华鑫期货
CHINA FORTUNE FUTURES

机器学习 第三期:【支持向量机模型】SVM

2020 年 1 月 16 日

量化策略主题报告


**崇盛棠** 高级分析师

投资咨询编号:Z0010966

从业资格编号:F0257938

邮箱:chongst@shhxqh.com

**田景山** 数据分析师

投资咨询编号:Z0015457

从业资格编号:F3045626

邮箱:tianjs@shhxqh.com

**施杨** 分析师助理

从业资格编号:F3057470


## ■ 支持向量机(SVM)介绍

支持向量机(Support Vector Machine,SVM)是由 Vapnik 等人在 1995 年提出来的一种被公认的比较优秀的分类模型。SVM 是一类按监督学习方式对数据进行二元分类的广义线性分类器,其决策边界是对学习样本求解的最大边距超平面。

## ■ 支持向量机(SVM)原理

超平面和最近的数据点之间的间隔被称为分离边缘,用 P 表示。SVM 的目标是找到一个特殊的超平面,使得这个超平面分离边缘 P 最大。在这个条件下,决策曲面称为最优超平面。基本上,SVM 的思想是建立在两个数学运算上,一个是输入向量到高维特征空间的非线性映射,另一个是构造一个最优超平面用于分离在上一步中发现的特征。

## ■ 支持向量机(SVM)应用

SVM 是应用最广泛的机器学习算法之一,它基于统计学习理论的模式识别,已经广泛应用数据挖掘、生物信息学、文本和手写识别等领域。在金融市场中,已有金融机构将 SVM 应用在智能选股,择时策略中,金融产品的价格通常有上涨和下跌两类,可以用两类支持向量机来预测。

## ■ 风险提示

机器学习量化策略的结果是对历史经验的总结,存在失效的可能。


**关于我们:**

客服电话:400-186-8822

公司地址:上海市黄浦区福州路 666 号
华鑫海欣大厦 21、22 楼

公司官网:http://www.shhxqh.com

# 内容目录：



# 图表目录：

本报告对机器学习中的一类算法---支持向量机（Support Vector Machine，SVM）进行探索。支持向量机是由 Vapnik 等人在 1995 年提出来的一种被公认的比较优秀的分类模型。SVM 是一类按监督学习方式对数据进行二元分类的广义线性分类器，其决策边界是对学习样本求解的最大边距超平面。SVM 是应用最广泛的机器学习算法之一，它已经广泛应用数据挖掘、生物信息学、文本和手写识别等领域。在金融市场中，已有金融机构将 SVM 应用在智能选股，择时策略中，金融产品的价格通常有上涨和下跌两类，可以用两类支持向量机来预测。

## 1、SVM 模型推导

### 1.1 线性可分决策面

线性可分的数据分布简单，可以找到一个超平面，直接在原始空间中将数据进行切分。设样本集为：

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

其中，$y$ 的取值为{+1, -1}。

**决策面方程**

分类的手段是取得一个符合条件的决策平面，平面可以用方程表示：

$$w^T x + b = 0$$

问题的解决，转化为找到符合条件的 $w$ 和 $b$，条件是：使得数据集边缘若干点，到这个平面的距离是最长的。

因两类数据分布在决策平面的两侧，所以，$y = -1$ 的样本点所在区域可表示为：

$$w^T x + b < 0$$

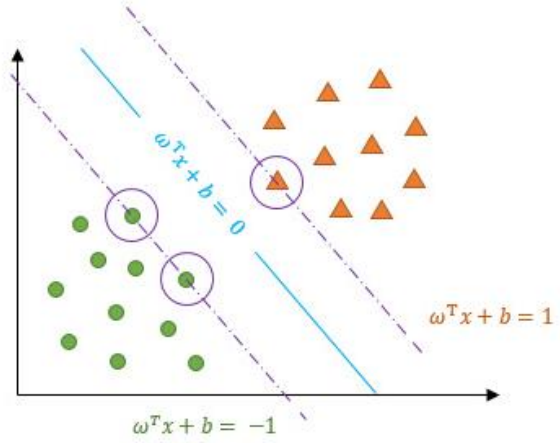
所以，$y = +1$ 的样本点所在区域可表示为：

$$w^T x + b > 0$$

那么支持向量所在的平面可以表示为：$w^T x + b = \pm A$

在后面的优化中，其实 $A$ 的结果并不影响结果，为计算方便，令 $A = 1$，即 $w^T x + b = \pm 1$，如图一所示：

图一：线性可分支持向量所在超平面图




*资料来源：华鑫期货研究所*

**计算最长距离**

有了决策面，接下来是计算支持向量到决策平面的距离，当该距离最长时所得到的参数，就是所需要的解。

由几何知识可知，点到平面的距离可以表示为：

$$\gamma = \frac{w^T x + b}{||w||}$$

此距离成为几何距离，存在正负。而算法的目标，是找到一个集合 $x$ (支持向量)，使得 γ 的值最小。因 $y$ 的取值为{+1, -1}，所以上式乘以 $y$，使该距离恒为正。所以最大距离可表示为：

$$\gamma_{max} = max\ (\frac{y(w^T x + b)}{||w||})$$

结合支持向量所在方程：

$$w^T x + b = \pm 1$$

故：

$$y(w^T x + b) = 1$$

则最大距离简化为：

$$\gamma_{max} = max\ (\frac{1}{||w||})$$

求解上式的最大值，等同于求解下式的最小值：

$$min\ (\frac{1}{2}||w||^2)$$

这里增加了一个 1/2 系数、和一个平方，是为了方便求导。一求导两者就相消了。这个式子有还有一些限制条件，完整的写下来，应该是这样的：

$$min\ (\frac{1}{2}||w||^2),\ s.t,\ y(w^Tx+b)\geq 1$$

*s.t.*后面的限制条件可以看作是一个凸多面体，我们要做的就是在这个凸多面体中找到最优解。求解该式，可以用拉格朗日乘子法去解，使用了 KKT 条件的理论。

**求最优解**

待求解式子的拉格朗日目标函数如下:

$$L(w,b,a)=\frac{1}{2}||w||^2-\sum_{i=1}^{n}\alpha_i(y_i(w^Tx+b)-1)$$

目标是让 $L(w,b,a)$针对 $a$ 达到最大值。

如何求解?

$L$ 是 $w$、$b$、$a$ 三个变量的函数，要得到使得 $L$ 最大的 $a$，需进行两步操作:

一、需要先排除掉 $w$ 和 $b$ 的影响，让 $L$ 关于 $w$、$b$ 最小化。如此一来，不管 $w$、$b$ 如何改变，$L$ 都不会再变小;

在可导的情况下，极值在导数为 0 的位置，令 L 关于 $w$、$b$ 的偏导数为 0，即:

$$\frac{\partial \mathrm{L}}{\partial \mathrm{w}}=0\ \text{和}\ \frac{\partial \mathrm{L}}{\partial \mathrm{b}}=0$$

求解上面的导数，得到:

$$w=\sum_{i=1}^{n}\alpha_i y_i\,x_i$$

$$\sum_{i=1}^{n}\alpha_i y_i=0$$

将上量式代入 $L(w,\ b,\ a)$，得到对偶问题的表达式

$$L(w,b,a)=\frac{1}{2}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

于是新的目标问题及限制条件为(对偶问题):

$$\begin{cases} max_a\left(\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j\right) \\ s.t.,\alpha\geq 0,\ \ i=1,2,\dots,n \\ \sum_{i=1}^{n}\alpha_i y_i=0 \end{cases}$$

这就是需要最终优化的式子，只关于α向量的式子。

二、接着再让 $L$ 关于 $a$ 取最大值。

上式最终的对偶问题，是一个凸二次规划问题，理论上用任何一个解决凸二次规划的软件包都可以解决。一般使用 SMO 算法，输入是样本，输出是 $a$。

SMO 基本思想，是不断执行如下两个步骤直至收敛:

- 选取一对参数$(\alpha_i,\ \alpha_j)$
- 固定α向量的其他参数，将$(\alpha_i,\ \alpha_j)$代入上述表达式进行求最优解获得更新后的$(\alpha_i,\ \alpha_j)$

解出α后，$w$、$b$ 也就确定下来，进而能得到决策面。

### 1.2 线性不可分决策面

上面讨论了线性可分的数据集的处理方式。但是，实际应用中的数据样本，可能更多的是线性不可分的，即不能找到一个可以将数据分类的超平面。

一般可以使用核函数将原始空间映射到一个高维空间，在高维空间对数据进行划分。理论上只要维度足够高，那么总能做到线性分类。

**决策面方程**

在线性可分的基础上，将样本 $x$ 进行一次变换，得到 $\phi(x)$，超平面变为:

$$w^T\phi(x)+b=0$$

**求最优解**

整个推导过程，与线性可分的基本一样，唯一不同的，是将各个公式中的 $x$，换成 $\phi(x)$。即得到对偶问题:

$$\begin{cases} max_a\left(\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\phi(x_i)^T\phi(x_i)\right) \\ s.t.,\alpha\geq 0,\ \ i=1,2,\dots,n \\ \sum_{i=1}^{n}\alpha_i y_i=0 \end{cases}$$

令　　$\phi(x_i)^T\phi(x_j)=K(x_i,x_j)$

这个式子所做的就是将线性的空间映射到高维的空间，这里的 K 函数就为核函数，核函数有很多种，比较典型的几种如下:

•线性核: $K(x_i,x_j)=x_i^Tx_j$

•多项式核：$K(x_i, x_j) = (x_i^T x_j)^d$， d≥1 为多项式的次数

•高斯核：$K(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right)$，σ>0 为高斯核的带宽

•Sigmoid 核：$K(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||}{2\sigma}\right)$，σ>0

•拉普拉斯核：$K(x_i, x_j) = \tanh(\beta x_i^T x_j + \theta)$，tanh 为双曲正切函数，$\beta$>0，$\theta$<0

最后一样能通过例如 SMO 算法等，实现求解。

## 2、SVM 模型对期货价格涨跌方向预测的实证分析

### 2.1 实现工具选择

选择 Python 第三方机器学习算法库 Scikit-learn(简记 Sklearn)的 svm 程序包。现分别采用支持向量机中线性分类器和分线性分类器，并对于两种分类器下各个参数的训练和测试效果的影响做实证分析。

### 2.2 数据集选取

采用近三年沪深 300 股指期货三个品种连续主力合约近三年(2017/1/14~2020/1/14)的行情数据，共计 730 个。其中前 80%的数据作为训练样本共 584 个，后 20%的作为测试样本共 146 个。

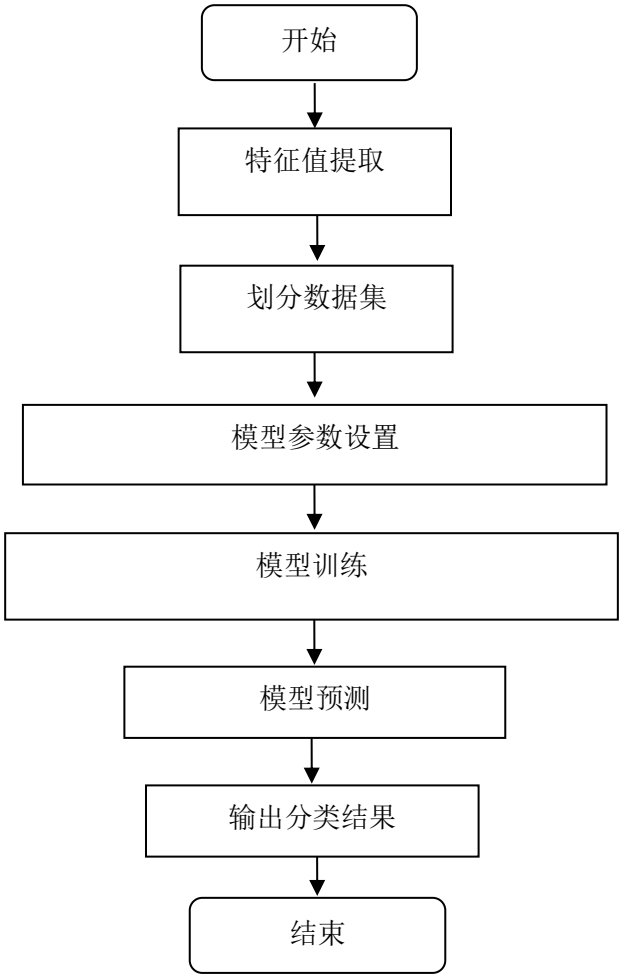
### 2.3 特征值提取

输入特征变量选择开盘价、收盘价、最高价、最低价、成交量、持仓量 6 个特征。对于训练样本，增加一列标签值，假如该天收盘价高于上一天收盘价，则意味价格上涨，标记为 1，假如该天收盘价低于上一天收盘价，则意味着价格下跌，标记为-1，不涨不跌标记为 0。

### 2.4 实证流程

期货价格预测流程如图二所示：

图二：SVM 模型预测流程图

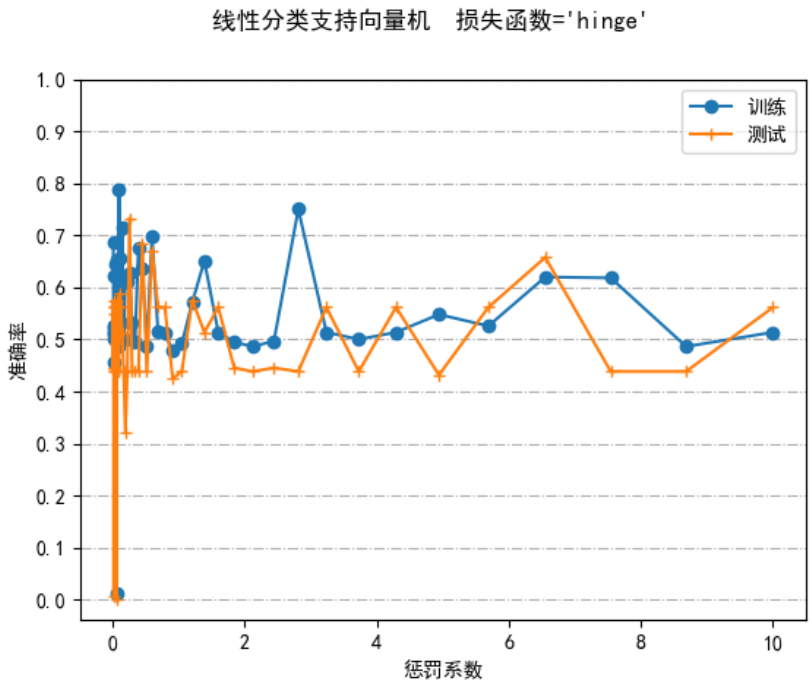



*资料来源：华鑫期货研究所*

### 2.4.1 线性分类 SVM 实证分析

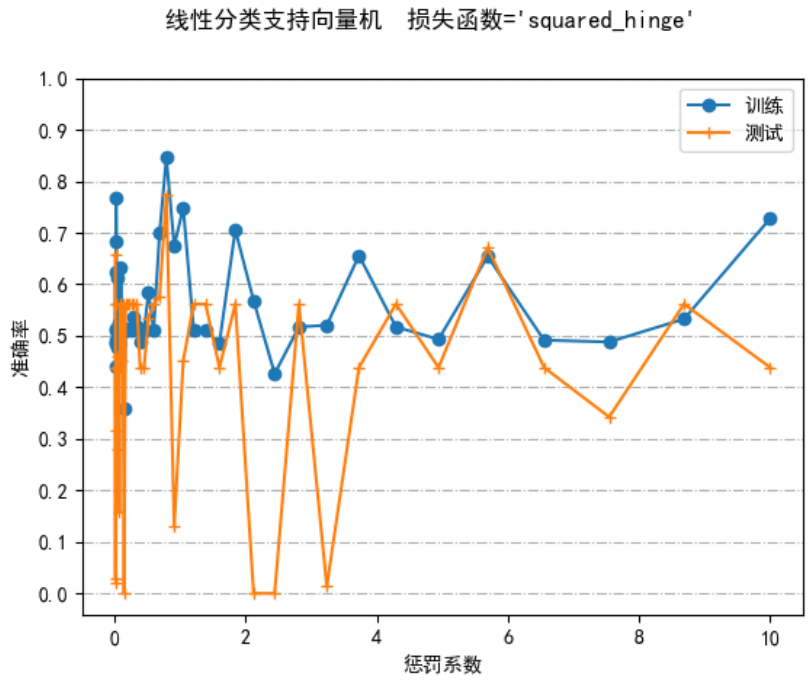
线性分类支持向量机中，损失函数可以有 hinge 和 squared_hinge，当值为 hinge 时表示合页损失函数（它是标准 SVM 的损失函数）；值为 squared_hinge 时表示为合页损失函数的平方。现在对其它参数取默认时，两种损失函数对训练和测试准确率的影响，如图三、四所示：

图三：hinge 损失函数下惩罚系数对模型准确率影响




*资料来源：华鑫期货研究所*

图四：squared_hinge 损失函数下惩罚系数对模型准确率影响



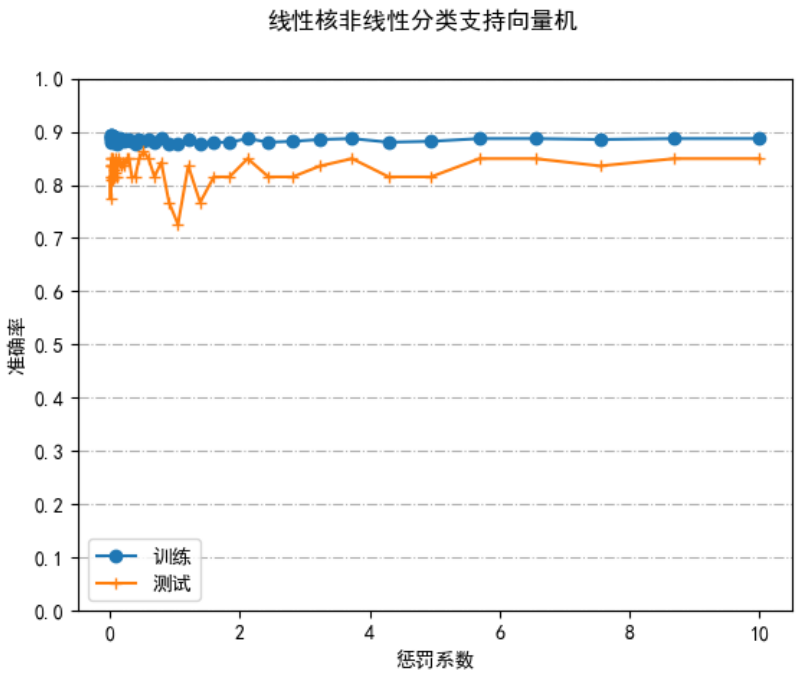

*资料来源：华鑫期货研究所*

从图中看出，线性分类器的损失函数对训练和预测精度影响不大，惩罚系数 C 衡量了误分类点的重要性，C 越大则误分类点越重要。

### 2.4.2 非线性分类 SVM 实证分析

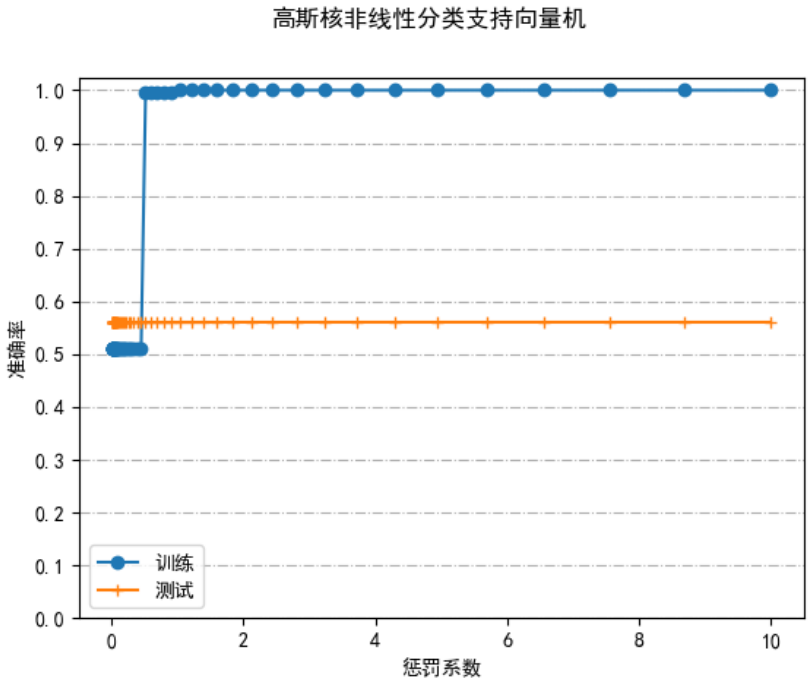
非线性分类 SVM，其核函数一般有线性核、多项式核、高斯核、sigmoid 核，现采用常用的高斯核与线性核，并分别分析参数对训练和测试准确率的影响做分析，效果分别如图五、六、七。

图五：线性核非线性分类 SVM 惩罚系数对模型准确率影响分析



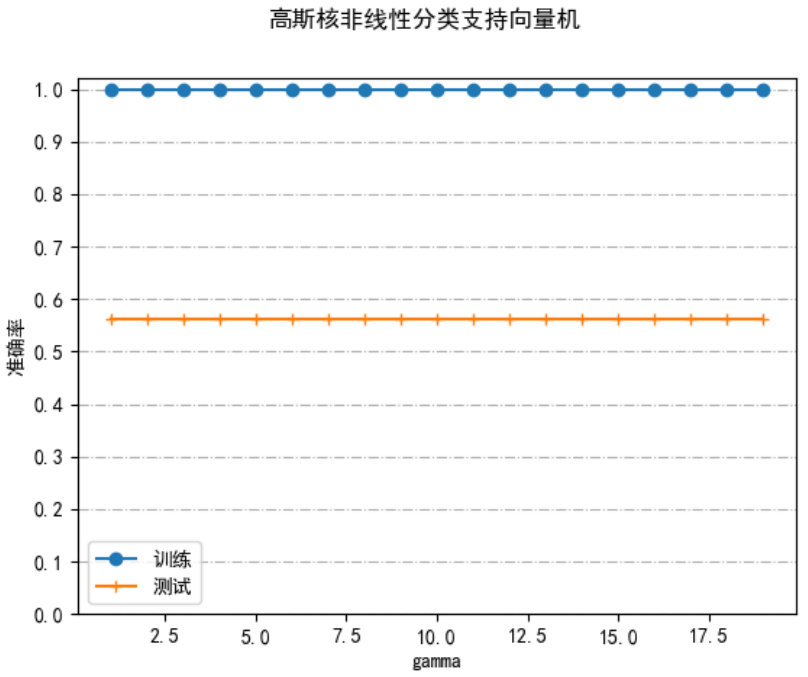

*资料来源：华鑫期货研究所*

图六：高斯核非线性分类 SVM 惩罚系数对模型准确率影响分析




*资料来源：华鑫期货研究所*

图七：高斯核非线性分类 SVMgamma 系数对模型准确率影响分析




*资料来源：华鑫期货研究所*

从图中看出，采用了线性核函数的非线性分类器训练和测试效果比线性分类器要好。非线性分类器对测试集的预测效果较为平稳。

## 3、交易策略

### 3.1 规则描述

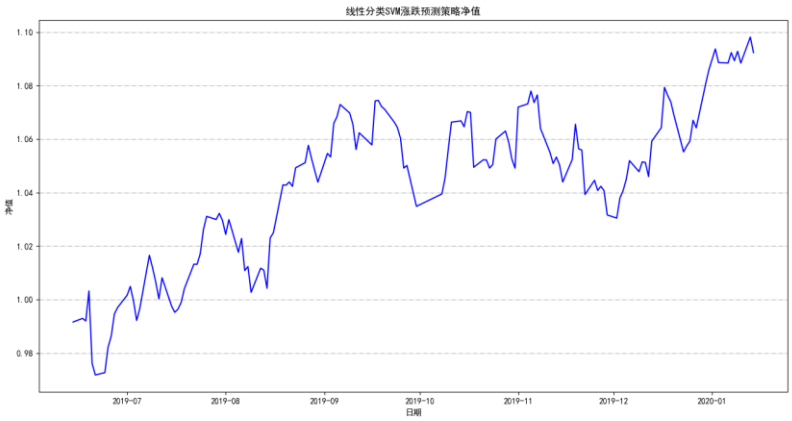
如果预测明收盘价涨，就以明开盘价买入开仓，明收盘价卖出平仓；

如果预测明收盘价跌，就以明开盘价卖出开仓，明收盘价买入平仓；

### 3.2 策略实现与评价

由实证过程可以看出，非线性分类线性核 SVM 的预测最好，非线性分类器对测试集的预测效果较为平稳。现分别对线性非线性分类器、线性核高斯核 SVM 进行策略实现，交易合约选取股指期货主力连续合约，回测时间 2019 年 6 月 14 日至 2020 年 1 月 14 日，回测结果净值曲线如图八、九、十:

图八：线性分类 SVM 预测交易策略净值




*资料来源：华鑫期货研究所*

测试总数:146，正确数量:82

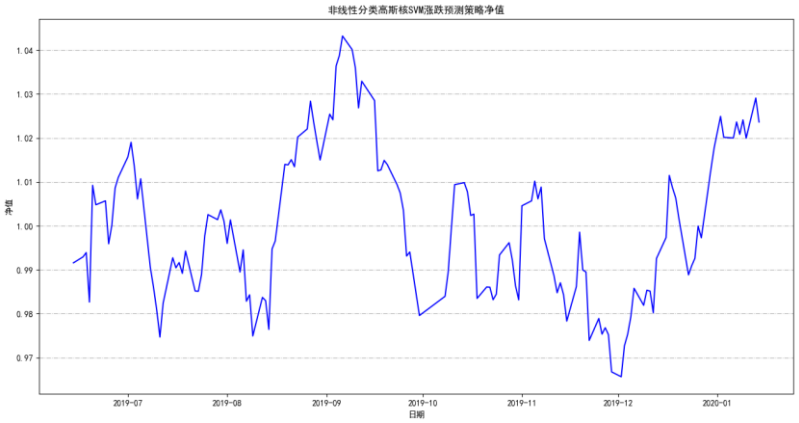
准确率:0.5616

年化收益率: 0.175

最大回撤: -0.049

夏普率: 1.138

图九：非线性分类高斯核 SVM 预测交易策略净值




*资料来源：华鑫期货研究所*

测试总数:146，正确数量:82

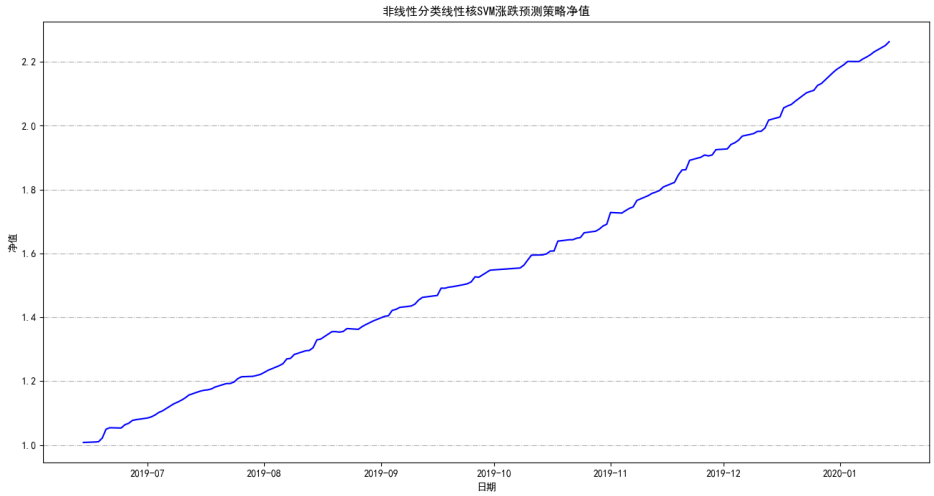
准确率:0.5616

年化收益率: 0.062

最大回撤: -0.074

夏普率: 0.187

图十：非线性分类线性核 SVM 预测交易策略净值




*资料来源：华鑫期货研究所*

测试总数:146，正确数量:118

准确率:0.8082

年化收益率: 1.390

最大回撤: -0.003
夏普率: 16.725

**4、总结**

支持向量机模型中，有较多参数需调，对于线性非线性分类器的选择，及核函数、惩罚系数、gamma 系数等都对预测的效果会产生影响。如果不能做好选择，可能得不到理想的分类效果，另外也可能有过拟合的现象。在实践中需要模型原理理解充分，对参数不断地进行尝试。

■ **分析师声明**

崇盛棠声明，本人具有中国期货业协会授予的期货从业资格，勤勉尽责、诚实守信。本人对本报告的内容和观点负责，保证信息来源合法合规、研究方法专业审慎、研究观点独立公正、分析结论具有合理依据，特此声明。

■ **本公司具备期货投资咨询业务资格的说明**

华鑫期货有限公司（以下简称"本公司"）经中国证券监督管理委员会核准，取得期货投资咨询业务许可。本公司及其投资咨询人员可以为期货投资人或客户提供期货投资分析、预测或者建议等直接或间接的有偿咨询服务。发布期货研究报告，是期货投资咨询业务的一种基本形式，本公司可以对期货及期货相关产品的价值、市场走势或者相关影响因素进行分析，形成投资评级等投资分析意见，制作期货研究报告，并向本公司的客户发布。

■ **免责声明**

本研究报告由华鑫期货有限公司撰写，报告中的信息均来源于已公开的资料，我公司尽可能保证可靠、准确和完整，但并不保证报告所述信息的准确性、完整性和时效性，也不保证我公司做出的建议不会发生任何变更。在任何情况下，我公司不就本报告中的内容对任何投资做出任何形式的担保，本报告不能作为道义的、责任的和法律的依据或者凭证，对于本报告中提供信息所导致的任何直接或间接投资盈亏不承担任何责任。



华鑫期货有限公司 研究所
公司地址：上海市黄浦区福州路 666 号华鑫海欣大厦 21、22 楼
公司官网：http://www.shhxqh.com
咨询电话：400-186-8822