

**《机器学习》课程实验报告**

**学 院 软件学院**

**专 业 软件工程**

**组 员**   **夏俊煊**

**学 号 201530613139**

**邮 箱 875453294@qq.com**

**指导教师**  **吴庆耀**

**提交日期** **2017年12月03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03日

## 3. 报告人:夏俊煊

## 4. 实验目的:

## 进一步理解线性回归和梯度下降的原理。

## 在小规模数据集上实践。

## 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。

## 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*

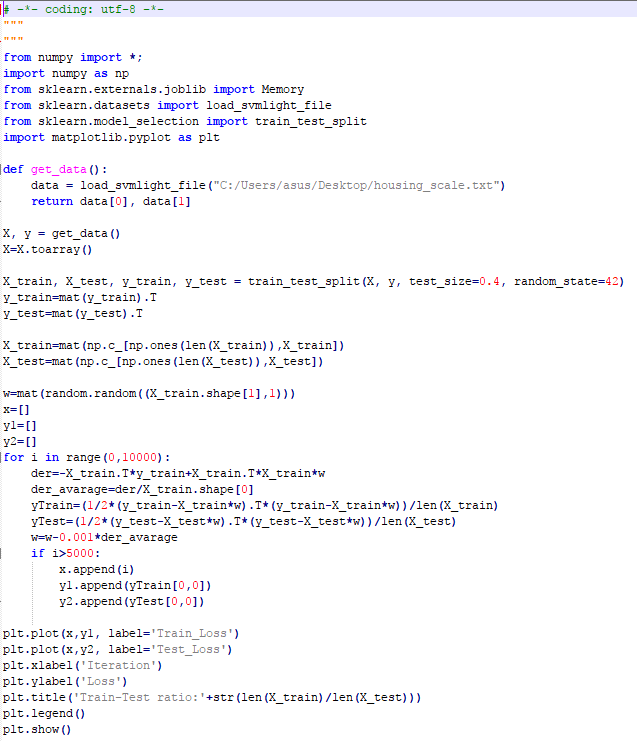
1. 读取实验数据，使用sklearn库的load\_svmlight\_file函数读取数据。
2. 将数据集切分为训练集和验证集，本次实验不切分测试集。使用train\_test\_split函数切分数据集。
3. 线性模型参数初始化，可以考虑全零初始化，随机初始化或者正态分布初始化。
4. 选择Loss函数及对其求导，过程详见课件ppt。
5. 求得**所有样本**对Loss函数的梯度。
6. 取梯度的负方向。
7. 更新模型参数。
8. 在训练集上测试并得到Loss函数值，在验证集上测试并得到Loss函数值。
9. 重复步骤5-8若干次，**画出随迭代次数的变化图**。

*线性分类和梯度下降*

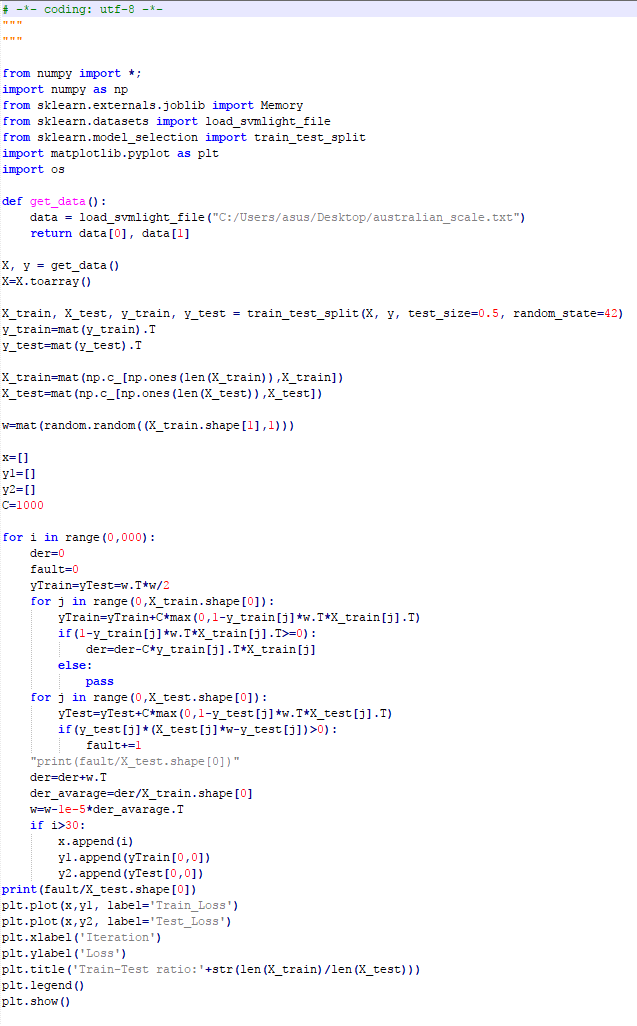
1. 读取实验数据，使用sklearn库的load\_svmlight\_file函数读取数据。
2. 将数据集切分为训练集和验证集，本次实验不切分测试集。使用train\_test\_split函数切分数据集。
3. 支持向量机模型参数初始化，可以考虑全零初始化，随机初始化或者正态分布初始化。
4. 选择Loss函数及对其求导，过程详见课件ppt。
5. 求得**所有样本**对Loss函数的梯度。
6. 取梯度的负方向。
7. 更新模型参数。
8. 选择合适的阈值，将计算结果**大于阈值的标记为正类，反之为负类**。在训练集上测试并得到Loss函数值，在验证集上测试并得到Loss函数值。
9. 重复步骤5-8若干次，**画出随迭代次数的变化图**。

## 7. 代码内容:

**线性回归：**



**线性分类：**

****

（针对线性回归和线性分类分别填写8-12内容）

## 线性回归：

## 8-1. 选择的评估方法（留出法，交叉验证，k折交叉验证等）:

评估方法：留出法-将数据集切分为训练集与测试集

## 9-1. 模型参数的初始化方法:

随机初始化

## 10-1.选择的loss函数及其导数:

Loss函数：

导数：

## 11-1.实验结果和曲线图:

## 超参数选择（η,epoch等）：η=0.001

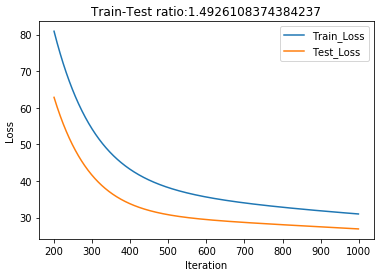
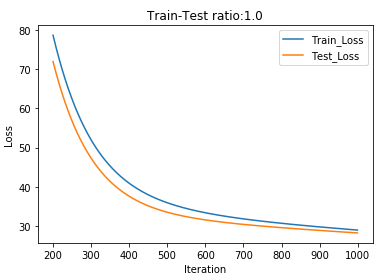
## 评估结果（根据选择的评估方法）：

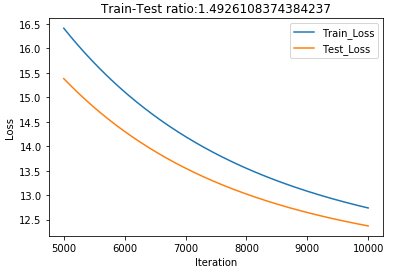
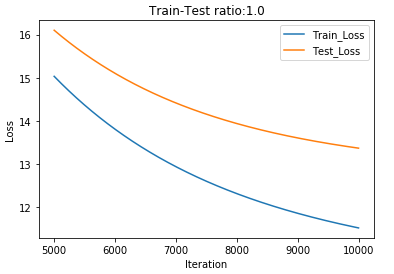
调节了多次train-test比例，迭代次数，与学习率，各个结果如下loss曲线图

## 预测结果（最佳结果）：

最好的结果见下图第四张，迭代了10000次

loss曲线图：





## 12-1.实验结果分析:

可以看到，对于Train-Test比为1：1的随着迭代次数的增加，Test\_Loss下降的趋势会减弱，即已经训练的大体上令人满意了，对于Train-Test比为1.5：1的情况，由于训练样本多，所以Train\_Loss会始终在Test\_Loss之上，但Test\_Loss结果还可以

**线性分类**

## 8-2. 选择的评估方法（留出法，交叉验证，k折交叉验证等）:

评估方法：留出法-将数据集切分为训练集与测试集

## 9-2. 模型参数的初始化方法:

随机初始化

## 10-2.选择的loss函数及其导数:

Loss函数：

导数：

## 11-2.实验结果和曲线图:

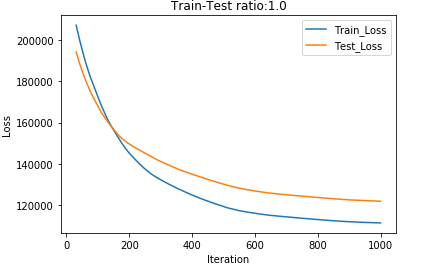
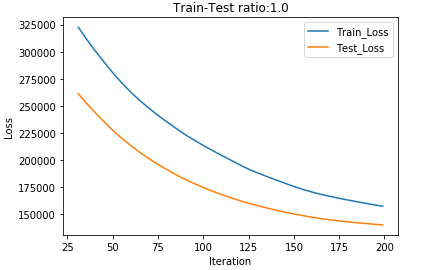
## 超参数选择（η,epoch等）：C=1000, η=0.00001

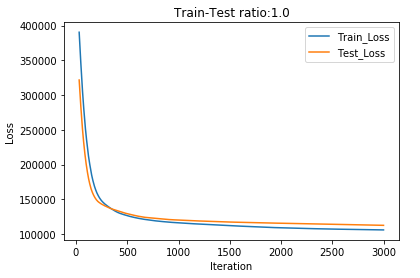
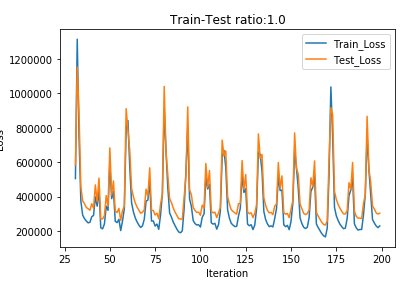
## 评估结果（根据选择的评估方法）：

多次调整了学习率，超参数C等，因为有多次循环，未能找出矩阵运算的版本，所以运行很慢，未大量调试

## 预测结果（最佳结果）：如图所示，见图二

## loss曲线图：





## 12-2.实验结果分析:

当学习率偏大时，如最后一张图，抖动会很明显，但同时准确率也会最快的到达最优值附近，为0.8左右，前一二张图分别为0.78与0.8左右

## 13.对比线性回归和线性分类的异同点：

线性分类要考虑间隔，最优化也比线性分类要困难一些，对数学要求也高

线性回归是为了得出一个模拟函数使得能预测出正确的y值，而线性分类则是考虑得出一个函数使得对于数据集的划分达到最优

## 14.实验总结：

这次实验有效的将所学应用到了实践，不仅稳固了这部分的知识，同时也开阔了视野，也学了一些python，收获良多