

**《机器学习》课程实验报告**

**学 院 软件学院**

**专 业 软件工程**

**组 员**  袁振宇

**学 号 201530613542**

**邮 箱 ken1024448582@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 8日

## 3. 报告人: 袁振宇

## 4. 实验目的: 通过实验理解应用线性回归、线性分类与梯度下降

## 5. 数据集以及数据分析： housing集 australian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，
5. 求得**所有样本**对Loss函数的梯度。
6. 取梯度的负方向，记为。
7. 更新模型参数，。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" \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. 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，记为。
7. 更新模型参数，。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:

见代码

## 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:

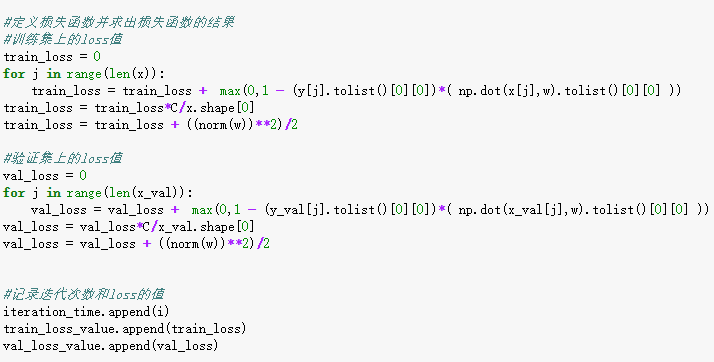
全零法

## 选择的loss函数及其导数:

线性回归：



线性分类：



## 11.实验结果和曲线图:

## 超参数选择（η,epoch等）：

（1）线性回归

theta=0.001

（2）线性分类

theta=0.001；C=0.5

## 评估结果（根据选择的评估方法）：

（1）线性回归

每次迭代更新theta，迭代次数固定

（2）线性分类

C不随迭代次数更新，theta进行更新

## 预测结果（最佳结果）：

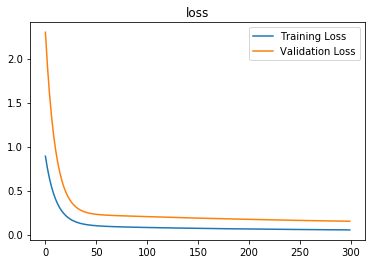
（1）线性回归

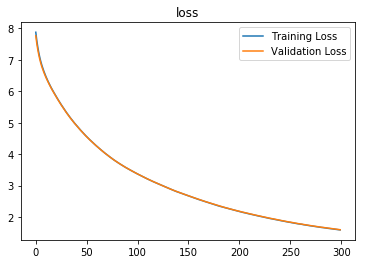
theta=0.0003

（2）线性分类

theta=0.001；C=0.5

## loss曲线图：





## 12.实验结果分析:

## 13.对比线性回归和线性分类的异同点：

共同点：数据都有特征和标记，都可以找出一个适当的模型进行拟合。都可以通过梯度下降的方法来求参数

不同点：线性回归是通过模型预测数据对应的值，线性分类侧重于用模型对数据进行分类。后者用到SVM进行调参。

以下截取于google：

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).

## 14.实验总结：

虽然对这部分知识有一定的理解和体会，但因为数学较差，理解起来有较大难度，学习起来很慢。而且是现学的python，只好一边参照别人的代码。通过这次实验，对这部分内容有了更深入的理解。理解了拟合的过程。