

计算机与信息技术学院

**实验名称 线性模型&神经网络&SVM综合实验报告**

**姓 名 龙敏**

**学 号 18281046**

**日 期 2020.12.2**

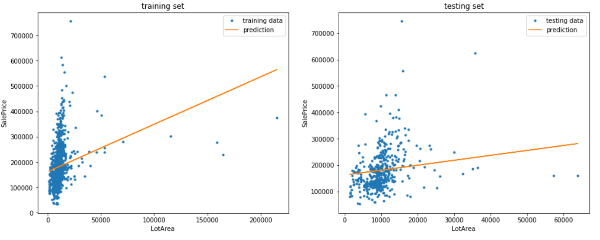
1. **LinearModel**
2. **一元线性回归**

使用一元线性回归模型对房价进行预测, 采用不同特征得到模型1, 2, 3.

1. 模型1，特征：LotArea

MAE: 53912.821831126574

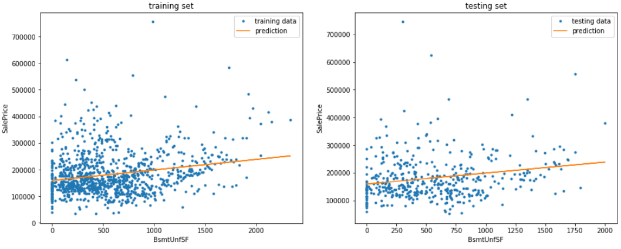
RMSE: 75795.63312280484



1. 模型2，特征：BsmtUnfSF

MAE: 53944.93997603965

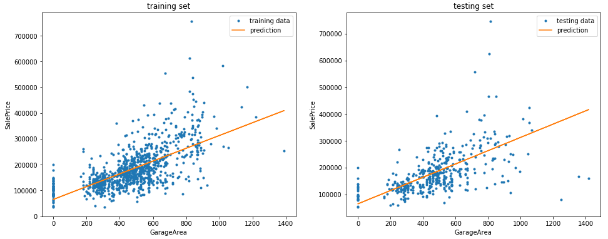
RMSE: 76895.8671879706



1. 模型3，特征：GarageArea

MAE: 43064.63752598681

RMSE: 65406.08811697361

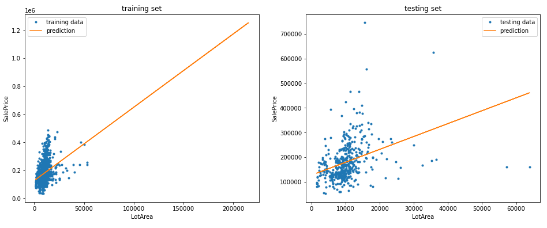


1. 剔除训练集中的离群值(outlier)，然后重新训练模型，观察模型预测性能的变化

模型4，特征：LotArea

MAE: 52423.413193762746

RMSE: 74965.38186222738

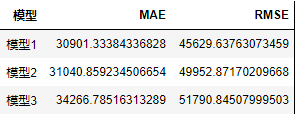


1. **多元线性回归**

以下三个模型使用特征如下, 各个模型的性能如下表所示

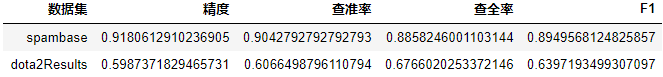
1. 模型1使用的特征：'MSSubClass','LotArea','OverallQual','OverallCond','YearBuilt','YearRemodAdd'
2. 模型2使用的特征：'BsmtFinSF1','BsmtFinSF2','BsmtUnfSF','TotalBsmtSF','1stFlrSF','2ndFlrSF'
3. 模型3使用的特征:

'LowQualFinSF', 'GrLivArea', 'BsmtFullBath', 'BsmtHalfBath', 'FullBath', 'HalfBath'



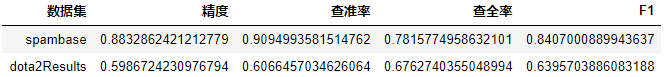
1. **对数几率回归**

使用对数几率回归, 将数据集spamx和dota2x的全部特征用于训练, 在两个数据集下得到以下两个结果如下表所示



1. **线性判别分析**

使用线性判别分析完成垃圾邮件分类问题和Dota2结果预测问题,经过训练后模型在两个数据集下性能如下表所示

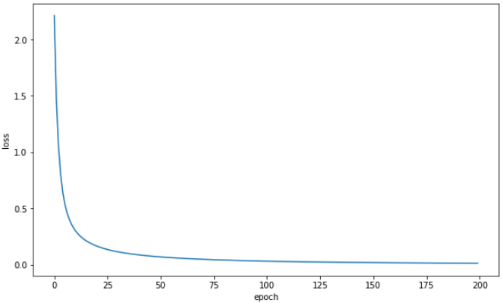


1. **综合任务-白酒质量分类(选做)**

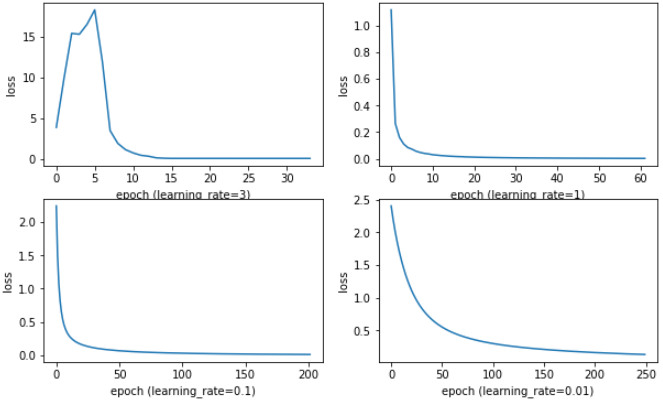
使用精度作为评价指标。使用一元线性回归、多元线性回归、对数线性回归等线性回归模型对葡萄酒质量进行预测，计算其十折交叉验证的精度。对比不同模型的预测能力结果如下表所示



1. **NeutralNetwork**
2. 使用sklearn.neural\_network.MLPClassifier完成手写数字分类任务, 精度达到了98%

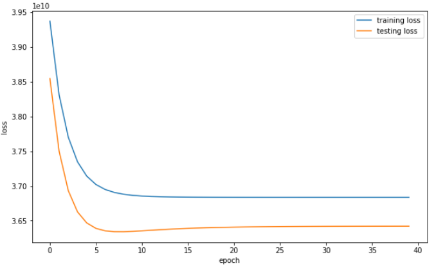


绘制学习率为3，学习率为1，学习率为0.1，学习率为0.01，四个模型的损失函数变化曲线(最大迭代轮数为250轮)

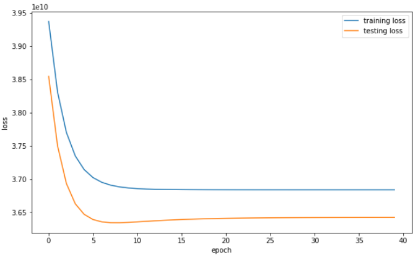


可见,选择过大的学习率，会使学习的过程产生过大扰动，而无法获得较好的学习效果.

1. **神经网络：线性回归**
2. 使用房价数据集, 基于'LotArea', 'BsmtUnfSF', 'GarageArea' 三个特征, 进行线性回归. 损失值的变化曲线如下图.

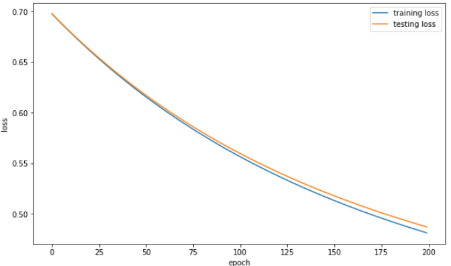


1. 对参数进行归一化处理，将其标准化，使均值为0，缩放到 [−1,1][−1,1]附近,重新训练模型，这次我们迭代40轮，学习率设置为0.1.



在测试集上的MSE=190843.24799686685

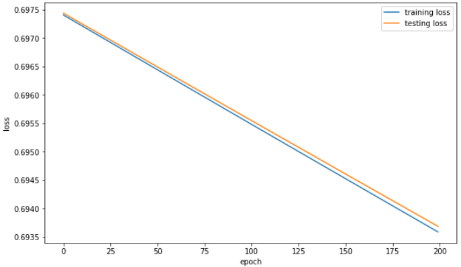
1. **神经网络：对数几率回归**
2. 使用sklearn.datasets.make\_moons生成半月形数据, 损失曲线如下:



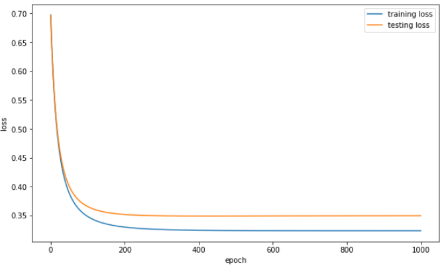
1. 初始化新的参数，学习率和迭代轮数按下表设置，绘制其训练集和测试集损失值的变化曲线，得到下表.



学习率=0.0001时损失曲线:

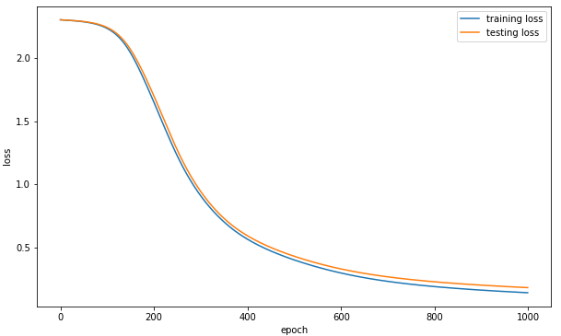


学习率=0.1时损失曲线:

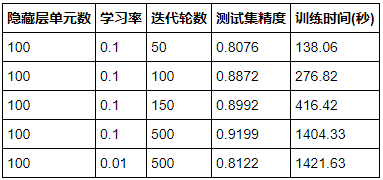


* 学习率较小时, 梯度下降步长较小, 下降很慢; 并且, 在小范围内(该图中epch=200), 梯度的变化不大, 因此呈线性下降. 但是由于它步长小,对局部最低点的逼近更加精确.
* 当学习率较大时, 梯度下降步长较大, 下降很快,但较难逼近局部最小值

1. **三层感知机**
2. 使用手写数字数据集, 设定隐藏层单元数设置为50，输出层单元数为10，学习率为0.03，迭代轮数为1000轮, 进行训练. 此模型训练耗时73.81s, 测试集上准确率94.99%; 损失值变化曲线如下.



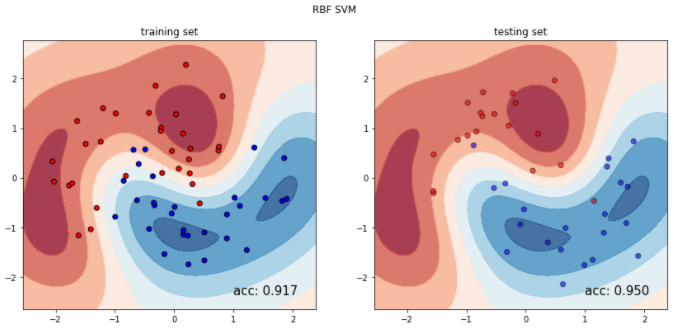
1. 使用kaggle MIIST手写数字数据集, 在不同超参数下进行训练,得到下表



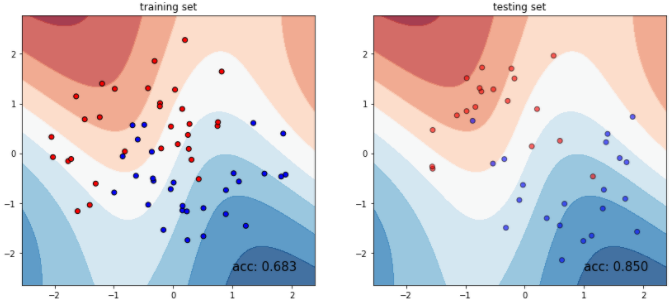
1. **SVM**
2. 支持向量机的核函数

使用sklearn.datasets.make\_moons数据集, SVM模型分别采用高斯核, sigmoid核,多项式核进行训练, 得到分类效果图.

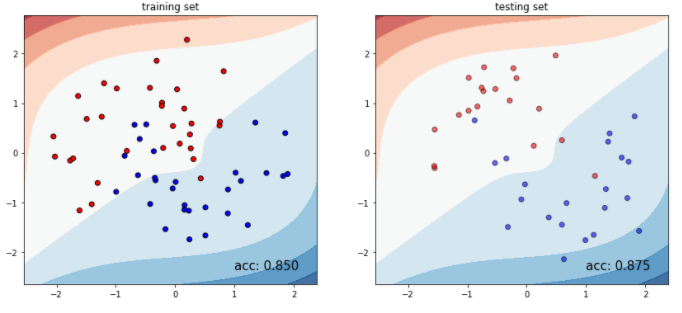
1. 高斯核



1. sigmoid核



1. **多项式核**



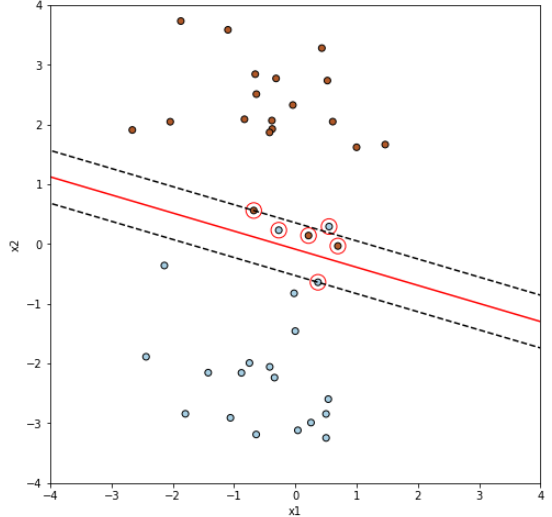
1. 引入核函数的原因

解决线性不可分问题时，需要把样本值映射到高维空间，在这个高维空间内使用分类问题的方法时，要计算向量内积，核函数可将求高维向量内积转为求低维向量内积，降低运算复杂度。

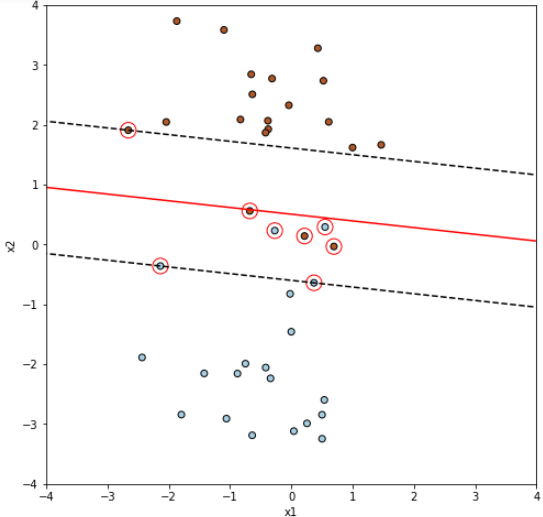
1. **支持向量机的软间隔**

数据通过随机函数生成数据集. SVM使用线性核，调整C(C=10, 1, 0.1, 0.01) 的值，绘制SVM分离超平面，间隔超平面以及支持向量.

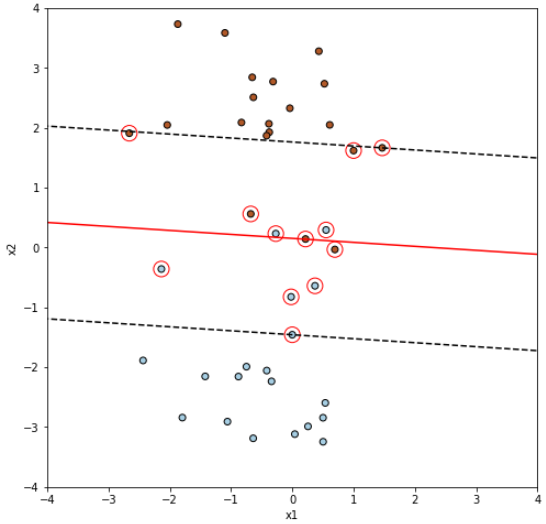
1. C=10



1. C=1



1. C=0.1



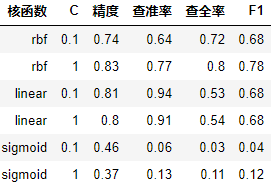
1. 加入软间隔原因

核函数使样本在更高维可分，但有出现过拟合的风险，为避免这一风险，引入软间隔，允许少量的样本分类出错，但这些样本又需尽可能的少。

1. C值对SVM影响

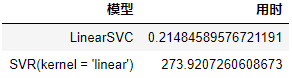
C是正则化常数，当C越大时其支持向量机的容错范围越小。

1. 使用支持向量机完成spambase垃圾邮件分类任务
2. 使用不同的核函数, 每种核函数的C值取0.1和1两种来建立模型,得到下图



1. 比较LinearSVC和SVR(kernel = 'linear')的运行时间

LinearSVC快很多, 如下图所示



1. 使用支持向量机完成kaggle房价预测问题
2. 使用不同的核函数, 每种核函数的C值取0.1和1两种来建立模型,得到下图

