

**《机器学习》课程实验报告**

**学 院 软件学院**

**专 业 软件工程**

**组 员**  胡永浩

**学 号 \_\_\_201530611630**

**邮 箱 1458732467@qq.com**

**指导教师**  **吴庆耀**

**提交日期** **2017年12 月 8 日**

## 1. 实验题目: 线性回归，线性分类和梯度下降

## 2. 实验时间：2017年12 月 2 日

## 3. 报告人:胡永浩

## 4. 实验目的:

1）进一步了解线性回归，线性分类和梯度下降的原理

2）在小规模数据集上实践

3）体会优化和调参的过程

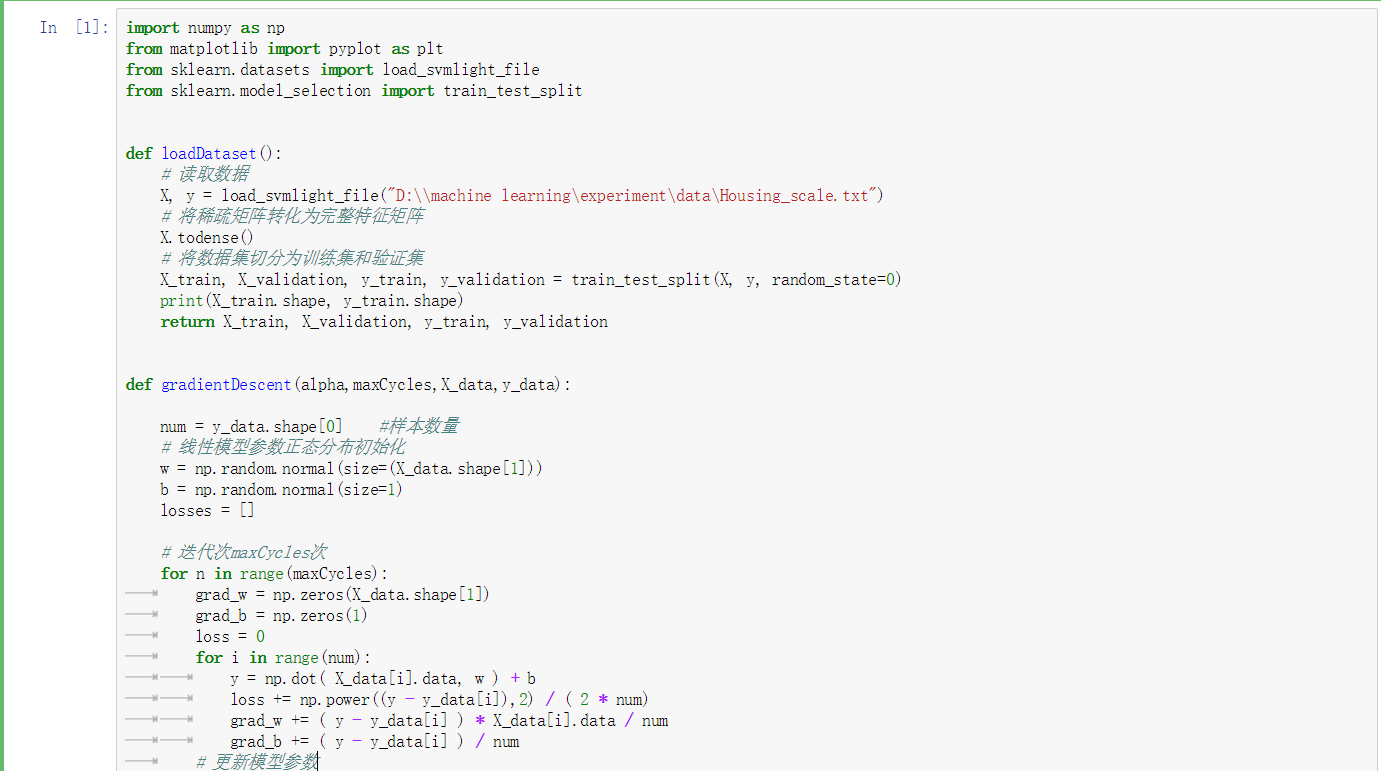
线性回归和梯度下降

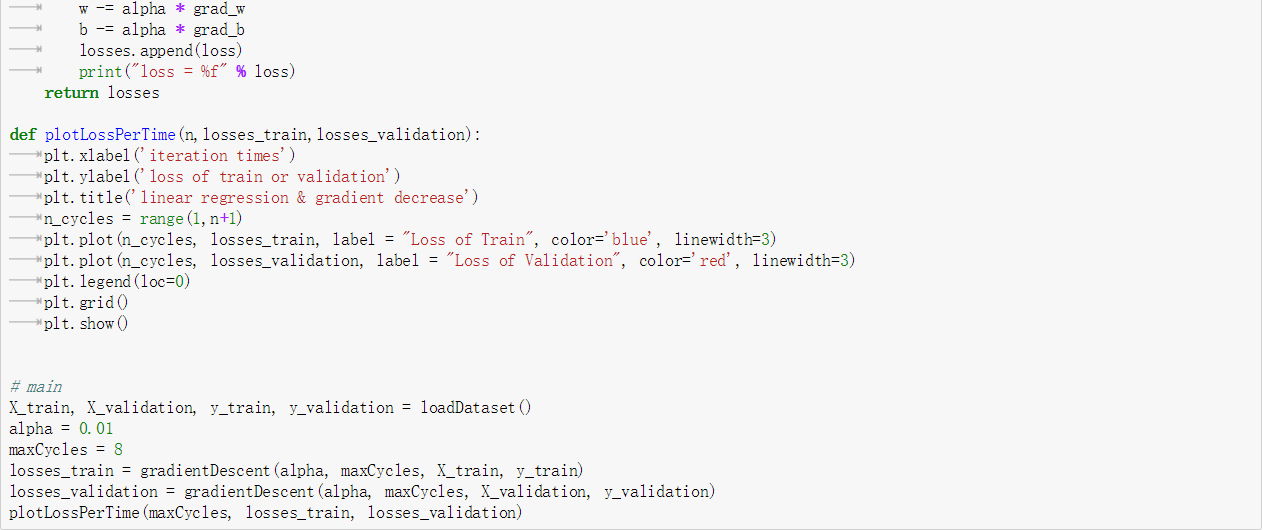
## 5. 数据集以及数据分析：线性回归使用的是[LIBSVM Data](https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/" \t "https://www.zybuluo.com/chenyaofo/note/_blank)中的[Housing](https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html" \l "housing" \t "https://www.zybuluo.com/chenyaofo/note/_blank)数据，包含506个样本，每个样本有13个属性。

## 实验步骤:

1. 读取实验数据，使用sklearn库的[load\_svmlight\_file](http://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_svmlight_file.html" \t "https://www.zybuluo.com/chenyaofo/note/_blank)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](http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html" \t "https://www.zybuluo.com/chenyaofo/note/_blank)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，Wt=Wt-1+nD。n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Ltrain，在验证集上测试并得到Loss函数值Lvalidation。
9. 重复步骤5-8若干次，****画出Ltrain和Lvalidation和随迭代次数的变化图****。

## 代码内容:





## 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:正态分布初始化

## 10.选择的loss函数及其导数:hinge

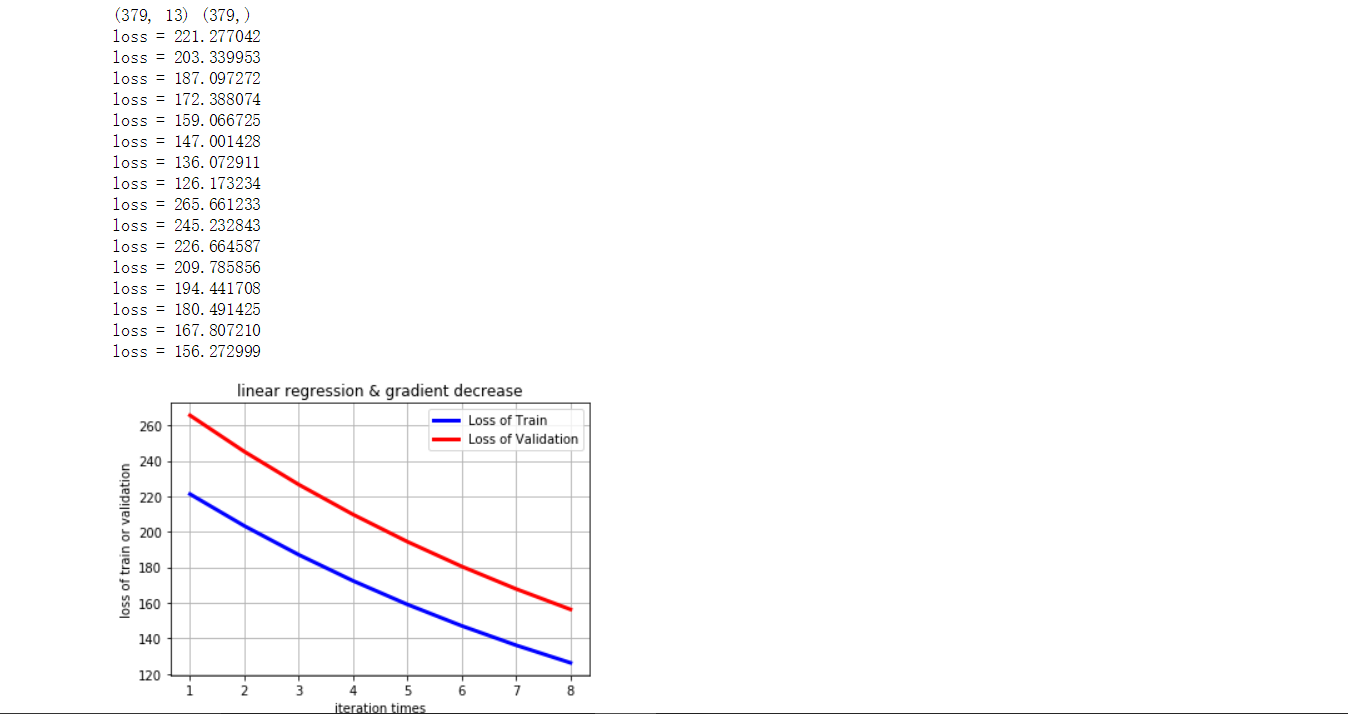
ℓ(θ)=1/2∑i=1m(hθ(x(i))−y(i))2=1/2∑i=1m(θTx(i)−y(i))2

## 11.实验结果和曲线图:

## 超参数选择（η,epoch等）：

## alpha = 0.01

## maxCycles = 8



## 实验结果分析:过拟合较严重

线性分类和梯度下降

5. 数据集以及数据分析：线性分类使用的是[LIBSVM Data](https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/" \t "https://www.zybuluo.com/chenyaofo/note/_blank)中的[australian](https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html" \l "australian" \t "https://www.zybuluo.com/chenyaofo/note/_blank)数据，包含690个样本，每个样本有14 个属性。

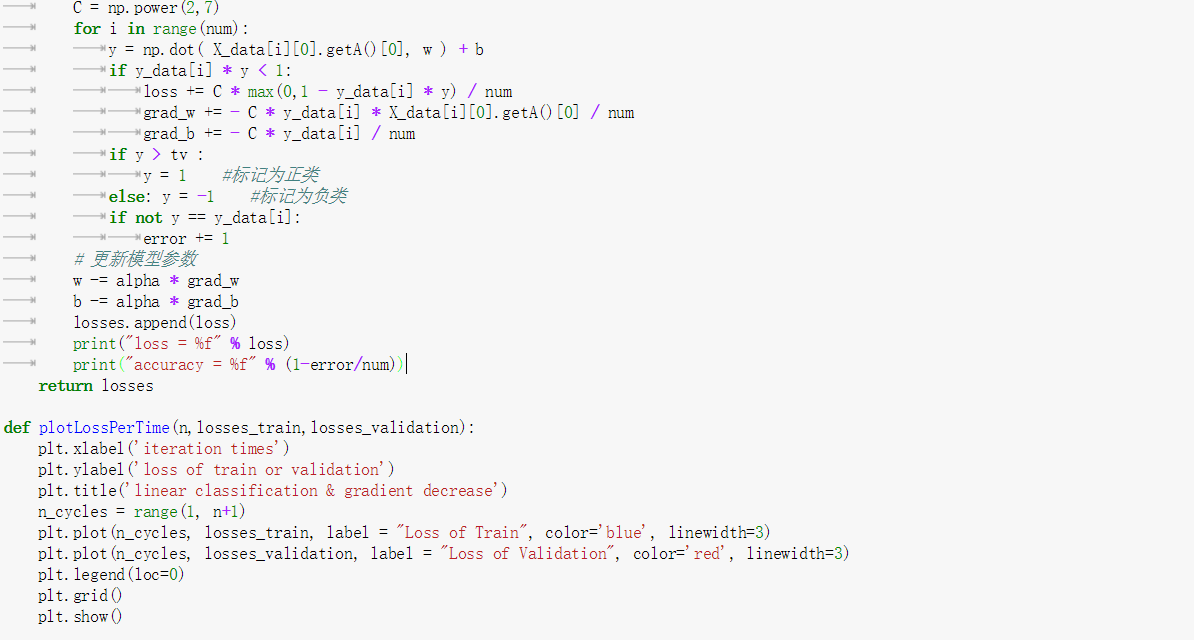
## 实验步骤:

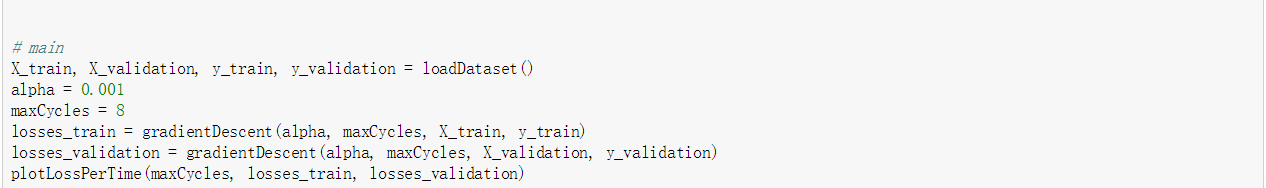
1. 读取实验数据，使用sklearn库的[load\_svmlight\_file](http://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_svmlight_file.html" \t "https://www.zybuluo.com/chenyaofo/note/_blank)函数读取数据。
2. 将数据集切分为训练集和验证集，本次实验不切分测试集。使用[train\_test\_split](http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html" \t "https://www.zybuluo.com/chenyaofo/note/_blank)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，Wt=Wt-1+nD。n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Ltrain，在验证集上测试并得到Loss函数值Lvalidtion。
9. 重复步骤5-8若干次，****画出Ltrain和Lvalidation随迭代次数的变化图****。

整理实验结果并完成实验报告（实验报告模板将包含在[示例仓库](https://github.com/chenyaofo/ML2017-lab-01" \t "https://www.zybuluo.com/chenyaofo/note/_blank)中）

## 代码内容:







## 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:正态分布初始化

## 10.选择的loss函数及其导数:hinge

ℓ(θ)=1/2∑i=1m(hθ(x(i))−y(i))2=1/2∑i=1m(θTx(i)−y(i))2

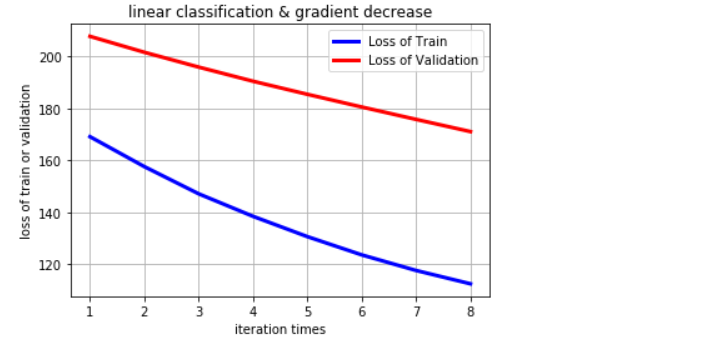
## 11.实验结果和曲线图:

## 超参数选择（η,epoch等）：

## alpha = 0.001

## maxCycles = 8





## 12.实验结果分析:过拟合较严重

## 对比线性回归和线性分类的异同点：

总的来说两个问题本质上都是一致的，就是模型的拟合（匹配）。 但是分类问题的y值(也称为label), 更离散化一些. 而且， 同一个y值可能对应着一大批的x,  这些x是具有一定范围的。

所以分类问题更多的是 (一定区域的一些x) 对应 着 (一个y).   而回归问题的模型更倾向于 (很小区域内的x，或者一般是一个x)  对应着  (一个y).

在把一个问题建模的时候一定要考虑好需求，让模型更好的与现实问题相对应。

## 14.实验总结：

初步了解了线性回归和线性分类的异同和调参的基本操作