

**《机器学习》课程实验报告**

**学 院 软件学院**

**专 业 软件工程**

**组 员**   **陈世鸿**

**学 号 201530611234**

**邮 箱 yingzi827165351@vip.qq.com**

**指导教师**  **吴庆耀**

**提交日期** **2017年12月7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:陈世鸿

## 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。

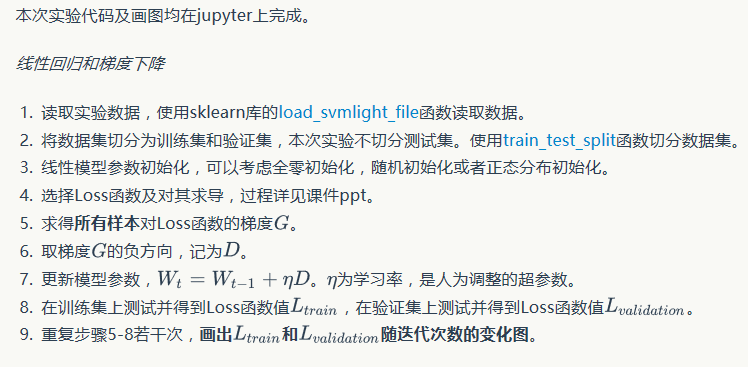
3. 体会优化和调参的过程。

## 5. 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。请自行下载scaled版本，并将其切分为训练集，验证集。

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。请自行下载scaled版本，并将其切分为训练集，验证集。

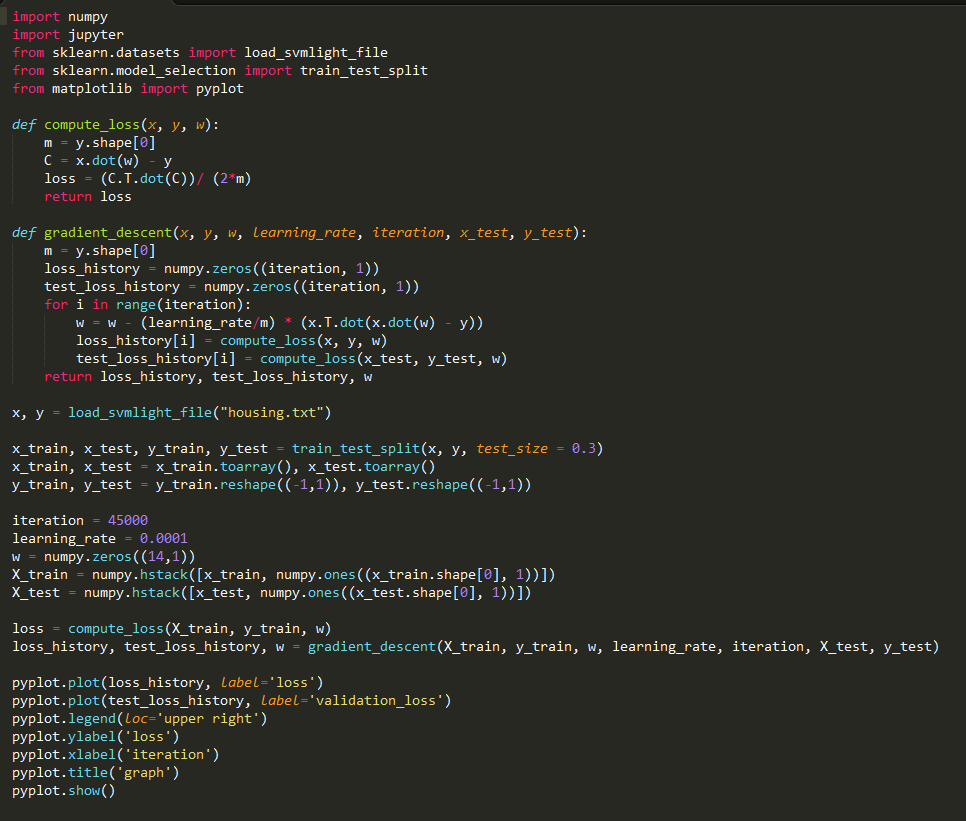
## 6. 实验步骤:



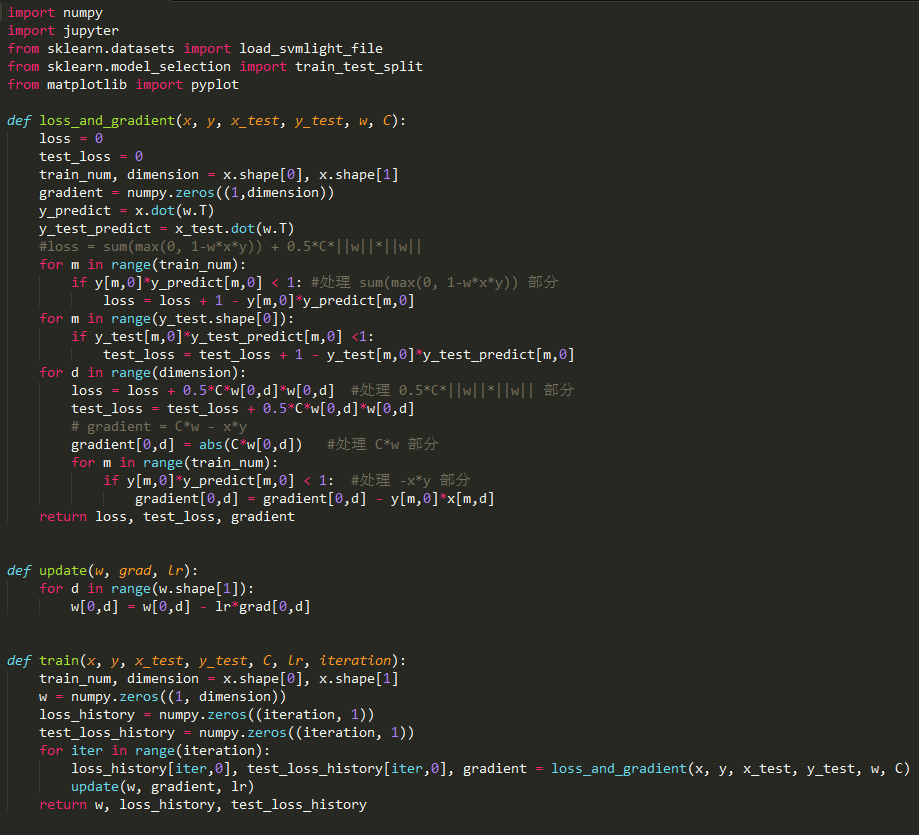


## 7. 代码内容:

RegressionExperiment:



ClassificationExperiment:



## 8. 选择的评估方法:

留出法。

## 9. 模型参数的初始化方法:

**Regression Experiment：**全零初始化。

**Classification Experiment：**全零初始化。

## 10.选择的loss函数及其导数:

**Regression Experiment：**

Loss函数：

导数：

**Classification Experiment：**

Loss函数：

导数：

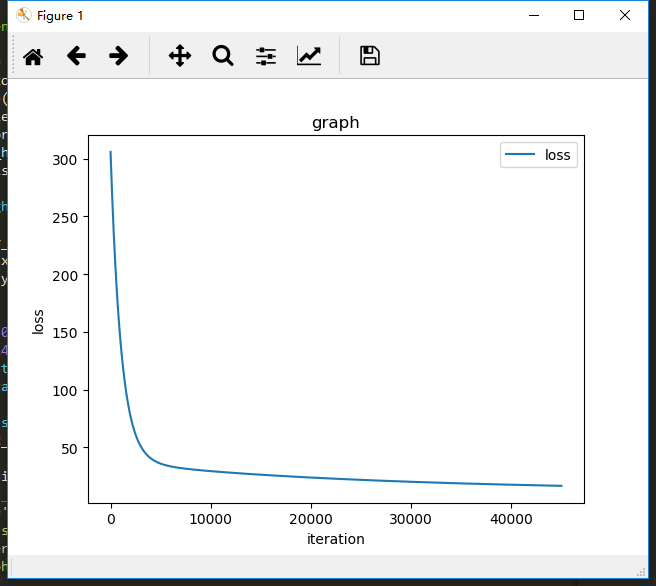
## 11.实验结果和曲线图:

超参数选择（η,epoch等）： **Regression Experiment：** η= 0.0001 epoch = 45000

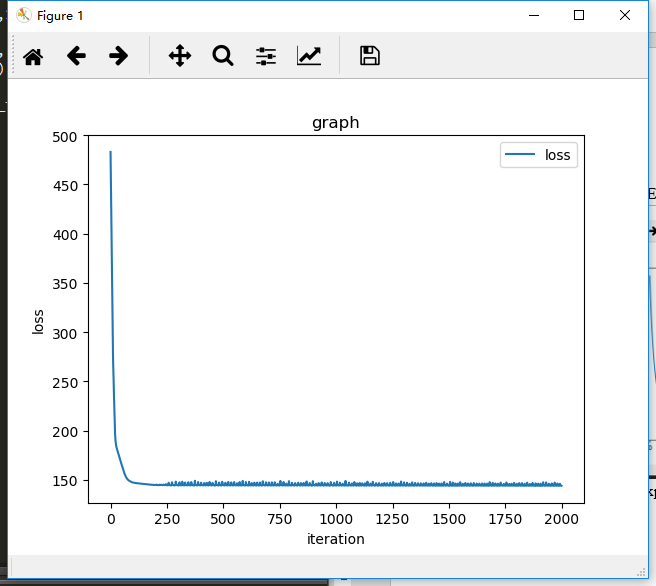
**Classification Experiment：** η= 0.0001 epoch = 2000 C = 0.3

## 评估结果（根据选择的评估方法）：

**Regression Experiment：**

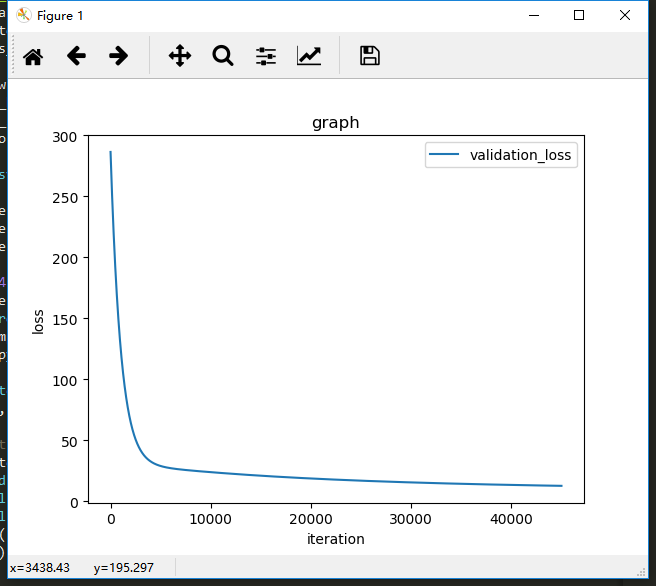


**Classification Experiment：**

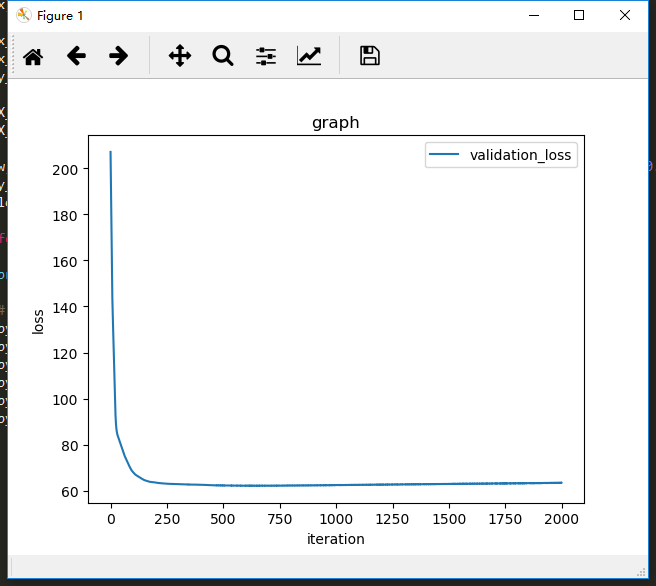


## 预测结果（最佳结果）：

**Regression Experiment：**

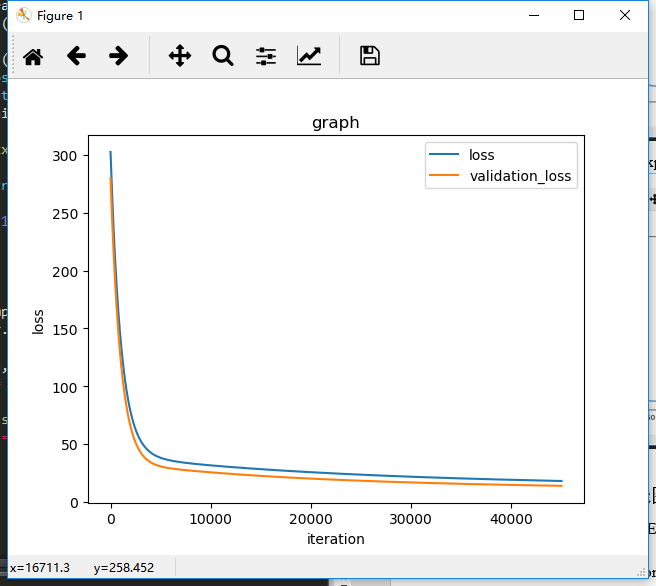


**Classification Experiment：**

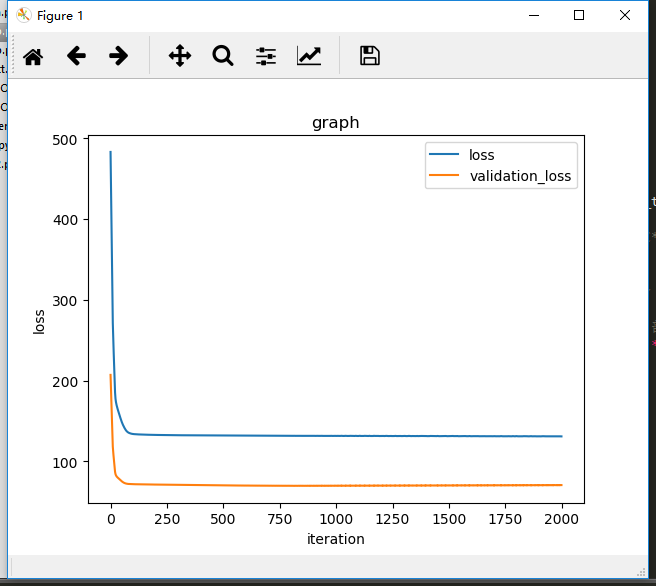


## loss曲线图：

**Regression Experiment：**



**Classification Experiment：**



## 12.实验结果分析:

在实验中，w被初始化为零，在每一次迭代更新w参数时，利用损失函数对w的导数，乘以学习率η得到新的w，重复此过程直至w收敛。在一定迭代次数后，w找到局部最优点，而在凸函数里面，局部最优点就是全局最优点，loss稳定在最小值。

## 13.对比线性回归和线性分类的异同点：

**同：**

对于这两类问题都有以下几个步骤,

* 如何选取一个合理的模型。
* 选择一个误差函数
* 采取一切可能的技术(e.g. 导数下降法，解极值方程法) 求出最好的模型参数

**异：**

总的来说两个问题本质上都是一致的，就是模型的拟合。 但是分类问题的y值, 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).

## 14.实验总结：

对公式要有正确的理解，不能盲目照搬；在实现算法的时候，还要考虑性能，因为样本的数量问题，优秀的算法能有效地节省时间。

对于线性回归，学习率的选取和迭代次数的选择尤为重要。如果学习率过大，会导致拟合得不好，无法收敛；学习率过小，收敛过慢，在本次实验中学习率经过多次调整，设为0.0001，有效解决问题。迭代次数过大，可能会导致过拟合，迭代次数过小，可能会出现欠拟合，而本次实验的loss在迭代次数为45000时稳定。

对于线性分类，惩戒因子C取为0.3，C过大会导致欠拟合，过小会导致过拟合。