**智能信息处理**

学 号：

小 组 组 长： 冯 振 兴

小 组 成 员： 魏 正 皓

叶 倚 丰

崔 玉 红

任 课 教 师： 聂 文 滨

教师所在学院： 信息工程学院

年 月

**目录**

[第1章 绪论 3](#_Toc58862184)

[1.1 上证指数开盘指数研究意义 3](#_Toc58862185)

[1.2 上证指数预测研究现状 3](#_Toc58862186)

[1.3 模型定义 4](#_Toc58862187)

[第2章 模型建立 6](#_Toc58862188)

[2.1 SVM模型 6](#_Toc58862189)

[2.2 函数间隔与几何间隔 6](#_Toc58862190)

[2.3 拉格朗日乘子法 7](#_Toc58862191)

[2.4 支持向量机回归 8](#_Toc58862192)

[第3章 Matlab实现 9](#_Toc58862193)

[3.1 数据分类的含义 9](#_Toc58862194)

[3.2 相关软件包： 10](#_Toc58862195)

[3.3 相关函数 10](#_Toc58862196)

[3.5 实验过程与结果 12](#_Toc58862197)

[第4章 总结 15](#_Toc58862198)

[4.1 各小组成员分工 15](#_Toc58862199)

[4.2小组成员的个人总结 15](#_Toc58862200)

基于支持向量回归的上证指数开盘指数预测分析

摘 要：随着股票市场投资活动的日益频繁，市场迫切需要一种有效的预测方法以帮助人们增加投资收益。本文通过SVM算法对1990年到2016年9月的上证指数进行回归预测，为了找到最优的回归参数，运用交叉验证的方法，可以找到在一定意义下最佳的参数。回归结果表明，本文的方法可以有效的拟合历年的上证指数，预测效果较为理想。

关键词：股票预测 SVM 交叉验证

第1章 绪论

1.1 上证指数开盘指数研究意义

股票市场的好坏已成为我国国民经济的晴雨表。对于我国经济的发展和社会的和谐稳定起着重要作用。从股票市场预测所面临的问题及其相应影响因素分析可知，我国的股票市场是具有典型的中国特色的、符合中国国情的、行为机制十分复杂，如果可以对股票市场进行准确的预测将会使这一切问题都迎刃而解。在预测我国股市行情时，最常运用的指标是上证指数，上证指数即上海证券交易所综合股价指数，是一种反映上海证券交易所股票总体走势的统计指标，对其进行科学预测可以更好地了解股市未来的波动情况。

目前，上证指数已发展成为包括综合股价指数、30指数、180指数、债券指数、A股指数、B股指数、分类指数等在内的股价指数系列。这些指数为投资者进行投资提供了重要的参考价值,投资者对不同品种证券进行投资时可以参照不同的证券指数。上证指数的影响指标众多。而在这些指标中，开盘价、最高盘价、最低盘价、收盘价和成交量、成交额这6个指标是对上证指数本身具有决定性影响，它们可以直观的反映出上证指数的价值属性，并可以直接影响到投资者对其是否购买的决定。其中开盘价对于投资的购买决策的影响是最大的。

1.2 上证指数预测研究现状

随着计算机和网络技术的飞速发展，非常多的学者开始研究指数预测，经过大量研究，现有股价预测理论已得到了飞速发展，目前来说比较热门的预测理论有以下几种:

(1)传统证券投资分析法:传统证券投资分析法具体的可分为技术分析法和基本分析法, 前者主要通过对国家政策、经济运行状况、行业发展情况、公司业绩等因素的分析，大致确定股票的内在价值，进而给投资者一定的投资指向性信息;而后者是一种完全依据股市行情变化而分析的方法，通过对股票市场历史数据的分析，找到单只股票乃至股市的变化规律，判断股票市场可能出现的转折点，进而给投资者提供信息指向的方法。

(2)基于时间序列的预测:所谓时间序列预测是根据被预测对象较长一段时间内发展变化的趋势和规律，建立起适当的数学模型，进而推算出被预测对象未来短时间内变化趋势的方法，它有一些典型的模型，比如ARIMA模型、RW模型、ARMA模型等[5]。但面对复杂多变的股票市场行情，这种方法仅能对较短时间内的股票价格进行有效预测，一般不超过时间序列的1/15，预测的周期越长，该方法的精准度则越低。

(3)基于神经网络的预测: 智能预测方法在股市预测中得到广泛应用并迅速占据了主导地位。神经网络凭借其优越的预测性能迅速被应用于股市预测当中。王芳等人[1]采用经BP神经网络进行了股票预测，很多学者在经典的神经网络基础上进行了改进，陈立勇等人[2]应用了粗神经网络对股市进行了预测研究，陈芳芳等人[3]引入了小波神经网络理论对时间序列做预测。王行愚等人[4]通过主成分分析等技术，改进了RBF神经网络，试验证明该模型具有收敛速度快、预测准确度高等特点。

然而神经网络自身存在难以克服缺陷，如网络结构复杂、过拟合、泛化能力不强等，证券价格预测结果与期望值之间有一定差距。支持向量机(Support Vector Machine，SVM) 是一种基于结构风险最小化原则的机器学习方法，较好地解决了非线性、过拟合和局部极值等难题，泛化能力优异，在各个邻域都有应用[8][9][10][11][12]，在经济预测领域也有研究[6][7]。本文尝试利用SVM建立一种证券价格预测模型，并采用交叉验证算法对SVM参数进行搜索，最后通过采用上证开盘指数数据对模型的预测性能进行测试，验证其有效性。

1.3 模型定义

模型目的：利用SVM建立的回归模型对上证指数每日的开盘数进行回归拟合。

模型假设：假设上证指数每日的开盘数与前一日的开盘指数，指数最高值，指数最低值，收盘指数，交易量，交易额相关，即把前一日的开盘指数，交易量，交易额作为当日开盘指数的自变量，当日的开盘指数为应变量。

算法流程如图1-1所示。

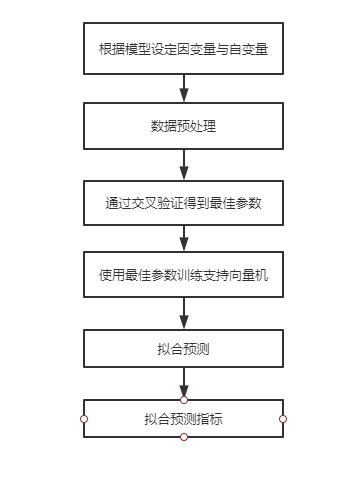


图 1-1

第2章 模型建立

2.1 SVM模型

对样本，传统的回归模型通常是基于模型输出与真实值输出之间的差别来计算损失，当且仅当与完全相同时，损失才为零。与此不同，支持向量回归假设我们能容忍与之间最多有的偏差，即当且仅当与之间的最大差别绝对值大于才计算损失。这相当于以为中心，构建了一个宽度为的间隔带，若训练样本落入次间隔带，则被认为时预测正确的。如图2-1所示

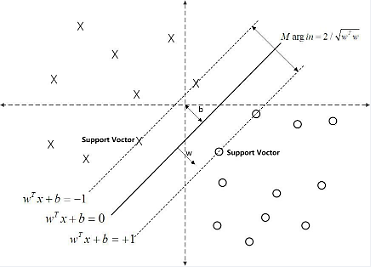


图2-1

2.2 函数间隔与几何间隔

一个点距离超平面的远近可以表示分类器预测的确信程度。在超平面确定的情况下，能够相对地表示点距离超平面的远近。而的符号与类标记的符号是否一致能够表示分类是否正确。所以在SVM中可以用来表示分类的正确性即确信程度。对于给定的训练数据集和超平面，定义超平面关于样本点的函数间隔为：

定义超平面关于训练数据集的函数间隔为超平面关于中所有样本点的函数间隔的最小值：

我们通过函数间隔可以表示分类预测的正确性及确信度，但仅仅通过函数间隔还不够，由于只要成比例地改变和，例如将其改为和，可以看出超平面本身并没有发生改变，但是函数间隔却变为原来的2倍。所以在使用SVM时，我们可以对超平面的法向量添加某些约束，例如将法向量规范化：，此时函数间隔变为几何间隔。

2.3 拉格朗日乘子法

支持向量机学习的基本想法是求解能够正确划分训练数据集并且几何间隔最大的分离超平面。对线性可分的训练数据集而言，线性可分的分离超平面有无穷多个，但是几何间隔最大的超平面是唯一的。这里的间隔最大化又称为硬间隔最大化。在二分类问题中，支持向量机会根据输入数据输出一个类别标签，对应的当分类判别函数时，输出分类标签，反之输出。在这里由于-1和+1仅仅相差一个符号，方便数学上的处理，可以通过一个统一公式来表示间隔或者数据点到分隔超平面的距离，同时不必担心数据类别标签是-1还是+1类。这样支持向量机的目标就是找出分类器定义的和，因此这里需要找到具有最小间隔的数据点即支持向量。一但找到具有最小间隔支持向量，我们就需要对该间隔最大化，如下式所示：

可以将上式进一步改写为：

此处函数间隔的值并不影响最优化问题的解，就可以取带入上面的最优化问题，于是可以进一步优化支持向量机学习的最优化问题：

对上式使用拉格朗日乘子法可得到其“对偶问题”，即对每条约束添加拉格朗日乘子，则该问题的拉格朗日函数可写为：

其中。令对和的偏导为零可得

将上两式带入拉格朗日函数将和消去，就可得到支持向量机目标函数的对偶问题：

通过以上公式可以看出当支持向量机的模型训练完成后，大部分的训练样本都不需要保留，最终的模型仅与支持向量有关。

2.4 支持向量机回归

在有了上述SVM分类的基础上，我们可以进一步推导出SVM在回归问题上的SVR模型。给定样本，我们可以学得一个形如上式的回归模型，使得与尽可能的接近，其中和同样也是模型待学习的参数。

对于样本，传统的回归模型通常直接基于输出模型输出**与**完全相同时，损失才为零。但在SVR中允许**与**之间最多有的偏差，即仅当**与**之间的差别绝对值大于时才计算损失。从而SVR问题可形式化为：

其中为正则化常数，为不敏感损失函数：

通过引入松弛变量和，可以将上式重写为：

类似于上一节中介绍的引入拉格朗日乘子的方法，通过引入乘子，由拉格朗日乘子法可以得到相应的拉格朗日函数，再令拉格朗日函数对和的偏导为零，再将得到的结果回代到上式中即可得到相应的对偶问题。最后我们可以得到SVR相应的解，如下式所示：

上式中的的样本即为SVR的支持向量，它们必落在的偏差之外。可以看出SVR的支持向量仅是训练样本的一部分，其解任然具有稀疏性。

第3章 Matlab实现

3.1 数据分类的含义

对于数据选择了上证指数(1990.12.29–2009.08.19)，这是一个的double型矩阵，记录了从1990年12月19日至2009年8月19日这段时间内4579个交易日每日上海证券综合指数的各种指标。数据的每一行表示每一天上证指数的各个指标，每一列分别表示了当天上证指数的开盘指数、指数最高值、指数最低值、收盘指数、当日交易量以及当日交易额。其中，开盘指数是指股票开市时成交的第一笔交易的价格，每日的开盘指数如图3-1所示。

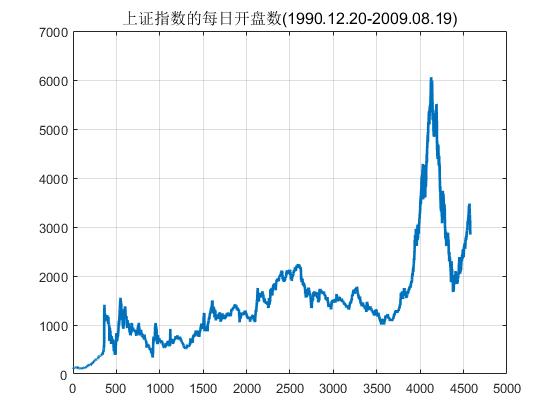


图 3-1

指数最高值与指数最低值分别指的是当日股票市场单笔交易成交价格的最高值与最低值。每日的指数最高值与每日最低值如图3-2所示。

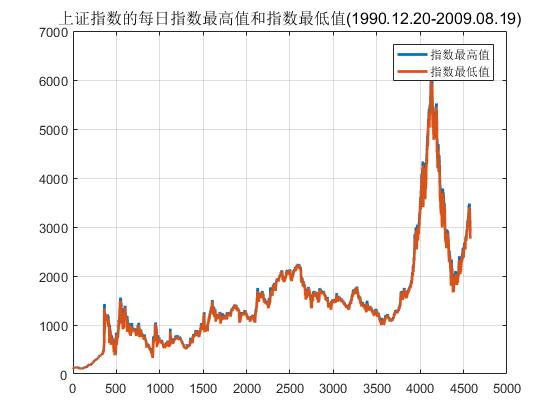


图 3-2

收盘指数是指当日指数收盘的具体数字，一般区分还在交易的盘中指数。上证指数每日的收盘指数如图3-3所示。

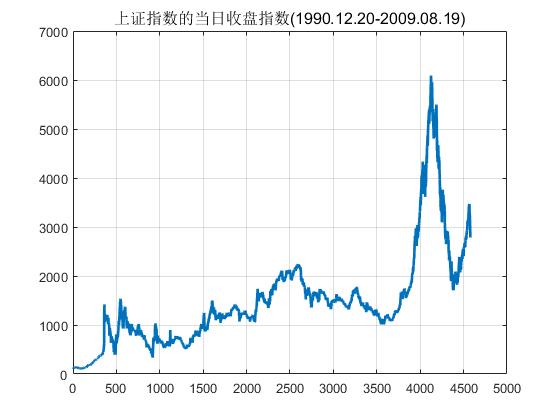


图 3-3

3.2 相关软件包：

在代码实现过程中，使用到了libsvm包，这是一个由台湾大学林智仁教授等开发设计的一个简单、易于使用和快速有效的SVM模式识别与回归的软件包。它可以解决分类问题（包括C-SVC、n-SVC），回归问题（包括e-SVR、n-SVR）以及分布估计（one-class-SVM）等问题，提供了线性、多项式、径向基和S形函数四种常用的核函数供选择，可以有效地解决多类问题、交叉验证选择参数、对不平衡样本加权、多类问题的概率估计等。

libsvm使用的一般步骤：

1. 按照libsvm软件包所要求的格式准备数据集；
2. 对数据进行简单的缩放操作；
3. 考虑选用RBF核函数*；*
4. 采用交叉验证选择最佳参数*C*与*g*；
5. 采用最佳参数*C*与*g* 对训练集进行训练获取模型；
6. 利用获取的模型进行测试与预测。

整体的算法流程图如图3-4所示：

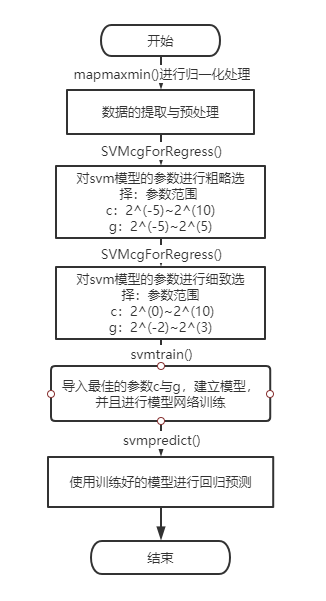


图3-4

3.3 相关函数

3.3.1 svmtrain

svmtrain函数通过实现对训练数据集的训练，实现SVM模型的获取。它的用法为

其中，为操作参数，可用的选项如下：

*-s* svm类型：设置svm的类型，默认为c-svc，其余的可选类型分别为：n-svc、one-class-svm、e-svr与n-svr；

*-t* 核函数类型：设置核函数类型，默认为RBF核函数，其余的可选类型分别为：线性核函数、多项式核函数以及sigmoid核函数；

-*d* d：核函数中的度设置，默认值为3；

*-g* g：设置核函数中的参数g，默认值为1/k；

*-c* 损失：设置c-svc、e-svr、n-svr的惩罚系数C，默认值为1；

-*n* n：设置n-svc、n-svr中的参数n，默认值为0.5；

*-h* 收缩：是否启用启发式，默认为1；

*-b* 概率估计是否计算SVC或SVR的概率估计，默认为0；

是要进行训练的数据集，是训练结束后产生的模型文件。

3.3.2 svmpreict

svmpredict函数通过训练得到的模型，对数据集进行预测。它的用法为

其中，的可用参数为-*b* probability\_estimates：是否需要进行概率估计预测，默认值为0。是由svmtrain函数训练得到的模型文件，是要进行预测的数据文件，是输出文件，表示预测的结果值。

3.3.3 SvmMcgForRegress

SvmMcgForRegress函数通过交叉验证得到最佳的模型训练参数c与g。它的用法为

其中，为训练集标签，为训练集，为惩罚参数*c*的变化范围最小值，为惩罚参数c的变化范围最的大值，为参数g的变化范围最小值，为参数g的变化范围最的大值，*v*为交叉验证的参数，默认值为1，为参数c进步的大小，默认值为1，为参数g进步的大小，默认值为1。

3.3.4 mapminmax

mapminmax函数为matlib自带的无量纲处理函数，其主要是用来对数据进行归一化处理，它把所有的数据都转换为一定区间之间的数，目的就是取消各维数据间的数量的差别，防止大数吃小数的现象。它的用法为

其中是归一化得到的数据，是描述数据的结构，其包含了一系列对归一化操作进行约束的参数，比如归一化映射的区间以及。他们的计算方式为

3.5 实验过程与结果

在具体实验过程中，首先提取数据集上证开盘指数的数据集，即进行load chapter14\_sh.mat，然后提取数据集中的相关信息。每日开盘指数、上证指数的可视化图像已在3.1节给出。

为了方便数据操作，同时消除指标之间的量纲影响，以解决数据指标之间的可比性，对元数据进行归一化操作，如图3-5所示。

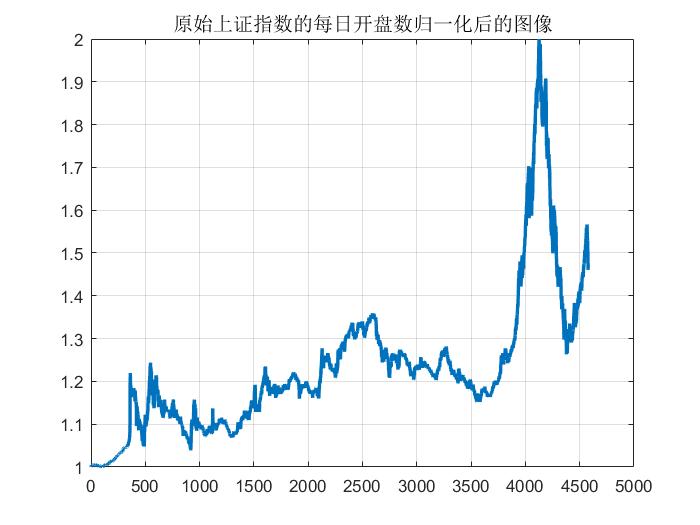


图 3-5

原始数据经过数据标准化处理之后，各指标处于同一数量级，适合进行综合对比评价。在matlab中对应的函数为:

在映射函数中将范围参数设置为1和2，即将原始数据映射到范围1-2之间。

在使用svmtrain函数时，需要对惩罚参数c惩罚参数和核参数g进行选取，在本实验中采取了一种较为简单的思想，即让c和g在某一范围内离散取值，取使得最终测试集分类的准确率最高的c和g为最佳参数。在matlab中对应的子函数为SVMcgForRegress.m。对应的参数c和g的选取过程如图3-6和3-7所示。

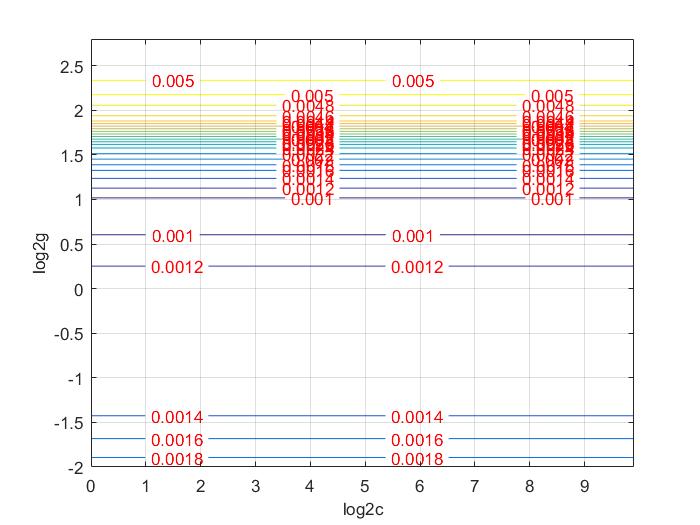


图 3-6

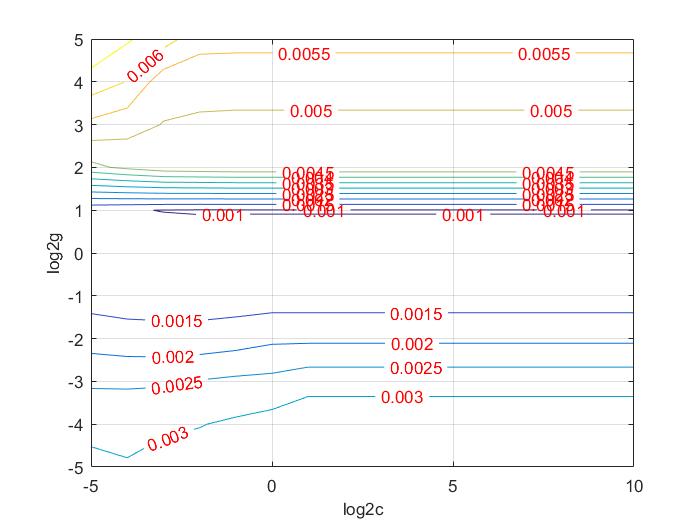


图 3-7

通过SVMcgForRegress训练得到最佳参数c和g，然后对SVM进行训练，再对原始数据进行回归预测，最终基于SVM神经网络的回归预测分析--上证指数开盘指数预测的实验结果如图3-8所示。

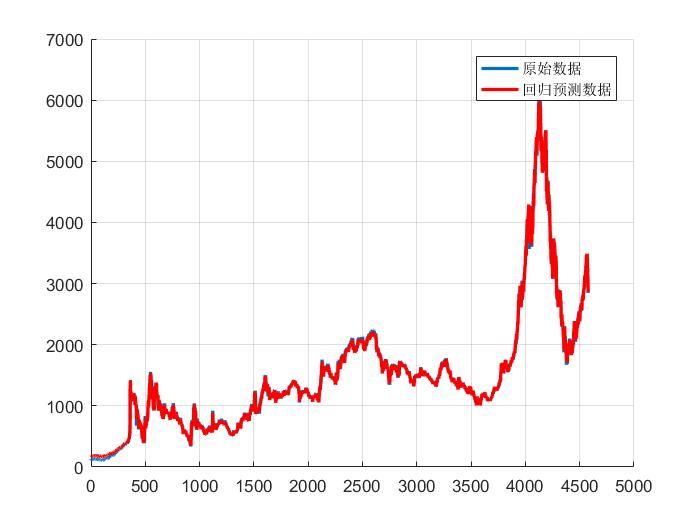


图 3-8

第4章 总结

4.1 各小组成员分工

冯振兴负责论文中的模型建立、实验过程与结果的撰写

魏正皓负责运行程序以及论文部分数据分类、相关函数的撰写

叶倚丰负责论文中的国内外研究现状的撰写

崔玉红负责论文中的问题研究意义的撰写

4.2小组成员的个人总结

4.2.1 冯振兴的个人总结

我在本次智能信息处理大作业中负责第二章模型建立、第三章中实验过程与结果部分的撰写。

通过学习支持向量机算法，认识到该算法具体的实现方式，对SVM和SVR有了一个较为深刻的认识。了解到该算法的优势在于：1）在高维空间中非常高效；2）即使在数据维度比样本数量大的情况下依然有效；3）由于算法在实际的使用中，仅依赖于支持向量，因此该算法能够高效利用内存；

实验过程中首先对数据集进行了预处理操作，通过定义子函数SVMcgForRegress.m实现交叉验证选择回归的最佳参数c&g，然后利用最佳参数训练SVM，最后拟合数据并预测。

但是在实际使用过程中也发现了一些实际问题，在选择核函数时，我们并不知道究竟那种核函数能够较好的拟合模型，因此需要通过不断的尝试去选取适合相应问题的核函数，如果特征数量比样本数量要大得多时，在选择核函数时要避免过拟合，这是正则化的作用尤为重要。

在对上证指数数据集进行处理时，首先对数据集进行了归一化处理，这是由于特征向量中的不同特征往往具有不同的量纲和量纲单位，会影响到数据分析的结果，对数据进行归一化作用能够较好的消除量纲之间的影响，同时归一化操作还能较好的消除奇异样本数据导致的不良影响。

通过这次大作业，让我进一步熟悉了论文的撰写，掌握了matlab实现机器学习中的经典学习算法SVM。了解了数据处理的作用和方法，对我未来的研究生生涯有了较为深刻的启发。

4.2.2 魏正皓的个人总结

在这次的大作业中，本人负责代码的调试与运行、关于论文中数据分类部分以及libsvm库中相关函数介绍的资料查找与撰写。

首先对于数据集部分，数据集是采用了上海证券交易所从1990年12月20日到2009年8月19日这二十年间关于开盘指数、指数最高值、指数最低值、收盘指数、当日交易量以及当日交易额。共有4579条数据，全部都是连续性数据。通过对这些数据进行可视化分析，可以发现对于开盘指数、收盘指数、指数最大值与指数最小值，他们的变化趋势非常的相似，可以说明这些特征具有高度的相关性。在日常的研究中，若需要对特征之间的相关性进行分析，通常来说直接观测的方法是没有依据的，一般的方法是使用相关性指标进行评估，例如常用的Pearson相关系数，适用于对连续性变量进行相关性分析。分类变量分析则可以使用卡方检验。

然后是关于代码的运行与调试。代码的运行与调试都是在MATLAB -R2016a的环境下进行。对于运行代码，最主要的问题是在库的安装。由于代码中使用的是第三方库libsvm，需要对其进行安装。并且libsvm是一个基于C的库，还需要安装相关的编译器，操作较为复杂 (libsvm的详细信息可见正文3.2节) 。代码的运行没有太大的问题，可以得到与教程近乎相同的结果。我在原代码的基础上增加了对数据的可视化描述，以便于对数据进行观察分析。

再是是关于使用到的函数，代码中主要使用到的函数有libsvm库中的，有MATLAB库中自带的，也有代码中自己定义的函数。svmtrain与svmpredict为svmlib库中的函数。前者输入数据为预处理过的数据集，输出的则是训练好的模型。在该函数中，可以对支持向量机的类型、核函数的类型以及一些细节进行指定。后者则是可以将输入数据通过模型进行运算，得到相应的结果。这两个函数内部都在调用libsvm中的svm.h的方法，svm.h内部封装了关于支持向量机的一系列基本函数。mapminmax为MATLAB自带的归一化函数，在此实验中，它将所有的数据进行处理，将其规划到1到2的范围内，方便了后面数据的收敛。SVMcgForRegress函数为自定义函数，其通过交叉验证的方式得到最佳的参数。

最后是关于此次实验与论文撰写的心得体会。支持向量机作为一种监督学习方式，历经了几十年的发展，目前已经高度成熟，并且也拓展出了支持向量回归，支持向量聚类等。这是我在以前学习了支持向量机的相关知识之后第一次将其与实际运用，使我对其有了更加深刻的理解。但是也使我意识到自己的学识还不够丰富，需要在接下来的研究生生涯中多加补充。

4.2.3 叶倚丰的个人总结

在本次SVM预测上证指数开盘指数的论文撰写中，我了解了上证指数和开盘指数的意义，上证指数指上海证券交易所的大盘指数，而开盘指数指的是第一笔交易的成交价格。学习到了SVM支持向量机的基本概念以及运作原理等，在国内外研究的撰写中，查找了相关的文献，包括IEEE文献，浙江大学的论文等，在查看资料的过程和撰写论文中也学到了该预测算法的设计及流程。了解了预测回归模型的建立过程，将前一天的开盘指数，指数最高值，指数最低值，收盘指数，交易量，交易额六个值作为自变量，当日的开盘指数作为因变量，对股票进行预测，预测股票指数。在论文的撰写中，除了学习到了新的SVM及相关知识，我还练习了论文的撰写，文献的查找，格式的调整等等。

这次论文的撰写使我受益颇多，让我深刻的意识到我所掌握的知识还很有限，还需要在平时多多积累知识，为以后的学习生活奠定基础。

4.2.4 崔玉红的个人总结

在本次学习中我了解到SVM不仅可以解决分类问题，还可以做回归预测分析，这一章节主要是以股票的上证指数开盘指数为例，通过对上证指数从每日实际的开盘数进行回归分析，最后得到相应的均方误差和平方相关系数。同时了解到上证指数是国内外普遍采用的用于反应上海证券市场总体趋势指标。开盘是指某种证券在证券交易所每个营业日的第一笔交易,第一笔交易的成交价即为当日开盘价按上海证券交易所规定。

本章以一个完整的SVM回归预测分析过程呈现，一个预测分析首先要进行模型的建立，然后MATLAB实现需要根据模型假设选定自变量和因变量，再进行数据预处理，参数选择是一个重要过程，对模型建立好坏起重要作用，最后进行训练模型和回归预测。股票指数是一个非常灵活的数据，要求精确度比较高，他的预测可变程度也比较大，但可以做一个短期的预测，对股票投资者还是具有非常大的参考价值。

通过本章学习，对于SVM的作用有了一个新的认识，同时了解到预测分析的具体流程，对于以后模型的建立有很大帮助，同时加深了自己对股票参数的一些了解。

参考文献

1. 王芳. 基于支持向量机的沪深300指数回归预测[D].山东大学,2015.
2. 陈立勇. 改进最小二乘支持向量机及其应用[D].华东交通大学,2014.
3. 陈芳芳. 基于支持向量机的股票价格预测研究[D].重庆理工大学,2016.
4. 王行愚.RBF神经网络在股市趋势预测中的应用[J].华东理工大学学报,2002(05):547-550.

1. [Some simulations and applications of forecasting long-memory time-series models](https://kns.cnki.net/kcms/detail/detail.aspx?filename=SJES13012100590227&dbcode=SJES&dbname=scholar_journal_SJES" \t "kcmstarget)[J] . Valdério Anselmo Reisen,Silvia Lopes.  Journal of Statistical Planning and Inference . 1999 (1)
2. Cao L，Tay F. Financial forecasting using support vector machines . NeuralComputing&Applications. 2001;10(2):184-192.
3. Tay F，Cao L. Application of support vector machines in financial time seriesforecasting. Omega. 200 1; 29:309—3 17.
4. 徐海峰,朱晶,王艳波.基于SVM法的大坝安全稳定预测模型分析及应用[J].黄河水利职业技术学院学报,2020,32(04):7-11.
5. Jose C, Goyal P, Aggrwal P, et al. Local deep kernel learning for efficient non-linear svm prediction[C]//International conference on machine learning. 2013: 486-494.
6. Singh S, Parmar K S, Makkhan S J S, et al. Study of ARIMA and least square support vector machine (LS-SVM) models for the prediction of SARS-CoV-2 confirmed cases in the most affected countries[J]. Chaos, Solitons & Fractals, 2020, 139: 110086.
7. Shi Y,Eberhat R C. A modified particle swarm optimizer [C].IEEE InternationalConference of Evolutionary Computation, Anchorage,Alaska,May 1998.
8. Shi Y,Elberhart R C. Fuzzy adaptive particle swarm optimization [A]. Proceeding ofCongress on Evolutionary Computation[C]. Seoul, Korea, 2001.