

**《机器学习》课程实验报告**

**学 院 软件学院**

**专 业 软件工程**

**组 员**  **周昭育**

**学 号 201530613900 \_**

**邮 箱 2605503083@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、线性分类与梯度下降

## 2. 实验时间：2017年 12月 2日

## 3. 报告人:周昭育

## 4. 实验目的:

1. 进一步理解线性回归、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。

## 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。使用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。使用scaled版本，并将其切分为训练集，验证集。

## 6. 实验步骤:

*线性回归和梯度下降*

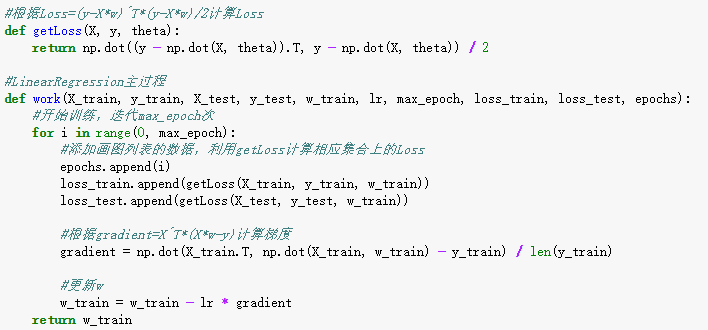
1. 读取实验数据，使用sklearn库的[load\_svmlight\_file](http://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_svmlight_file.html)函数读取数据。
2. 将数据集切分为训练集和验证集，本次实验不切分测试集。使用[train\_test\_split](http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html)函数切分数据集。
3.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

*线性分类和梯度下降*

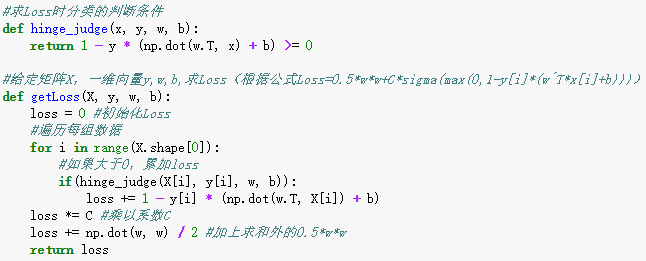
1. 读取实验数据，使用sklearn库的[load\_svmlight\_file](http://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_svmlight_file.html)函数读取数据。
2. 将数据集切分为训练集和验证集，本次实验不切分测试集。使用[train\_test\_split](http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html)函数切分数据集。
3.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

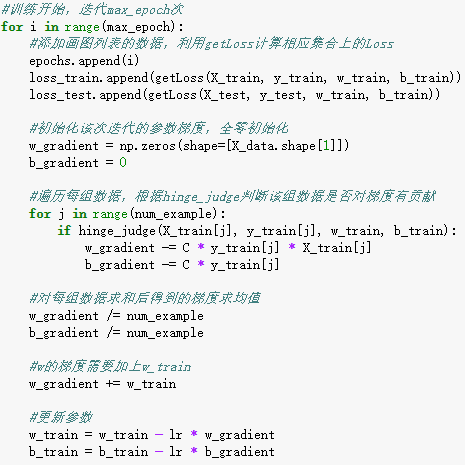
## 7. 代码内容:

*线性回归*



*线性分类*





## 8. 选择的评估方法（留出法，交叉验证，k折交叉验证等）:

使用留出法，将数据集随机分为训练集和验证集，本次实验不设测试集，验证集的比例为33%

## 9. 模型参数的初始化方法:

使用全零初始化

## 10.选择的loss函数及其导数:

*线性回归*

**Loss函数：**



**Loss函数的导数：**

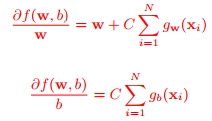


*线性分类*

**Loss函数：**



**Loss函数的导数：**



## 11.实验结果和曲线图:

## 超参数选择（η,epoch等）：

*线性回归*

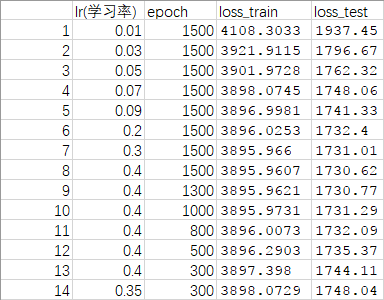
学习率η=0.35, 迭代次数epoch=300

*线性分类*

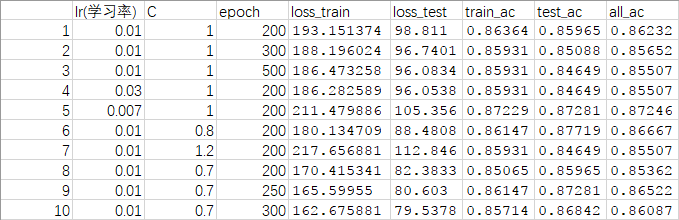
学习率η=0.01, 迭代次数epoch=250，C=0.7

## 评估结果（根据选择的评估方法）：

*线性回归*



线性分类



## 预测结果（最佳结果）：

*线性回归*

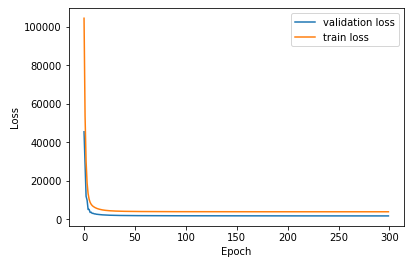
第14组参数，Loss\_train=3898.0729, Loss\_validation=1748.04，曲线较为平滑，迭代次数适中。

*线性分类*

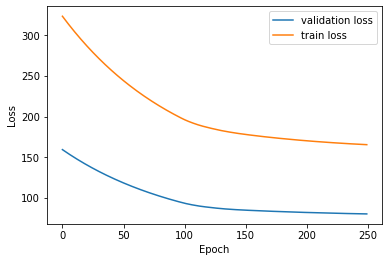
第9组参数，Loss\_train=165.59955,Loss\_validation=79.5378,所有数据集分类准确率为0.86522，与其他参数组相比，Loss较低，准确率较高，综合结果最好。

## loss曲线图：

*线性回归*



*线性分类*



## 12.实验结果分析:

*线性回归*

线性回归实验中，学习率在0.4左右达到一个低的极点，一定范围内，学习率较高时，Loss下降的速度会增加，但Loss曲线变得陡峭，降低学习率可以增加曲线平滑度。

线性回归的梯度下降速度很快，在300次以后，Loss的变化已经不明显。

本次实验选用评估结果中的第14组参数为最佳参数，这是Loss与曲线平滑度以及收敛速度的权衡。

*线性分类*

线性分类实验中，学习率在0.01左右达到一个低的极点，一定范围内，Loss的变化与线性回归的变化规律类似。

该实验中一个重要的结果是Loss的减小不意味着准确率的增加，这是由于Loss函数与错误样本的偏离程度有关，而准确率与错误样本的偏离程度无关。因此，本次实验选择评估结果中的第10组参数为最佳参数，这组参数在Loss与准确率两方面都有不错的表现。

## 13.对比线性回归和线性分类的异同点：

*相同点*

两者训练的都是线性模型，进行的都是线性变换。

*不同点*

线性回归的输出是连续的值，而线性分类的输出是离散的分类结果

## 14.实验总结：

本次实验中，我进一步了解了线性回归、线性分类和梯度下降的原理，初步体会了调参的过程。由于对于python不够熟悉，实验过程中出现了较多意外情况，但经过本次实验，我也增加了一些python知识。总体上说，本次实验达到了实验目的。