

**《机器学习》课程实验报告**

**学 院 软件学院**

**专 业 软件工程**

**组 员**  颜常霖

**学 号 201530613368**

**邮 箱 693812981@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:颜常霖

## 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：

## 线性回归使用的是[LIBSVM Data](https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/" \t "https://www.zybuluo.com/chenyaofo/note/_blank)中的[Housing](https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html" \l "housing" \t "https://www.zybuluo.com/chenyaofo/note/_blank)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](https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/" \t "https://www.zybuluo.com/chenyaofo/note/_blank)中的[australian](https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html" \l "australian" \t "https://www.zybuluo.com/chenyaofo/note/_blank)数据，包含690个样本， 每个样本有14 个属性。

## 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 "https://www.zybuluo.com/chenyaofo/note/_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" \t "https://www.zybuluo.com/chenyaofo/note/_blank)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G。

（6）取梯度G的负方向，记为D。

（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 "https://www.zybuluo.com/chenyaofo/note/_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" \t "https://www.zybuluo.com/chenyaofo/note/_blank)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。

（5）求得所有样本对Loss函数的梯度。

（6）取梯度的负方向，记为。

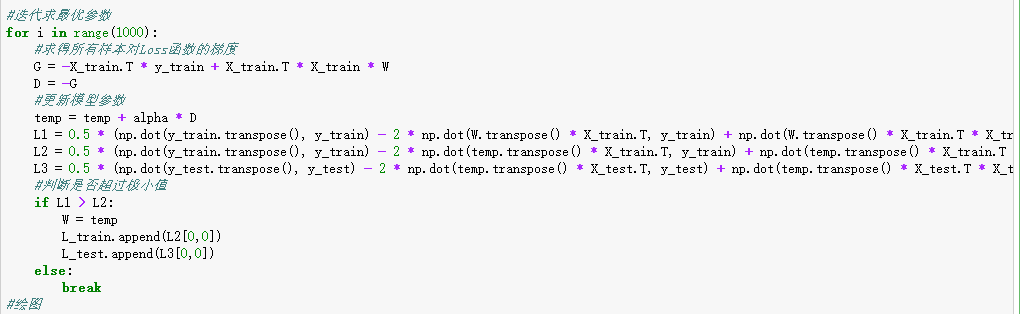
（7）更新模型参数，。为学习率，是人为调整的超参数。

（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。

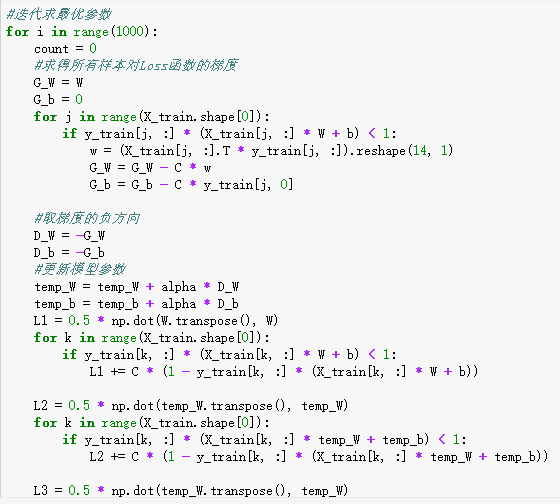
（9）重复步骤5-8若干次，画出和随迭代次数的变化图。

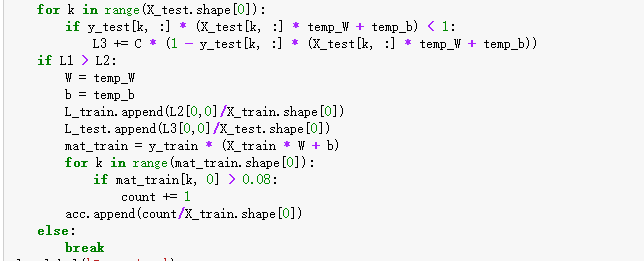
## 代码内容:

**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）

## 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：**

线性回归选择的评估方法为“留出法”，将数据集housing\_scale划分为两个互斥的集合，其中一个集合作为训练集X\_train,另外一个作为测试集X\_test。其中X\_train有339个样本，X\_test有167个样本。用X\_train进行训练后，如果模型在X\_test上的loss值越小，则模型越准确。

**线性分类：**

线性分类选择的评估方法为“留出法”，将数据集australian\_scale划分为两个互斥的集合，其中一个集合作为训练集X\_train,另外一个作为测试集X\_test。其中X\_train有462个样本，X\_test有228个样本。用X\_train进行训练后，如果模型在X\_test上的loss值越小，则模型越准确。用X\_train进行训练后，如果模型在数据集australian\_scale上有100个样本分类错误，那么其错误率为（100/690）\*100 % =14.5%,相应的，精度为1-14.5%=85.5%

1. **模型参数的初始化方法:**

**线性回归：**

线性回归采取的初始化方法为随机初始化，参数为0-10之间的整数。

**线性分类：**

线性分类采取的初始化方法为：w采用随机初始化，参数为0-1之间的浮点数。b初始化为0

## 选择的loss函数及其导数:

**线性回归**：

loss函数为：L=1/(2n)∑(y - h(w))^2=1/(2n)∑(y - wx)^2，其中n为样本数目，y为实际值，h(w)为预测值。

对w偏导：G = ∑(y - wx)x

**线性分类：**

**loss**函数为：L=1/n(1/2||w||^2 + C∑max(0, 1 - y(wx+b)))，其中C为大于0的常数,n为样本数目。

对w偏导：G\_w = C∑(w - yx)，当1 - y(wx + b)>0时。

对b偏导：G\_b = w - C∑y

## 11.实验结果和曲线图:

## 超参数选择（η,epoch等）：

线性回归：

超参数有：alpha。alpha为学习率，用于改变梯度下降中每次更新参数的间隔大小，是人为调整的超参数。

**线性分类：**

超参数有：alpha。alpha为学习率，用于改变梯度下降中每次更新参数的间隔大小，是人为调整的超参数。

C。C为正则项，可以用于防止过拟合的问题，需要用验证集来调整。

## 评估结果（根据选择的评估方法）：

线性回归：

alpha=0.01,迭代1000次，参数W为[ 0.09598164, 0.00305735, 0.84442861, 0.28463503, 0.36651456,0.74063828, 0.86785056, 0.35625434, 0.45247167, 0.76469104,0.5100941 , 0.89182554, 0.12235597]，损失值L\_test=278.897

alpha=0.001,迭代1000次，参数W为[-13.94465749, 1.60134057, 0.22324644, 0.89825139,-5.28026692, 9.67068899, -0.24677922, -9.91055097,3.67157681, -1.55594172, -4.56221415, 2.7938031 ,

-10.28974931], 损失值L\_test=12.606

alpha=0.0001,迭代1000次，参数W为[-12.65344542, 0.18914598, -0.94648512, 0.79247301,-4.54219238, 9.04816901, 0.19658528, -8.30573458,2.58811252, 0.14636265, -4.7095447 , 3.11827741,

-11.28281087]， 损失值L\_test=13.1

alpha=0.00001,迭代1000次，参数W为[-7.65904094, -2.74840286,-2.10740569, -1.64858577, -3.1066541 ,4.88289221, 1.25159884, -3.73174196, -0.15351954, -0.42785619,-1.86042904, 6.18826819, -7.27544039]，损失值L\_test=20.575

线性分类：

alpha=0.01,C=0.1，迭代1000次，参数W为[-0.01180035, 0.03474702, 0.08946935, 0.19912866, 0.19243165,-0.0154485 , -0.09222824, 1.03798636, 0.04110177, 0.31955688,0.00415604, 0.33708264, -0.19030014, 0.31823056]，损失值L\_test=0.0349，精确度acc=85.9307%

alpha=0.001,C=0.1，迭代1000次，参数W为[-0.00184085, -0.00226013, 0.038786 , 0.03584927, 0.03740587, -0.00302553, -0.02906222, 1.00695469, 0.00578807, 0.1040481 ,-0.00247172, 0.07809443, 0.01319898, 0.30099118]，损失值L\_test=0.0332，精确度acc=85.9307%

alpha=0.0001,C=0.1，迭代1000次，参数W为[ 0.35332559, 0.28588302, 0.34056459, 0.60795821, 0.42707524,0.55033597, 0.11749666, 1.03185668, 0.50425987, -0.14464365,0.21912302, 0.27088665, 0.25487923, 0.1118488 ]，损失值L\_test=0.0425，精确度acc=84.6320%

## 预测结果（最佳结果）：

线性回归：

alpha=0.001,迭代1000次，参数W为[-13.94465749, 1.60134057, 0.22324644, 0.89825139,-5.28026692, 9.67068899, -0.24677922, -9.91055097,3.67157681, -1.55594172, -4.56221415, 2.7938031 ,

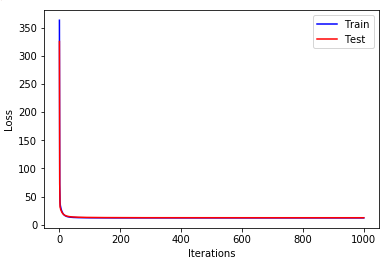
-10.28974931], 损失值L\_test=12.606

线性分类：

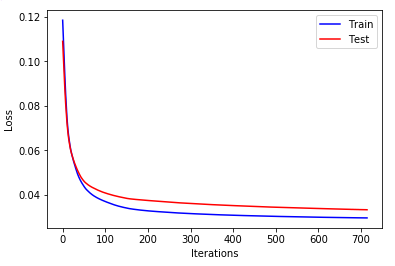
alpha=0.001,C=0.1，迭代1000次，参数W为[-0.00184085, -0.00226013, 0.038786 , 0.03584927, 0.03740587, -0.00302553, -0.02906222, 1.00695469, 0.00578807, 0.1040481 ,-0.00247172, 0.07809443, 0.01319898, 0.30099118]，损失值L\_test=0.0332，精确度acc=85.9307%

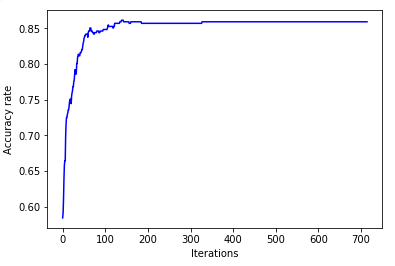
## loss曲线图：

线性回归：



线性分类：





## 实验结果分析:

**线性回归：**

当学习率alpha太大（>0.01）时，参数更新间隔大，超过了极小值，是的损失值大。

当学习率ahpha等于0.001时，可以得到最佳结果。当学习率alpha大于0.001时，结果略差与0.001时的结果。

**线性分类**：

当学习率ahpha等于0.001时，可以得到最佳结果。实验结果的准确度都在85%左右。

## 13.对比线性回归和线性分类的异同点：

线性回归和线性分类本质上都是一致的，都是模型的拟合。 但是分类问题的y值更离散化一些。而且，同一个y值可能对应着多个x,  这些x是具有一定范围的。

线性分类问题更多的是一定区域的多个x对应着一个y。而回归问题的模型更倾向于小区域内的x对应着一个y。

## 14.实验总结：

通过这次实验，使我进一步理解线性回归，线性分类和梯度下降的原理。

让我学会了如何获取和使用数据集进行实验。

让我体会了使用训练集和验证集进行优化和调参的过程。