

**《机器学习》课程实验报告**

**学 院 软件学院**

**专 业 软件工程**

**组 员**   **李可欣**

**学 号 201530611999**

**邮 箱 609186106@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 上午9:00-12:00

## 3. 报告人: 李可欣

## 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。**

## 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。**

## 实验步骤:

**1）线性回归和梯度下降：**

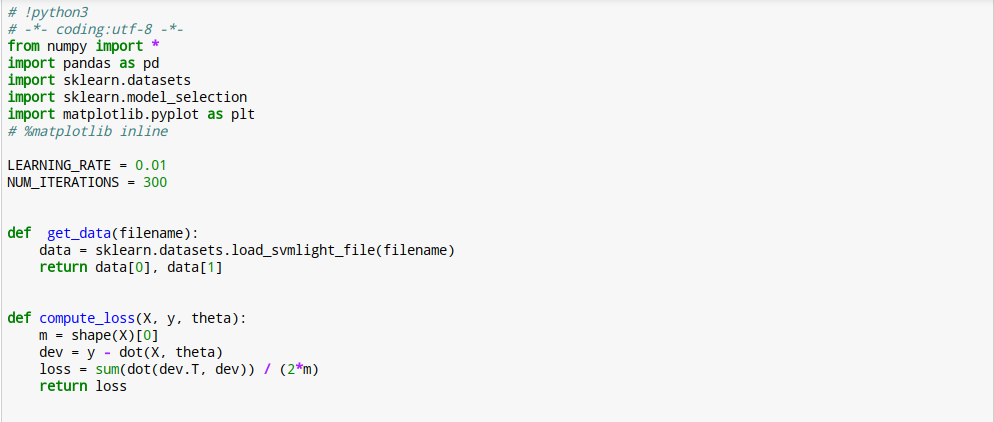
* 读取实验数据，使用sklearn库的load\_svmlight\_file函数读取数据。
* 将数据集切分为训练集和验证集，本次实验不切分测试集。使用train\_test\_split函数切分数据集。
* 线性模型参数初始化，可以考虑全零初始化，随机初始化或者正态分布初始化。
* 选择Loss函数及对其求导，过程详见课件ppt。
* 求得所有样本对Loss函数的梯度G。
* 取梯度G的负方向，记为D。
* 更新模型参数，w(t)=w(t-1)+nD。为学习率，是人为调整的超参数。
* 在训练集上测试并得到Loss函数值Ltrain，在验证集上测试并得到Loss函数值。
* 重复步骤5-8若干次，画出Ltrain和Lvalidation随迭代次数的变化图。

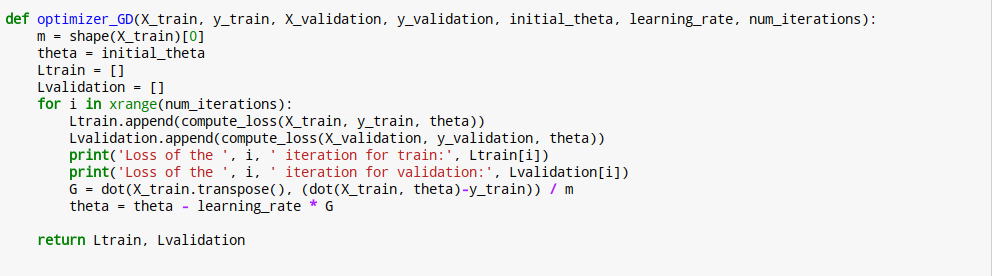
**2）线性分类和梯度下降：**

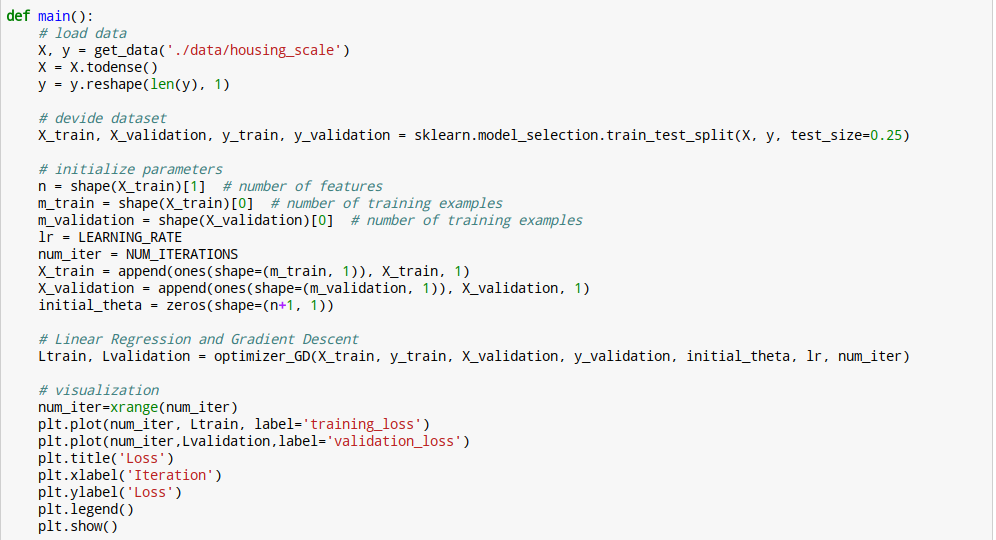
* 读取实验数据，使用sklearn库的load\_svmlight\_file函数读取数据。
* 将数据集切分为训练集和验证集，本次实验不切分测试集。使用train\_test\_split函数切分数据集。
* 支持向量机模型参数初始化，可以考虑全零初始化，随机初始化或者正态分布初始化。
* 选择Loss函数及对其求导，过程详见课件ppt。
* 求得所有样本对Loss函数的梯度G。
* 取梯度G的负方向，记为D。
* 更新模型参数，w(t)=w(t-1)+nD。为学习率，是人为调整的超参数。
* 在训练集上测试并得到Loss函数值Ltrain，在验证集上测试并得到Loss函数值。
* 重复步骤5-8若干次，画出Ltrain和Lvalidation随迭代次数的变化图。

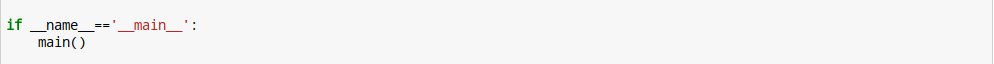
## 代码内容:

* **线性回归和梯度下降：**



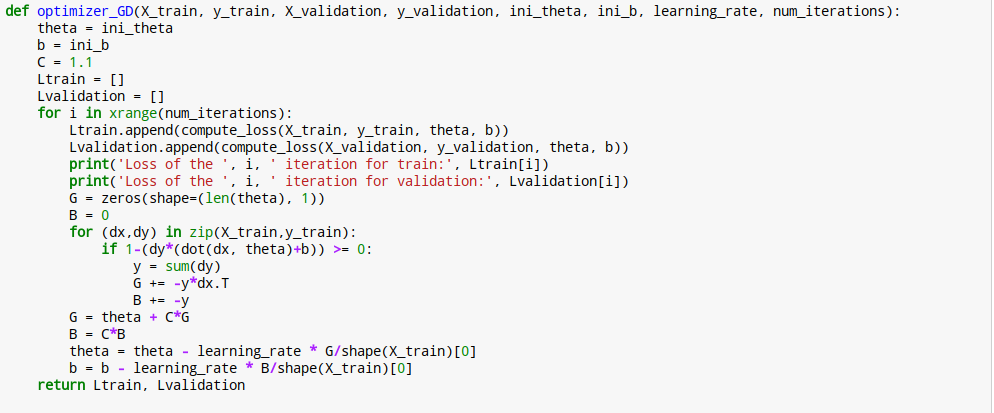


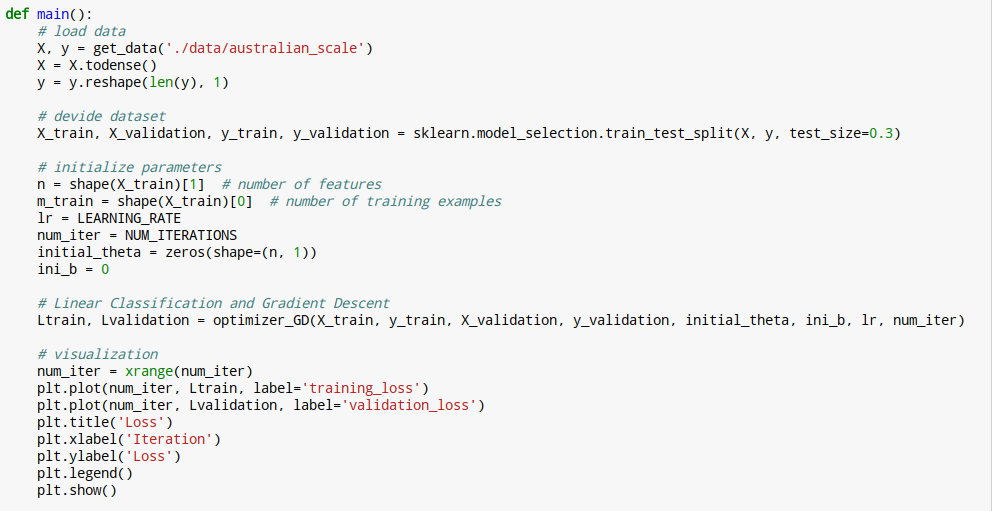


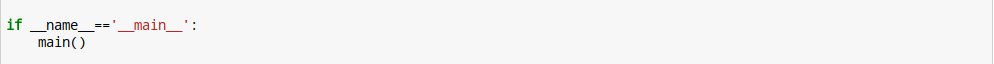


* **线性回归和梯度下降：**









## 选择的评估方法（留出法，交叉验证，k折交叉验证等）:

* **线性回归和梯度下降：留出法**
* **线性回归和梯度下降：留出法**

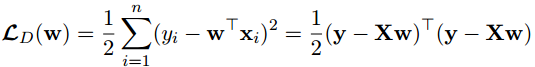
## 模型参数的初始化方法:

* **线性回归和梯度下降：全零初始化**
* **线性回归和梯度下降：全零初始化**

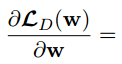
## 选择的loss函数及其导数:

* **线性回归和梯度下降：**

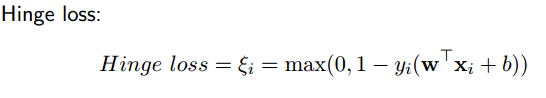
Loss函数:



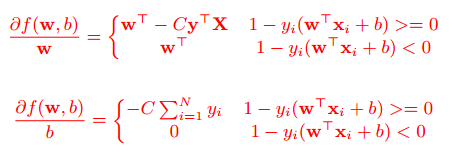
导数计算：



* **线性回归和梯度下降：**



导数计算：



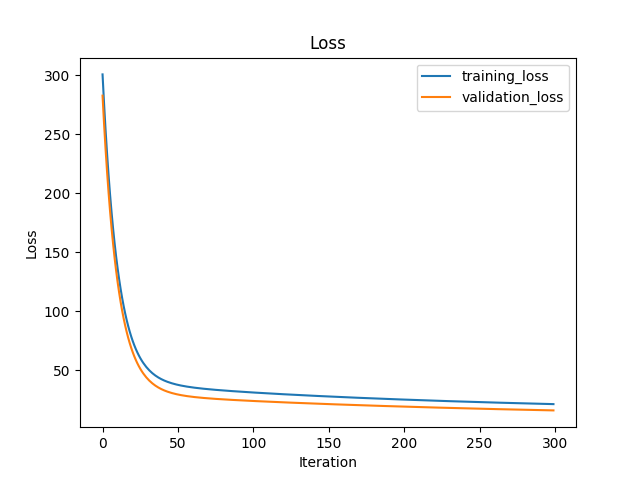
## 实验结果和曲线图:

* **线性回归和梯度下降：**

Learning rate: 0.01

Iteration = 300

theta = 0



* **线性回归和梯度下降：**

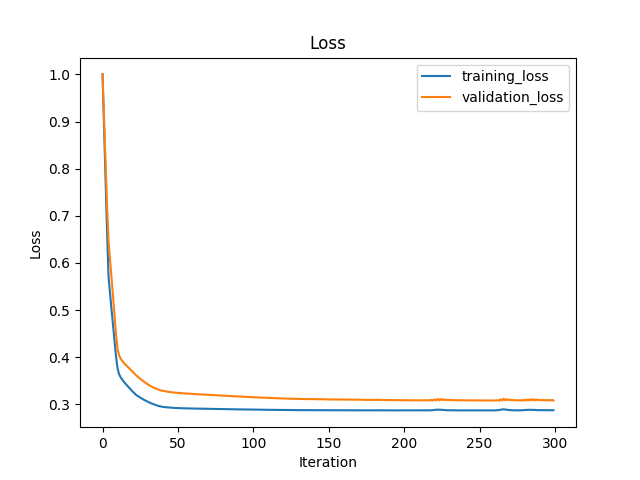
Learning rate: 0.1

Iteration = 300

theta = 0

b = 0

c =1.1



## 实验结果分析:

回归问题中结果显示validation\_loss比training\_loss还偏大一些，与一般常见结果不相一致。

分类问题的结果显示与一般常见结果相符。

## 13.对比线性回归和线性分类的异同点：

* 相同点：线性回归和线性分类这两个问题的本质是一致的，就是模型的拟合。
* 不同点：分类问题的y值更离散化一些，同一个y值可能对应着一批具有一定范围的x。回归问题的模型更倾向于很小区域内的或者是一个x对应一个y。

## 14.实验总结：

机器学习中，线性回归和线性分类是最基本的两类问题。对于这两类问题，有以下三个基本步骤：

* 选取一个合理的模型（此实验中均采用线性模型）。
* 选择一个合适的误差函数（此实验中回归问题的loss function为最小二乘误差，分类问题为hinge loss）。
* 采取合适的技术（gradient descent）求出最好的模型参数，使损失函数最小化。