线性模型也理解更深刻了，对之后的实验也有帮助。体验了调参的过程，对参数的设置也有了一些经验。