第1章 绪论

1.1 绪论(介绍支持向量机)

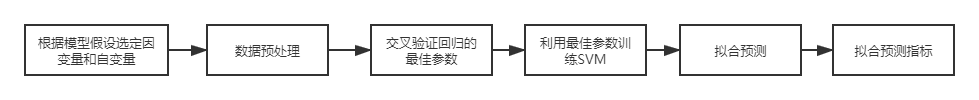
支持向量机是一种二分类模型。它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；支持向量机还包括核技巧，这使它成为了实质上的非线性分类器。支持向量机的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的损失函数的最小化问题。

1.2 问题定义

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。

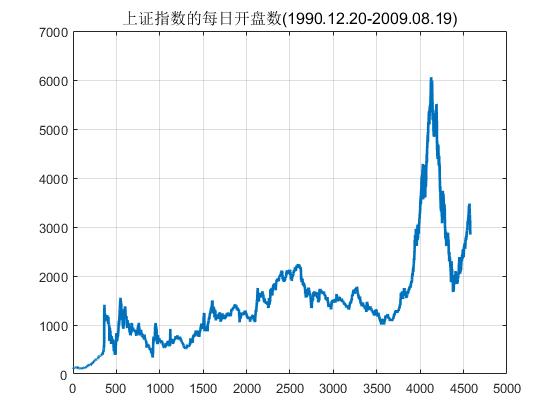
模型假设：假设上证指数每日的开盘数与前一日的开盘指数，指数最高值，指数最低值，收盘指数，交易量，交易额相关，即把前一日的开盘指数，交易量，交易额作为当日开盘指数的自变量，当日的开盘指数为应变量。

算法流程如图所示。

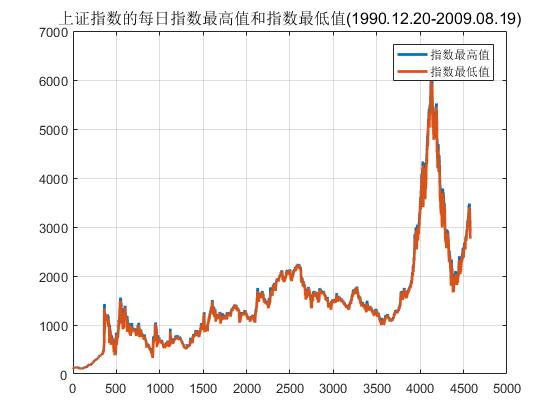


1.3 数据分类的含义

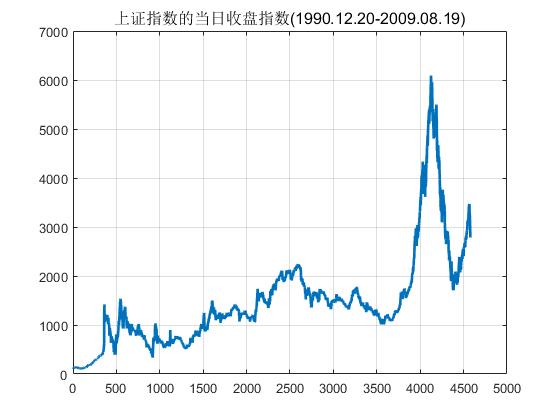
对于数据选择了上证指数(1990.12.29–2009.08.19)，这是一个的double型矩阵，记录了从1990年12月19日至2009年8月19日这段时间内4579个交易日每日上海证券综合指数的各种指标。数据的每一行表示每一天上证指数的各个指标，每一列分别表示了当天上证指数的开盘指数、指数最高值、指数最低值、收盘指数、当日交易量以及当日交易额。其中，开盘指数是指股票开市时成交的第一笔交易的价格，每日的开盘指数如图-所示。



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-所示。



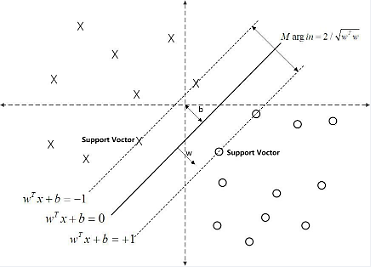
收盘指数是指当日指数收盘的具体数字，一般区分还在交易的盘中指数。上证指数每日的收盘指数如图-所示。



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(Support Vector Regression)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。



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：

类似于上一章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

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2 K(x,y) e x y = -g - ；
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。

使用到的函数

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，options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 度：核函数中的度设置，默认值为3

-g g：设置核函数中的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

training\_set\_file是要进行训练的数据集，model\_file是训练结束后产生的模型文件。

1. 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，options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。Model\_file是由svmtrain函数训练得到的模型文件，test\_file是要进行预测的数据文件，output\_file是输出文件，表示预测的结果值。

1. 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，train\_label为训练集标签，train为训练集，cmin为惩罚参数c的变化范围最小值，cmax为惩罚参数c的变化范围最的大值，gmin为参数g的变化范围最小值，gmax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，cstep为参数c进步的大小，默认值为1，gstep为参数g进步的大小，默认值为1。