

**《机器学习》课程实验报告**

**学 院 软件学院**

**专 业 软件工程**

**组 员**  **邓泽帅**

**学 号 201530611425**

**邮 箱 1321730442@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: 邓泽帅

## 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：

1. 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。
2. 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:

*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" \t "_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)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，记为G。
7. 更新模型参数， W\_t=W\_t-1 + ηD。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，在验证集上测试并得到Loss函数值。
9. 重复步骤5-8若干次，**画出和随迭代次数的变化图**。

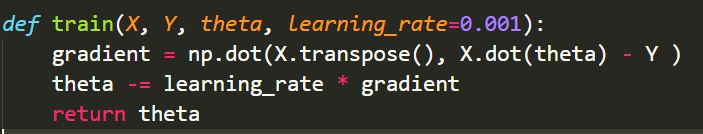
*线性分类和梯度下降*

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 + ηD。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，在验证集上测试并得到Loss函数值。
9. 重复步骤5-8若干次，**画出和随迭代次数的变化图**。

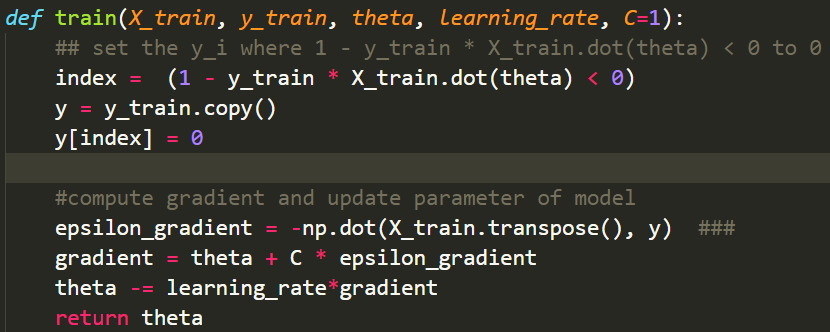
## 7. 代码内容:

（针对线性回归和线性分类分别填写8-12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）:

线性回归：离差平方和

线性分类：交叉验证

## 模型参数的初始化方法:

线性回归：全零初始化

线性分类：随机初始化

## 选择的loss函数及其导数:

线性回归： loss：

梯度

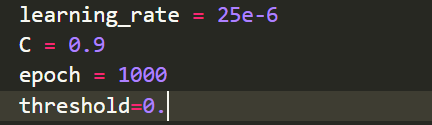
线性分类： loss：

梯度

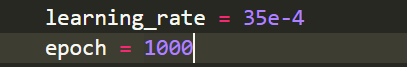
## 11.实验结果和曲线图:

## 超参数选择（η,epoch等）：

线性回归：



线性分类：



## 评估结果（根据选择的评估方法）：

## 预测结果（最佳结果）：

线性回归：

minimal train loss: 11.4968753689

minimal test loss: 10.4454852119

线性分类：

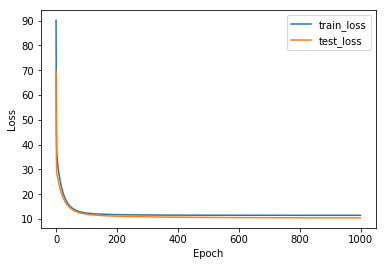
minimal train loss: 0.295144458228

minimal test loss: 0.330699250912

minimal accuracy of classification: 0.872807017544

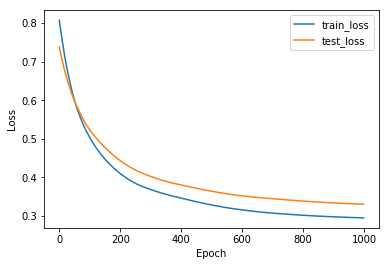
## loss曲线图：

线性回归：

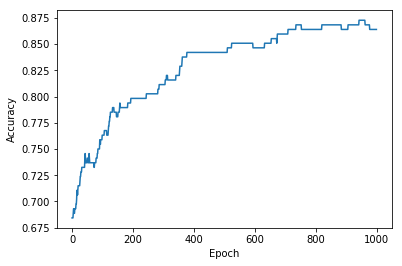


线性分类的loss曲线

线性分类：



线性分类的loss曲线



线性分类的分类准确度

## 12.实验结果分析:

线性回归的训练集和验证集的loss相差不大，下降的趋势大体相同，最终的loss比较合理；

线性分类的训练集和验证集的loss相差不大，下降的趋势大体相同，但是分类的准确度上升的不是很平滑，有的地方甚至会有所下降。总的来说，最终结果还不错，准确度也达到87.2%。

## 对比线性回归和线性分类的异同点：

相同点：

* 1. 两种机器学习算法都是使用数据学习出具体的模型，然后根据模型得出预测的结果。
  2. 算法的核心都是先找出目标函数，然后根据目标函数找一个评估的算法，然后再使用评估算法进行梯度下降更新模型的参数。

不同点：

* 1. 线性回归与现行分类算法的评估不一样

## 实验总结：

结果本次实验，对线性回归和线性分类有了更加深刻的了解，在实验中，遇到的问题都将是以后宝贵的经验