

**《机器学习》课程实验报告**

**学 院 软件学院**

**专 业 软件工程**

**组 员**  苏越

**学 号 201530612774**

**邮 箱 1090211523@qq.com**

**指导教师** 吴庆耀

**提交日期** **2017年 12 月 2 日**

## 1. 实验题目: 线性回归、线性分类与梯度下降

## 2. 实验时间：2017年 12 月 2 日

## 3. 报告人: 苏越

## 4. 实验目的: （1）进一步理解线性回归和梯度下降的原

## （2）在小规模数据集上实践。

## （3）体会优化和调参的过程

## 5. 数据集以及数据分析：

线性回归使用的是[LIBSVM Data](https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/)中的[Housing](https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html#housing)数据，包含506个样本，每个样本有13个属性。

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。

## 6. 实验步骤:

线性回归和梯度下降:

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，本次实验不切分测试集。使用train\_test\_split函数切分数据集。

3.线性模型参数初始化，可以考虑全零初始化，随机初始化或者正态分布初始化。

4.选择Loss函数及对其求导，过程详见课件ppt。

5.求得所有样本对Loss函数的梯度G。

6.取梯度G的负方向，记为D。

7.更新模型参数，。η为学习率，是人为调整的超参数。

8.在训练集上测试并得到Loss函数值L\_train，在验证集上测试并得到Loss函数值L\_validation。

9.重复步骤5-8若干次，画出L\_train和L\_validation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，本次实验不切分测试集。使用train\_test\_split函数切分数据集。

3.支持向量机模型参数初始化，可以考虑全零初始化，随机初始化或者正态分布初始化。

4.选择Loss函数及对其求导，过程详见课件ppt。

5.求得所有样本对Loss函数的梯度G。

6.取梯度G的负方向，记为D。

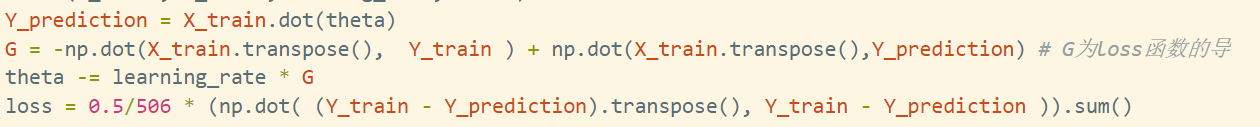
7.更新模型参数，。η为学习率，是人为调整的超参数。

8.选择合适的阈值，将计算结果大于阈值的标记为正类，反之为负类。在训练集上测试并得到Loss函数值L\_train，在验证集上测试并得到Loss函数值L\_validation。

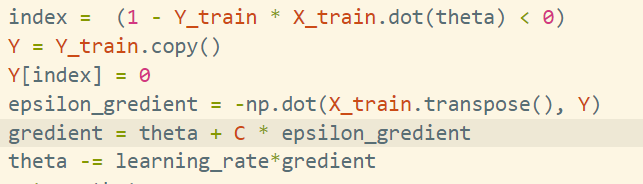
9.重复步骤5-8若干次，画出L\_train和L\_validation随迭代次数的变化图。

## 7. 代码内容:

线性分类（梯度下降部分）：



线性回归（梯度下降部分）：



## 8. 选择的评估方法（留出法，交叉验证，k折交叉验证等）:

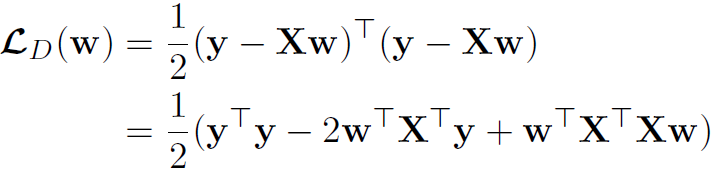
留出法

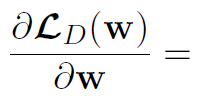
## 9. 模型参数的初始化方法:

初始化为0

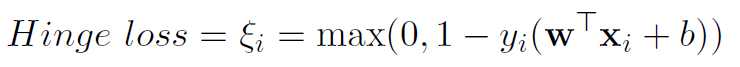
## 10.选择的loss函数及其导数:

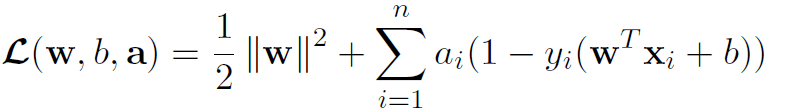
线性回归:





线性分类：





## 11.实验结果和曲线图:

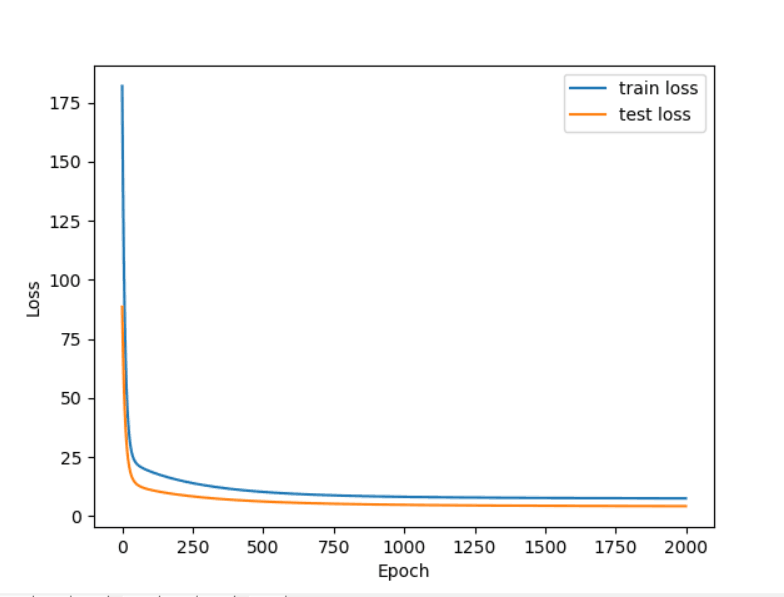
线性回归：

## 超参数选择（η,epoch等）：η=0.00003 epoch=2000

## 评估结果（根据选择的评估方法）：loss= 7.92335855527

## 预测结果（最佳结果）：loss=7.93356478447

## loss曲线图：



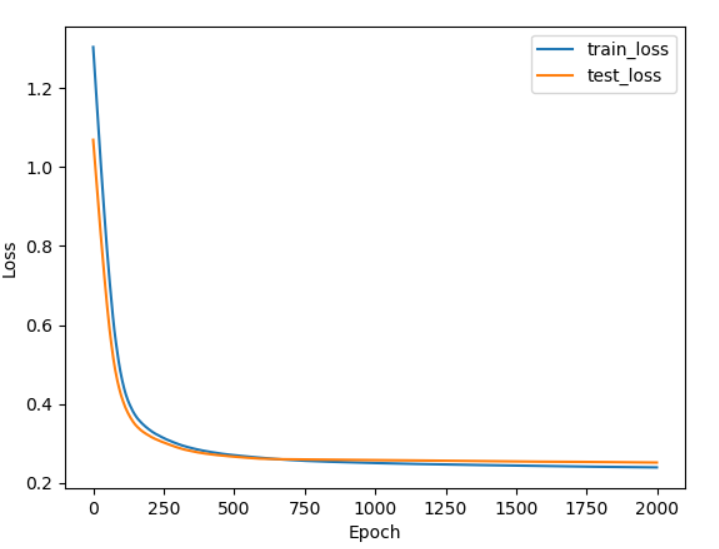
线性分类：

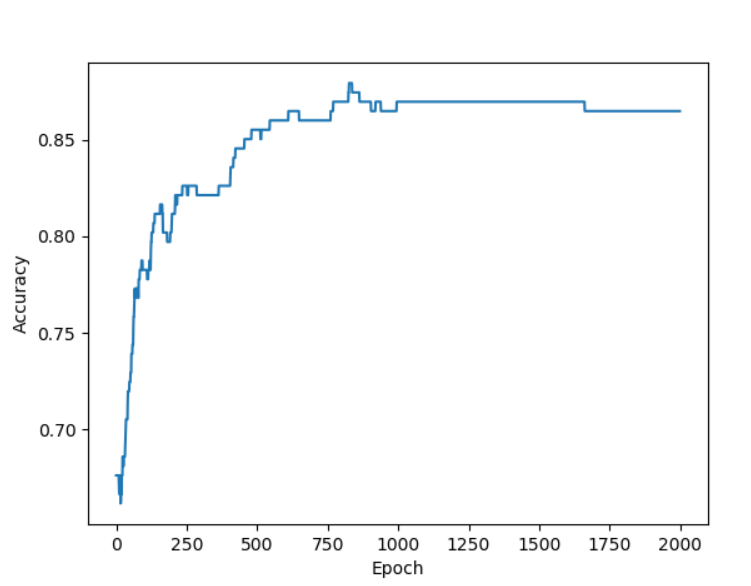
## 超参数选择（η,epoch等）：η=0.00003 epoch=2000

## 评估结果（根据选择的评估方法）：loss= 0.25143792056

## 预测结果（最佳结果）：loss=0.265069194832

## loss曲线图：





## 12.实验结果分析:线性回归和分类训练集和验证集的loss最后都会收敛。

## 13.对比线性回归和线性分类的异同点：连续变量预测为回归，离散变量预测为分类，一般来说，这两个问题本质上是相同的，即模型的拟合。但分类问题的Y值又称为标号，更为离散。此外，相同的y值可能对应于大量x，这些x具有一定的范围。因此，分类问题是在某些方）与y相对应的。回归问题模型更倾向于x在小区域对应于y。

## 14.实验总结：线性回归和线性分类都是模型的拟合，线性回归是连续的变量，线性分类是离散的变量。通过这次实验，感受到了超参数对评估结果的影响，以及数据的处理和算法的数学运算都对结果的求解影响重大。