长江证券 CHANGJIANG SECURITIES

机器学习

机器学习白皮书系列之一：监督学习的方法介绍及金融领域应用实例

2017-7-27

金融工程 | 专题报告

分析师 覃川桃
☎ （8621）68751782
✉ qinct@cjsc.com.cn
执业证书编号：S0490513030001

联系人 陈洁敏
☎ （8621） 68751787
✉ chenjm5@cjsc.com.cn

联系人 杨靖凤
☎ （8621）68751636
✉ yangjf@cjsc.com.cn

相关研究

《陆港通系列(一)：外资动向中的 Alpha》2017-7-24

《富时中国 A50 指数投资价值分析》2017-6-8

《基金的绩效归因方法分析及应用》2017-6-6

报告要点

■ 机器学习系列报告

本系列报告试图系统全面性的介绍各种不同的机器学习方法，并且结合具体的在投资研究领域应用实例、交易策略及 code 示例，说明其应用情景和实现方法。机器学习的方法可以分为以下几类：监督学习、无监督学习、深度学习及其他机器学习方法（例如强化学习），对应到具体的模型上数量则更是繁多，目前大部分机器学习模型并未广泛的应用在投研领域，因此本系列主要偏重于在投研领域有应用潜力的模型及方法。此篇将以介绍监督学习方法为主。

■ 监督学习模型之回归类模型及其应用

与普通线性回归不同，监督学习中的惩罚回归模型和非参数回归，可以分别用于处理输入变量中存在大量线性相关性关系及非线性关系时的情况。惩罚回归模型中金融领域使用得较多的有 Lasso 回归、岭回归和弹性网络回归；具有代表性的非参数回归模型则有：K 最近邻、LOESS 及卡尔曼滤波器。同时，也用到两个实例来说明了惩罚回归模型在拟合中的优势，以及卡尔曼滤波器使用时对于趋势判断、状态分辨的灵敏性。

■ 监督学习模型之分类模型及其应用

回归模型可以通过模型拟合进行样本外数据预测，得到具体的预测值。但是在金融领域很多问题不需要得到具体的值，得到目前的状态类型或者相对强弱位置即可。因此，分类模型应用非常广泛。此篇，我们将介绍以下分类算法：逻辑回归、支持向量机（SVM）、决策树、随机森林以及隐马尔可夫模型。前面四种模型我们将会给出具体的择时和选股上的实例，隐马尔科夫模型我们则是验证其对于国内 A 股市场的状态划分是否有效。

■ 监督学习模型的总结和比较

我们介绍的几种模型的共同特点是模型中都会要求有一个训练期（样本内）和预测期（样本外），通过训练期来找到最优参数，拟合非线性关系，然后在预测期内进行应用。不同模型的主要应用情景不同，具体可以参考我们给出的不同模型的实例。后面将会陆续介绍非监督学习和深度学习的方法及具体应用情景，并从模型中延展开来，持续追踪人工智能和大数据领域的发展状况及应用实践。

风险提示：
1. 模型在使用中存在建模风险；
2. 本文举例均是基于历史数据不保证其未来收益。

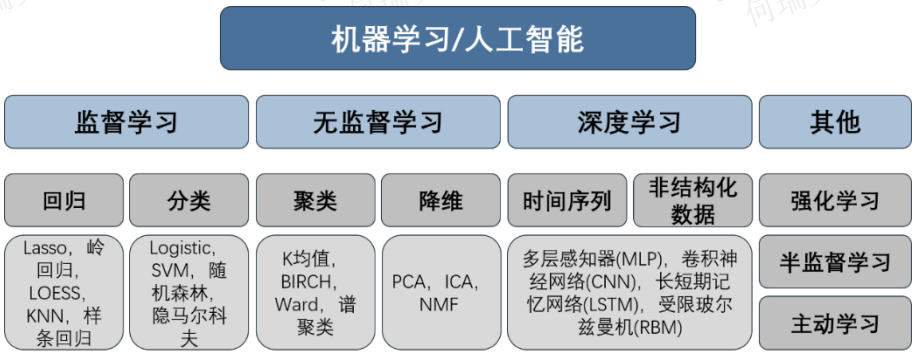
目录

图表目录

机器学习方法概述

本系列报告试图系统全面性的介绍各种不同的机器学习方法，并且结合具体的在投资研究领域应用实例、交易策略及 code 示例，说明其应用情景和方法。机器学习的方法可以分为以下几类：监督学习、无监督学习、深度学习及其他机器学习方法（例如强化学习），具体使用到的模型见图 1。此篇将以介绍监督学习方法为主。

图 1：机器学习/人工智能方法介绍



资料来源：JP Morgan，长江证券研究所

机器学习方法结合统计学和计算机两个领域，也可以根据具体方法与两个领域交叉度的高低，分为传统机器学习和深度学习两个大类。传统机器学习是统计学的延伸，金融领域多应用此类方法。深度学习例如卷积神经网络等则主要是应用于卫星图像处理、自然语言识别和其他非结构化数据分析。深度学习和强化学习在时间序列分析和投资组合构建中也有巨大的前景，深度学习有望在提升价格序列模型识别和收益预测的准确性上对于传统量化方法进行补充和改良，强化学习则为自动化交易的速度和有效性提供了保障。目前海内外也尝试将人工神经网络用于资产趋势的判断中，并且取得了一定的效果，但是与使用较为简单的随机森林等机器学习模型相比并未发挥出其优势。

传统机器学习在金融上的应用

监督学习和无监督学习统称为传统机器学习。一个机器学习算法通常是通过给定的数据来进行模型学习，选择出合适的参数，并且随着提供的数据量的不断增大模型的效果也逐步提升。监督学习方法更能体现出机器学习的概念，通过将样本数据截取出部分作为训练期，在训练期中明确输入指标（X）及对应的标签（Y），这样机器学习算法在模型学习的过程中相当于是基于给定的参考模式，在对应输出标签的“监督”下来选择合适的参数。

可以通过一个简单的例子来说明传统统计模型与机器学习模型的区别，例如我们在分析市场收益率和一些宏观指标如房地产投资增速、PPI、PMI、利率、宏观景气度等的关系时，传统的统计学采取的是线性回归的方式来计算市场收益率在这些指标上的 beta 值，通过机器学习方法，我们可以用一些更加先进的回归模型，通过尽量排除异常值的影响，考虑到众多输入变量之间的相关性及指标与输出结果之间的非线性，最后得到更为稳健的结果，常用的有 Lasso 回归及 Logistic 回归等。

对于无监督学习，在金融领域比较常用的是主成分分析（PCA）或独立成分分析（ICA），其并不存在输出变量作“监督”。例如在进行多因子分析时，我们的因子池里面有接近

八十多个子类因子，通过 PCA 可以将众多的因子进行特征抽取，最后将股票收益率归因于具有代表性的 8~9 个特征之上。另一种常用到的无监督学习方法是聚类分析，通过样本数据之间某种相似性将一组样本划分成几个小组。

深度学习的应用

在近些年来，除了传统机器学习方法之外，受到人脑思考模式启发的深度学习也得到了长足的发展。科学家观察到人的判断是由大脑之中连接起来的独立神经元共同作用的结果，每个神经元都可以接收不同来源的电流刺激，结合以往经验，给定不同电流刺激以相应的权重并粗略计算其加权平均值，通过与预定的阈值相比较，来决定“激活”或“忽略”这些刺激。计算机领域的专家发现通过复制这种构架，可以用于解决一些监督学习和无监督学习问题，并且将这种多层神经网络构架称之为深度学习。

过去几十年来，机器学习领域取得的卓越成就大多来自于深度学习方法的应用。图像识别、语音识别、语言翻译及自动驾驶都依赖于新的深度学习算法，其在大数据上处理优势和强大的自学习能力使其在投研领域具有巨大吸引力，但是目前的实践应用非常有限。

机器学习模型的应用情景

使用机器学习方法解决问题的第一步是需要在众多的模型中找到适合的一类。针对不同的情景，选择的最优机器学习模型也会存在差异。下面表格中我们列举出一些常用机器学习方用来解决的情景问题，有些情景可以有明确的金融实例可以对应，同时也列举了每种情景下常用到的机器学习方法。本篇后文也会通过特定的金融实例来具体介绍其中监督学习方法的具体使用。

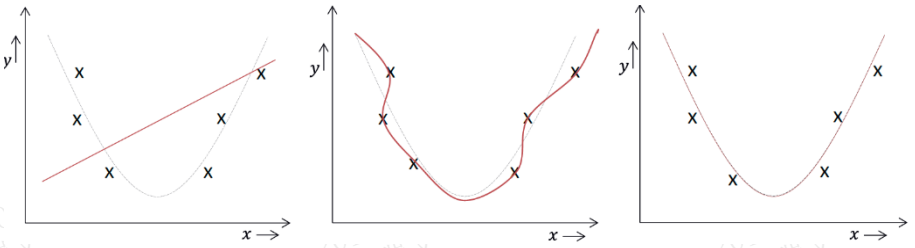
表 1：情景问题、具体的金融实例及其对应的机器学习方法

问题	金融实例	机器学习方法
给定输入变量，预测资产价格的方向	使用技术指标对于对应的指数进行择时	SVM、Logistic 回归、Lasso 回归等
一种资产的剧烈变动如何影响其他资产	美元指数的变动对于美国国债收益率及黄金走势的影响	格兰杰因果检验、脉冲响应函数
一种资产走势是否偏离其他相关资产	黑色系商品走势的分化	一对多分类
找出资产价格的驱动因素	行业中有效因子的筛选	PCA、ICA
目前市场状态判断	对于利率上行或下行周期判断	隐马尔科夫、Soft-max分类
一个事件发生的概率/是否会发生	高送转事件的预测	决策树、随机森林、Logistic回归
在噪音数据中寻找信号	资产周期的分析	低通滤波器、SVM
一篇文章或一段文字的感情色彩、主题	公司公告的舆情分析	词袋分析、词频-逆向文件频率（TF-IDF）
有哪些常见的市场压力指标		K-means聚类分析
计算图像中某物体数量		卷积神经网络
最优执行速度		基于部分可观察马尔科夫过程的强化学习
基于大量输入数据预测波动率		受限玻尔兹曼机、SVM

资料来源：JP Morgan Macro QDS，长江证券研究所

选择了合适的模型之后，还面临着参数个数选择及参数优化的问题，此时需要权衡模型的“方差”和“偏差”，如下图，线性回归过于简单，不足以解释数据点，具有较大的“偏差”，我们称为“欠拟合”。而使用高阶多项式降低偏差则导致模型“过拟合”，在新的数据点，模型预测准确性会降低，我们称为“方差”较大。

图 2：欠拟合、过拟合及完美拟合图例



资料来源：Wind，长江证券研究所

模型的复杂程度提高的时候能够降低偏差但是会增大方差，因此合适的模型需要优化的是方差与偏差之和。一般可以用优化的参数个数来衡量模型的复杂程度，模型的拟合程度则可以通过交叉验证的方法来判断。

交叉验证的方式是采取分组采集子样本集的方式，将原始数据分作训练集和验证集，在训练集中调试参数然后将模型用于验证集，通过验证集中的拟合结果与真实值之间的误差来衡量。当然交叉验证的方式并不适合于所有的机器学习模型，实践中很多编程工具也会提供各种模型拟合程度判断统计量及验证方法，都要求我们事前对于模型原理有基本的了解。

监督学习模型之回归

监督学习可以进一步分为回归和分类。回归方法试图根据输入变量来预测输出变量的值，分类方法则尝试将输出结果分到不同类别。与普通线性回归不同，监督学习中的惩罚回归模型和非参数回归，可以分别用于处理输入变量中存在大量线性相关性关系及非线性关系时的情况。下面将就几种常见的模型做具体介绍及示例。

惩罚回归模型

最具代表性也是在金融领域使用最多的惩罚回归模型有 Lasso 回归、岭回归和弹性网络回归，与普通线性回归相比这三种模型在寻找最优回归系数β上做了进一步改良，使模型稳健性更强。

Lasso 回归、岭回归和弹性网络回归

对于普通的线性回归模型（OLS），我们通过最小化预测值与真实值之间的误差来求得输入变量的系数β，假设预测值 y 为输入值 $x_1, x_2, ..., x_n$ 的线性组合，系数β通过下面方法估计：

$$\text{OLS: 最小化 } \sum\left(y-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)\right)^2$$

此时容易产生较为分散或数值较大的β，可以添加一个反映我们对数值较大的β厌恶的惩罚项，来防止这种情况发生，即：

Lasso 回归：最小化 $\sum\left(y-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)\right)^2+\alpha\sum_{i=1}^{n}|\beta_i|$

模型会将不需要和非常大的β设置为零。加上系数绝对值的惩罚项称为 L1 正则化，这种改进的线性回归称为 Lasso 回归。通过对系数的优化，Lasso 实质上进行了特征选择。

Lasso 的目标函数可以理解如下：

- 当α = 0 时，得到普通线性回归系数。
- 随着α 的增加，选择的特征越来越少，最终得到最重要的一个特征。

这里，α 被称为模型的参数。 同理，如果我们在目标函数中添加β的平方，就得到了岭回归。

岭回归：最小化 $\sum\left(y-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)\right)^2+\alpha\sum_{i=1}^{n}\beta_i^2$

加上系数平方的惩罚项称为 L2 正则化，这种改进的线性回归称为岭回归。 弹性网络回归是 Lasso 和岭回归的混合体。

弹性网络回归：最小化 $\sum\left(y-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)\right)^2+\alpha_1\sum_{i=1}^{n}|\beta_i|+\alpha_2\sum_{i=1}^{n}\beta_i^2$

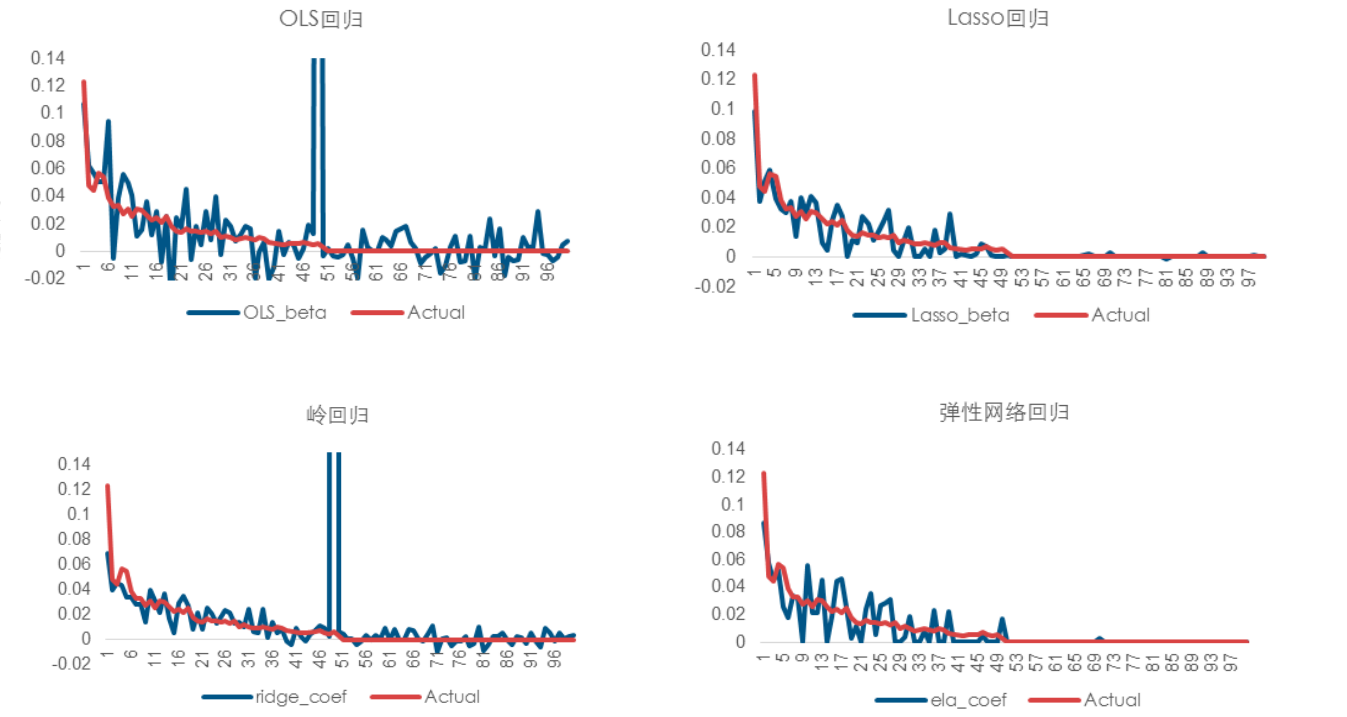
三种模型在进行参数估计时的逻辑一脉相承，选择哪种模型则跟样本特征相关性较大，一般最常用的是 Lasso 模型，基本上能够达到找出核心变量、排除变量共线性的作用。

惩罚回归模型应用实例

惩罚回归模型的典型应用实例是在存在众多相关性较高的变量中筛选出对于被解释变量解释力度最大的变量及对应的模型。例如在我们之前的一篇报告《因子轮动系列（二）：宏观周期与因子投资时钟》中，针对规模因子和价值业绩因子、波动率因子和 beta 因子的轮动，将几个宏观变量进行重要性排名，最后选出最为核心的两个影响因素，使用的就是 Lasso 回归。

为了比较几种惩罚回归模型的效果，展示其在实践应用中相对于传统线性回归（OLS）模型的优势，我们以上证 50 指数收益率以及其影响变量为例。已知上证 50 的收益率可以由其成分股收益率完全解释，我们用总共 100 只股票收益率作为解释变量，其中 50 只为上证 50 的成分股，另外 50 只随机选择作为干扰项。与真实权重相比，几种模型的拟合效果如下：

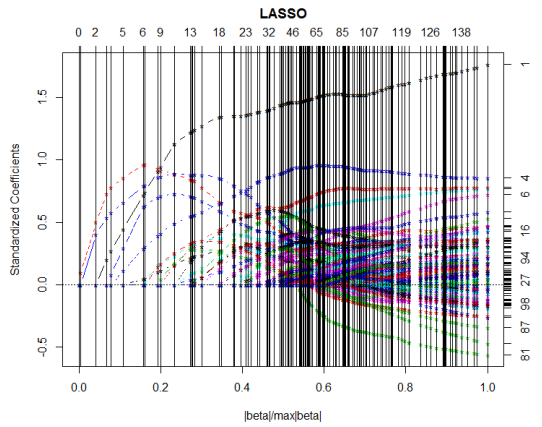
图 2：惩罚回归模型的拟合效果比较



资料来源：Wind，长江证券研究所

上图中变量的 beta 值用纵坐标来表示，横坐标对应 100 个变量。红色的线代表此段时间区间内各变量的真实值（取平均），在存在干扰变量的情况下，OLS 回归出的系数波动较大，且存在极端值和负数情况。相比而言 Lasso 回归和弹性网络回归都较强的排除了后面 50 个干扰变量的影响，与真实结果较为接近。Lasso 模型中也可以通过调整 α 的取值对影响权重较大的变量做进一步筛选，但在本例中应用不大因此不详细展开。

图 3：Lasso 模型中 beta 选择的轨迹图



资料来源：Wind，长江证券研究所

每一个模型在优化的过程中都有一个 beta 变化路径，如图 3 所示。

惩罚回归模型由于可以一定程度上排除非相关性变量及共线性强的变量的干扰，在用来做样本外预测时稳健性强于传统的线性回归模型。所以其另外一个具体应用可以是在做

多因子分析时，例如我们做某个行业内的因子模型，可以使用惩罚回归模型，筛选出对于此行业内选股较为有效的指标，然后对于行业收益率进行样本外预测。

非参数回归

监督学习中的回归方式也可以分为参数或非参数方法。参数回归中，模型由一系列参数描述，如根据历史数据估计的线性回归的β。在非参数回归中，我们假定相似输入具有相似输出，然后直接识别类似的历史样本以达到预测目的，针对新加入的数据点，通过搜索历史数据，并找到所谓的“最近邻”的 K 个相似样本点。最后通过对“最近邻”的输出取平均。

非参数回归提供了一种利用过去类似事件来预测未来的方法，在金融领域，输出变量与输入变量通常不是线性相关的，这使得线性回归及其扩展（如岭回归和 Lasso 回归）不适用。本文将介绍两种典型的非参数回归模型：K 最近邻和 LOESS。

K 最近邻和 LOESS

K 最近邻（KNN）回归：通过找出一个样本的 k 个最近邻居，将这些邻居的输出变量 y 的平均值赋给该样本，并将其用作我们的预测。K 最近邻可以捕获数据中的非线性属性。但是其主要缺点在于对异常值极度敏感。

局部线性回归 LOESS：依据 KNN 方法，对于每个新样本点，基于 K 个最近邻的数据进行线性回归，并使用拟合出的系数预测输出值。

线性回归和 K 最近邻法可以看作是经典机器学习的两个极端。线性回归可能“低估”数据，因此偏差较高，方差较低。而 K 最近邻可能“过拟合”，因此方差较高，偏差较低。KNN 中是通过调整 K 值来平衡“偏差”及“方差”。举一个较为极端的例子，当 K 取 1 的时候相当于对每个训练样本都划分了一个微小的区域，这很可能导致在未知样本上出现高错误率，产生样本内的过拟合。

关于 K 最近邻的方法在金融领域应用最常见的一个例子是用于选股中，例如结合因子数据，可以将股票按照收益率特征分为强势和弱势组合，进行预测时通过输入股票的因子数据找到其下期所处的组合类别，通过此种方式来构建股票组合。

动态系统-卡尔曼滤波

不同于静态的线性回归模型，卡尔曼滤波器考虑到β系数随时间缓慢变化的动态过程，其常常用于统计交易和波动率估计。在卡尔曼滤波器中，β系数在一定范围内连续变化，可迭代估计。如果我们将这个变化范围离散到一组有限的值内，则可以导出一个隐马尔科夫模型（HMMs），HMMs 将在分类模型中做进一步介绍。

卡尔曼滤波器在 1960 年由 Kalman 提出，其将一系列具有不确定性的观测值进行组合，从而估计和预测一个动态系统的参数。该算法通常分两步进行。第一步，对现在状态进行估计并得到估计误差。第二步，结合下一个观测值和误差得到新的预测（通过对先前的估计和误差以及新的观测和误差给予适当权重）。

动态系统由状态空间模型（或动态线性模型，DLMs）描述，其中有两个组成部分：

1）状态进化：

随时间不断变化不可观察的变量称为系统的状态，状态的变化方程为具有高斯噪声的线性表达式：

$$X_t = FX_{t-1} + w_t, w_t \sim N(0, Q)$$

给定先前的状态X_{t-1}，我们能够估计当前状态X_t，但由于外部随机因素影响，存在不确定性。

2）测量：

虽然我们不能直接观察状态X_t，但是我们可以得到状态X_t的测量Z_t，只不过测量仍然伴随着高斯噪声：

$$Z_t = HX_t + v_t, v_t \sim N(0, R)$$

测量可以用于估计状态，但有不确定性。

卡尔曼滤波器结合上述信息，给出最优状态变量服从高斯分布$\mathrm{N}(X_{t|t}, P_{t|t})$，其均值和协方差分别为：

$$X_{t|t} = X_{t|t-1} + K_t(Z_t - HX_{t|t-1})$$

$$P_{t|t} = P_{t|t-1} - KHP_{t|t-1}$$

其中$K = P_{t|t-1}H^T(HP_{t|t-1}H^T + R)^{-1}$是卡尔曼增益。公式虽复杂，但其推导仅依赖于条件高斯分布：

$$\begin{pmatrix} X \\ Z \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Z \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

$$X \mid (Z = z) \sim N(\hat{\mu}, \hat{\Sigma})$$

其中，$\hat{\mu} = \mu_X + \Sigma_{12}\Sigma_{22}^{-1}(z - \mu_Z)$，$\hat{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$。

卡尔曼滤波应用实例

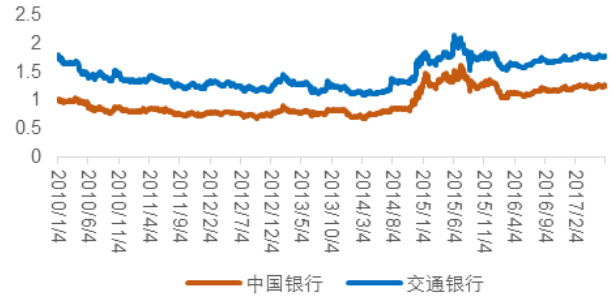
卡尔曼滤波实际应用中最著名的例子是在阿波罗任务中航天器导航，近些年来非线性的卡尔曼滤波也广泛的应用于自动驾驶。在金融领域，卡尔曼滤波器常用于趋势估计、信号去噪或描述资产与市场之间的动态关系。

一个常见的金融上的应用实例是做配对交易。海外由于 ETF 的种类繁多，交易费用低，所以常用 ETF 来做配对交易，在 ETF 配对策略中使用卡尔曼滤波器估计β，寻找协整

关系中的背离机会，相比于传统的回归方式有一定的增强作用。我们这边为了说明卡尔曼滤波器的效果，也选用一个配对交易的例子，但是由于国内限制条件较多，此例仅做理论上的方法实践及效果展示。

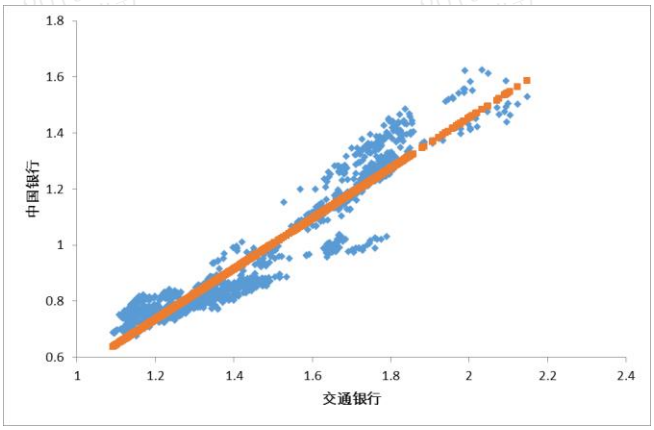
假设可以对于股票通过融资融券进行卖空操作，我们选取一对相关性较强的股票组合：中国银行和交通银行对卡尔曼滤波器在处理这种问题时的优势做详细说明。

图 4：中国银行和交通银行价格（取自然对数）走势图



资料来源：Wind，长江证券研究所

图 5：中国银行和交通银行价格（取自然对数）走势图



资料来源：Wind，长江证券研究所

从上面图 3 和图 4 可以看出，在 2010 年 1 月到 2017 年 6 月这段区间，两只股票之间的走势具有较强的相关性，当然也可以通过协整检验来验证其协整关系。

通过估计两只股票价格序列的β，来监控他们之间相关关系的变化：

$$S_{t,1} = \beta_t S_{t,2} + v_t, v_t \sim N(0, \sigma_v^2)$$

我们可以进一步假设β不是常数，而会随着时间变化。为了简单起见，我们假设β的变化服从随机游走：

$$\beta_t = \beta_{t-1} + w_t, w_t \sim N(0, \sigma_w^2)$$

这是一个动态线性回归问题，在状态空间中，β_t是状态变量，系统如下，使用到的状态方程和测量方程见上面卡尔曼滤波介绍部分，根据观察值（股票的价格），我们可以使用卡尔曼滤波器来估计β值。这一单变量例子中的卡尔曼增益为

$$K_t = \frac{S_{t,2}}{S_{t,2}^2 + \gamma^{-1}}, \gamma = \frac{\hat{P}_{t|t-1}}{\sigma_v^2}$$

γ表示信噪比（SNR）：状态方差与测量误差的比值。如果信噪比较小，则测量结果是嘈杂并且无效的，因此对先验信息$\beta_{t|t-1}$的加权较大。如果信噪比大，观察值加权应较大。

如果信噪比非常小，观察值于我们并没有用（因为它是嘈杂的），我们只用了先验信息：

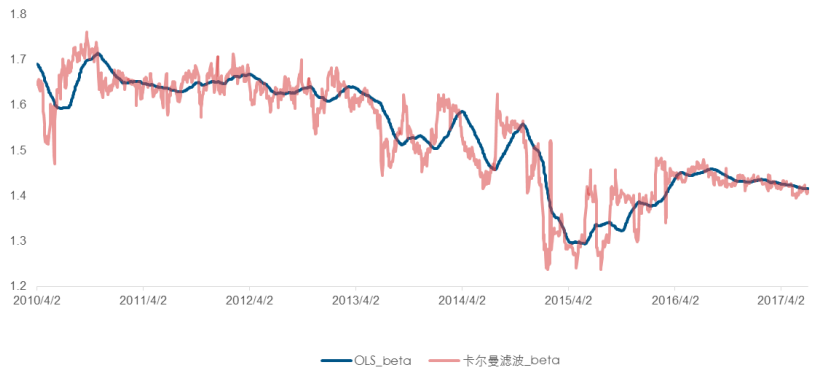
$$\hat{\beta}_{t|t} \approx \hat{\beta}_{t|t-1}$$

除了卡尔曼滤波器，我们显然也可以使用普通的线性回归：

$$S_{t,1} = \beta_t S_{t,2} + v_t, v_t \sim N(0, \sigma_v^2)$$

我们使用卡尔曼滤波器，同时也用过去 60 个交易日滚动窗口的线性回归对β进行估计。两者对比，不难看出，卡尔曼滤波器更灵敏。事实上，卡尔曼滤波器与指数平滑技术密切相关，它给最近的观测值更多的权重，并且可以根据测量的“噪声”调整权重。

图 6：卡尔曼滤波器和 OLS 回归估计的 beta 值



资料来源：Wind，长江证券研究所

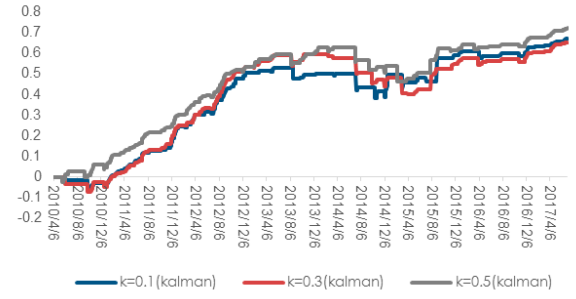
此交易信号只依赖于残差v_t，残差应当围绕均值 0 波动。在每个交易日结束时，我们得到新的股票收盘价来更新我们对 β 的估计，然后计算残差：$v_t = S_{t,1} - \beta_t S_{t,2}$。

我们把残差的不确定性记为σ_t，我们可以用它来确定残差的大小是否足以触发我们的策略：

如果$v_t \geq k\sigma_t$，我们做多β_t单位股票 2，同时做空股票 1；

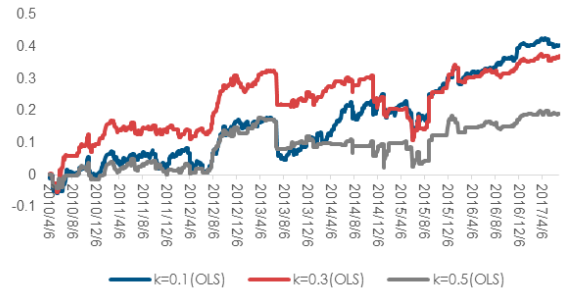
如果$v_t \leq -k\sigma_t$，我们做空β_t单位股票 1，同时做多股票 2。

图 7：卡尔曼滤波器做配对交易时的净值收益率



资料来源：Wind，长江证券研究所

图 8：OLS 做配对交易时的净值收益率



资料来源：Wind，长江证券研究所

K 分别取值为 0.1、0.3 及 0.5，得到策略收益率如上图所示。使用卡尔曼滤波器得到收益率略高于 OLS 模型，在不同 k 值下稳定性也更好。但是整体而言，策略并未产生较多的超额收益。

极限梯度提升（XGBoost）

“提升”是指迭代地组合弱“学习器”（即具有弱预测能力的算法）以形成具有强预测能力的算法。Boosting 从弱学习器开始（通常是回归树算法），记录学习器的预测与实际输出之间的误差，在每一次迭代中，它都能根据误差来改善前一迭代步骤中的弱学习器。如果误差项在损失函数负梯度方向上，则该方法被称为“梯度提升”。极端梯度提升（XGBoost）是指 Chen 和 Guestrin 的优化实现，是处理金融时间序列数据的一种流行监督学习算法。

回归树与决策树类似，只不过在每个子叶节点我们得到的是连续的数值非离散的类标签。输入一个大小为 m 的向量到一个有 T 片叶子的回归树模型中，输入变量与叶子节点的映射关系由函数：$q: R^m \to \{1, \ldots, T\}$表示。

用函数 w 表示在叶子上的得分，那么第 k 个数为函数$f_k(x) = w_{q(x)}$，其中$w \in R^T$。对于大小为 n，样本为(x_i, y_i)，$x_i \in R^m$，$y_i \in R$的训练集，树集成模型将使用 k 个函数的和来预测最终输出结果：

$$\hat{y}_i = \emptyset(x_i) = \sum_{k=1}^{K} f_k(x_i)$$

为了进行模型中的一组函数训练，定义正则项如下：

$$L(\emptyset) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

树集成模型以加法方式进行优化，用$\hat{y}_{it}$表示增强迭代第 t 阶段第 i 个训练样本的预测，那么最小化的目标函数可以写为：

$$L_t = \sum_{i=1}^{n} l\left(y_i, \hat{y}_{i,t-1} + f_t(x_i)\right) + \Omega(f_k)$$

通过二阶泰勒式展开，进行优化最后得到集成的树模型f_k。为了防止过度拟合，XGBoost 允许进行树的修剪和样本特征抽样（如随机森林模型）。

由于极限梯度提升方法经常和树模型一起使用，在此处不给出单独的实例，可以参考随机森林部分的实例。

监督学习之分类

监督学习中分类方法的目标是把观察值分为不同类别。在金融领域中我们经常希望对资产的趋势进行预测，但是很多时候采取回归的方法得到的预测值准确度较低，此时就可以采取分类的方式，一方面可以提升准确度，另外一方面某些情况下对于状态的预测比绝对值上的预测意义更大。此篇，我们将介绍以下分类算法：逻辑回归、支持向量机

（SVM）、决策树、随机森林以及隐马尔可夫模型。其中逻辑回归和支持向量机的原理在之前的报告《大类资产配置之机器学习用于股票资产资产的趋势判断》中有详细介绍及数学推导。

逻辑回归

逻辑回归即 Logistic 回归（又称 logit）利用给定的历史样本，预测事件发生的概率，在我们上篇报告中也用到逻辑回归对于股票资产的趋势进行预测，根据一系列宏观及估值类的指标预测月度股票走势，将其走势分为“上涨”、“下跌”两类，也可以根据涨跌幅度对于分类结果进行进一步细分。逻辑回归是对普通线性回归的简单变换。我们首先得到输入变量的线性组合，然后给出一个函数将该数映射到 0 和 1 之间。

支持向量机

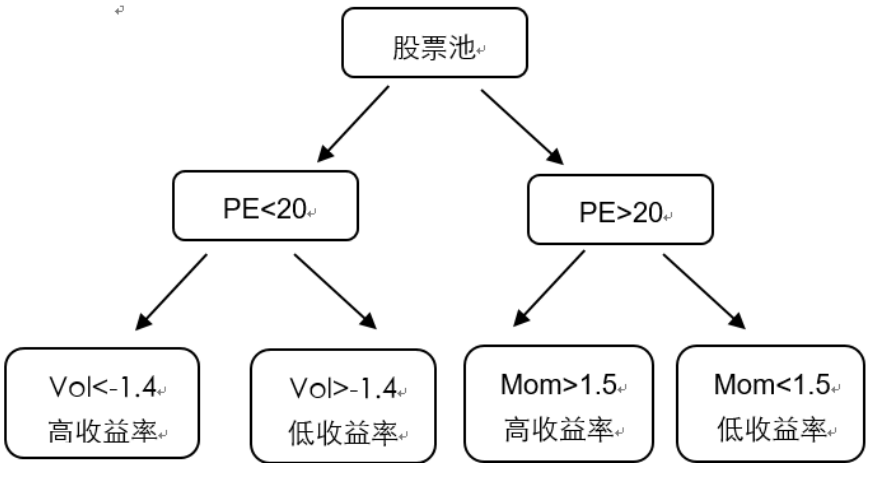
支持向量机因为其在使用及参数优化上的便利性成为最常使用的分类算法之一，常见的使用情景是在资产价格变化方向的预测上。假设我们有一系列的指标用于资产趋势预测，普通分类模型完成的任务是找到一组线性组合，当资产呈现上涨趋势时线性组合的值较大（或较小），反之，当资产呈现下跌趋势时对应的线性组合的值较小（或较大），支持向量机想要找到使得上涨或下跌趋势对应的线性组合的值区分度最大的结果。

决策树和随机森林

决策树模型本质与工商管理和金融分析中普遍应用的流程图类似，要得到最终的结果，需要解答中间一系列问题。根据每一步的对于问题的解答，来选择树的分叉方向。决策树的最终结果会受到中间问题的顺序的影响，一般将影响最重要的问题放在最前面。

决策树模型是用于非线性指标分类的最简单的模型之一，举一个简单的在金融领域应用例子，比如我们想要进行因子选股，不同于传统的多因子模型，我们认为因子暴露度与个股收益率之间存在非线性关系，这时可以用决策树模型，选用动量(Mom)、波动率(Vol)、PE 三个因子，通过下面的树模型决策过程得到股票组合。

图 9：决策树模型股票筛选示例



资料来源：Wind，长江证券研究所

通过上面的简单例子也可以发现，决策树模型拟合的核心在于寻找最优变量和分裂阈值，以最小化特定的损失函数。损失函数可以定义为子叶节点的不纯度，通常使用 Gini 系数或者熵度量。实际使用中通过参数调整来确保树模型预测准确度防止过拟合，例如：

- 最大深度（Max depth）：决定决策树的最大深度
- 节点样本个数（Node size）： 每一个节点至少有 N 个观察样本

也可以通过修建枝叶即决策树构建好后，用单一叶节点代替整个字数或者用一个数字代替一颗子树来防止模型由于太过“茂盛”产生过拟合。

决策树模型虽然逻辑简单使用方便，但使用过程中稳健性较差，样本产生一些小变化就有可能导致拟合出完全不同的树模型，因此作为预测模型单独使用效果不佳，可以用于观察不同变量之间的交互影响及形成集合模型。

随机森林就是依据决策树模型构建的一种典型的集合型算法，可以用于解决单个决策树模型预测时方差较大的问题。通过对原始数据进行随机样本划分，每一棵决策树都依据部分样本进行单独判断，最后的结果通过众多树模型投票得出，与简单树模型相比这种做法可以降低预测的方差。

为了防止构建出来的树模型性之间相关性较高，每一棵树模型都是从总共 p 个变量中随机选择 m 个分裂变量，根据 m（m<p）个变量来构建模型。此时，要度量每个变量对于结果的重要性，可以通过记录每个节点根据此变量分裂后不纯度的下降程度。另外一个衡量方法是部分依赖图，根据分类的几率的自然对数来绘制。仍然使用上面介绍的用三个因子分类的例子：

$$f(x) = \log \frac{\Pr(\text{收益率低}|x)}{\Pr(\text{收益率高}|x)}$$

其中 x 指动量(Mom)、波动率(Vol)或 PE 三个因子的值。

分类模型在选股及择时上的应用实例

上文介绍的几种分类模型是目前最常使用的几种模型，在国内外关于这几种模型的研究和实践案例分析的参考资料也较多。包括我们之前的报告中就有涉及，使用 logistic 模型、SVM 模型及人工神经网络模型，依据宏观指标以及资产的估值指标对于股票资产进行趋势判断。

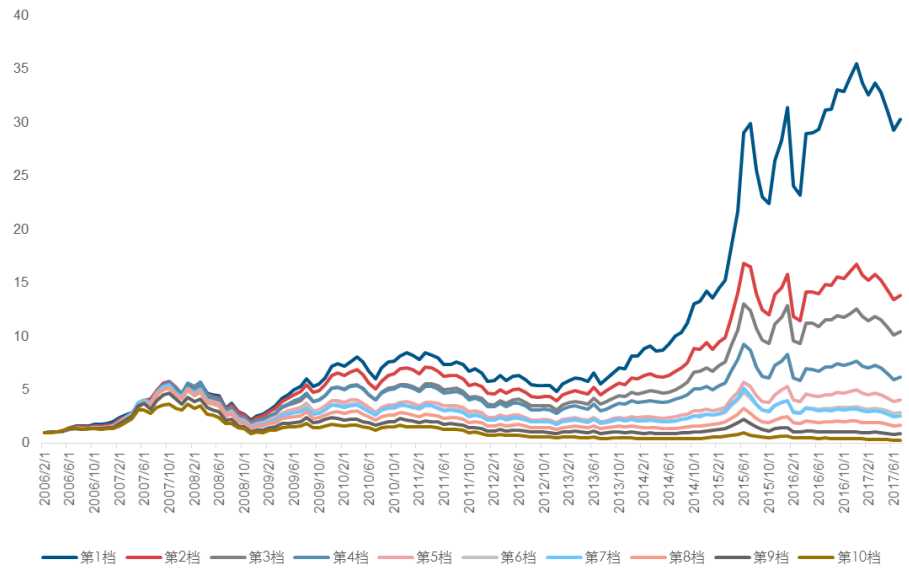
SVM 模型应用于沪深 300 内选股

关于 Logistic 模型及 SVM 模型的应用实例，在除了进行这种择时的判断，就目前大家比较关心的多因子选股，两种模型也都可以实现相关的功能。

选股的逻辑基本大同小异，本文示例中以月度为单位，选择过去 12 个月的沪深 300 股票的因子暴露度及股票下期收益率作为训练数据来训练模型，其中将股票的下期收益率按照高低分为 10 档，第一档为强势股，第十档为弱势股，对应标签 1~10。训练好的模型用于拿到下期因子数据后的预测中，得到股票的对应标签。

根据上面的方法，我们选择沪深 300 为股票池，剔除上市不满 1 年的次新股、ST 股，考虑到涨跌停情况及交易费用，回测区间选择 2006 年 2 月份到 2017 年 7 月，按照月度进行换仓。SVM 根据数据特征可以选择不同的核函数，在进行多因子选股时线性核的效果最好。

图 10：SVM 在沪深 300 内多因子选股分档效果

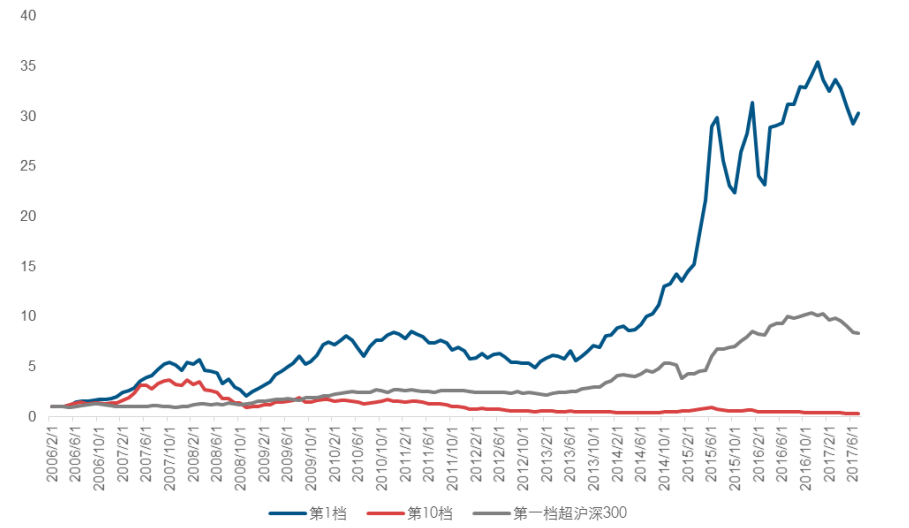


资料来源：Wind，长江证券研究所

常见的核函数的选择有高斯核函数、多项式核函数和线性核函数，比较而言，线性核函数用于多因子选股的效果最好。

使用线性核进行选股分档效果如上图，高低组之间的收益分化明显。

图 11：SVM 沪深 300 内多因子选股与基准比较的效果



资料来源：Wind，长江证券研究所

分年收益表现如下：

表 2：SVM 选股分年效果

日期	收益率	超额收益率	夏普比率	最大回撤	超额最大回撤
2006	96.81%	-2.65%	3.77	0.27%	21.65%
2007	175.97%	5.52%	3.52	14.72%	16.92%
2008	-49.71%	47.68%	-0.82	62.99%	8.14%
2009	172.78%	38.67%	4.37	12.60%	10.25%
2010	10.34%	26.13%	0.95	24.97%	7.90%
2011	-29.50%	-5.98%	-1.05	31.66%	7.47%
2012	-4.73%	-11.42%	0.42	23.13%	10.62%
2013	47.40%	59.60%	1.94	15.46%	1.85%
2014	66.33%	9.67%	3.25	4.91%	27.97%
2015	131.61%	119.36%	2.43	25.16%	0.00%
2016	7.43%	21.10%	0.72	5.01%	2.08%
2017	-10.06%	-18.81%	-0.62	13.02%	15.23%
平均	**34.53%**	**20.26%**	**1.38**	**62.99%**	**27.97%**

资料来源：Wind，天软，长江证券研究所

分年回测效果如上，今年以来表现不佳，整体而言回测区间内有正超额收益，但是与传统的多因子模型相比并未表现出较大优势。

随机森林依据多因子数据的择时

使用每日的因子收益率数据，由于不同指数的行业权重分布偏差，某些行业对于指数的走势有较大决定性作用，因此除了常见的几种大类风格因子还会使用到行业因子收益率数据。行业因子按照中信一级行业进行分类。

预测的主要信息如下：

预测标的的选择：可以针对主要指数包括沪深 300、中证 500 和中证 1000 进行择时。

模型的选择：主要考量随机森林模型的预测结果。

预测的结果：选取训练期之后，将当期因子收益率和下期三种指数的涨跌幅（上涨为 1，下跌为 0）进行训练，得到的模型用于下期指数涨跌结果的预测。采取滚动的方式进行。

表 3：随机森林择时准确度

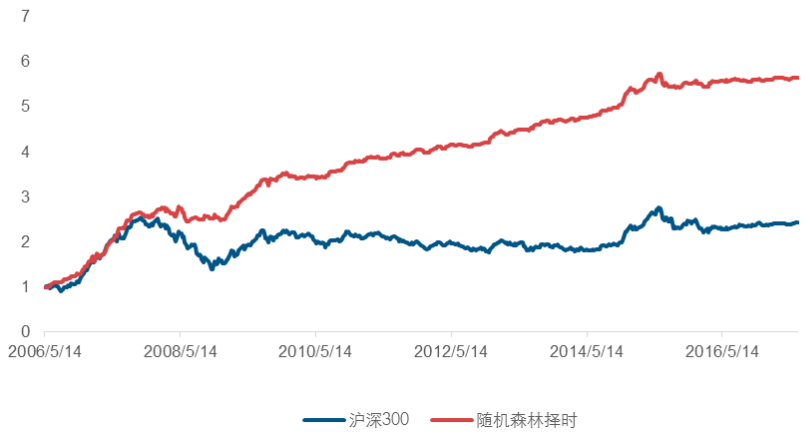
输入的指标	模型/周期	标的指数	准确度
全部因子	随机森林/60周	沪深300	65.20%
		中证500	69.24%

	中证1000	72.76%

资料来源：Wind，长江证券研究所

从预测的准确度来看，利用多因子数据进行周度择时在三种指数上的择时效果都较好，我们也从沪深 300 择时的收益角度来做具体分析。

图 12：随机森林沪深 300 择时效果



资料来源：Wind，长江证券研究所

表 4：随机森林沪深 300 择时分年效果

日期	择时收益率	超额收益率	基准最大回撤	择时最大回撤	夏普比率
2006	115.14%	-1.53%	13.79%	2.10%	7.02
2007	184.14%	34.57%	17.78%	12.28%	5.86
2008	-18.67%	45.74%	71.27%	29.89%	-0.55
2009	137.08%	57.58%	25.88%	12.78%	5.21
2010	33.76%	44.30%	29.02%	7.36%	2.01
2011	21.79%	45.69%	30.60%	5.43%	1.45
2012	45.20%	37.37%	22.40%	4.59%	3.36
2013	34.00%	42.91%	21.60%	7.25%	2.34
2014	81.00%	35.56%	9.29%	4.26%	4.29
2015	18.14%	11.55%	39.58%	27.30%	0.66
2016	-1.08%	9.76%	22.69%	13.65%	-0.07
2017	11.20%	-1.89%	4.19%	2.89%	1.96
平均	**48.37%**	**38.98%**	**71.27%**	**29.89%**	**2.03**

资料来源：Wind，长江证券研究所

分年来看，从 06 年到 17 年 6 月份基本上每年都能够获取超额收益，相对于基准的回撤也较小，盈亏比有 1.4，综合考虑盈亏比和预测准确度，比传统的技术性择时指标更为稳健。

我们在此就不再展开分析随机森林在中证 500 和中证 1000 上的择时效果。

隐马尔科夫模型

在卡尔曼滤波介绍中，我们提到卡尔曼滤波方法可以用于估计动态系统中的β系数，如果β的变化是离散的话，那不同β可以看做是代表不同的“状态”。隐马尔科夫模型（HMMs）类似于卡尔曼滤波，假设下个状态的发生只跟现在的状态有关（即隐藏状态服从离散马尔科夫过程）。HMMs 具有很强的实用性，因为在很多实际问题中，我们对于识别一些不能直接观察到的事件非常感兴趣，例如现在市场是处于一个趋势向上还是趋势向下状态，而这些问题可以通过其他能够观察得到的变量得到（例如市场收益率、波动率等）。

在上世纪 90 年代 HMMs 被广泛的应用于语音识别中，近些年来，在生物信息领域例如基因序列分析上应用较多。在金融领域，HMMs 主要用于市场状态的刻画。假设市场只有上涨和下跌两个状态，一个隐马尔科夫模型可以表述为市场状态的马尔科夫过程。意味着，如果现在市场是处于上涨状态，那么持续上涨状态的概率为 80%，转变为下跌状态的概率为 20%。市场收益率的分布是一个依据现在市场状态的条件概率分布：$(r|state)\sim N(\mu_{state},\sigma^2_{state})$。

由于市场收益率是有历史数据的，我们可以通过收益率来推导出不同时期的市场状态的似然性。HMMs 中的参数是通过 EM 算法优化这种似然性得到，估计的参数有五个，两个状态集合及三个概率矩阵，包括每个状态的初始概率、状态转移概率、在当前状态下的概率、观察值（例如收益率）在各个状态下的均值和方差。

图 13：HMMs 模型的估计结果（转移矩阵）

```
Initial state probabilties model
pr1 pr2
  1   0

Transition matrix
        toS1  toS2
fromS1 0.940 0.060
fromS2 0.031 0.969
```

资料来源：Wind，长江证券研究所

图 14：HMMs 模型的估计结果（均值和方差）

```
Response parameters
Resp 1 : gaussian
    Rel.(Intercept) Rel.sd
St1           0.051  0.101
St2          -0.005  0.057
```

资料来源：Wind，长江证券研究所

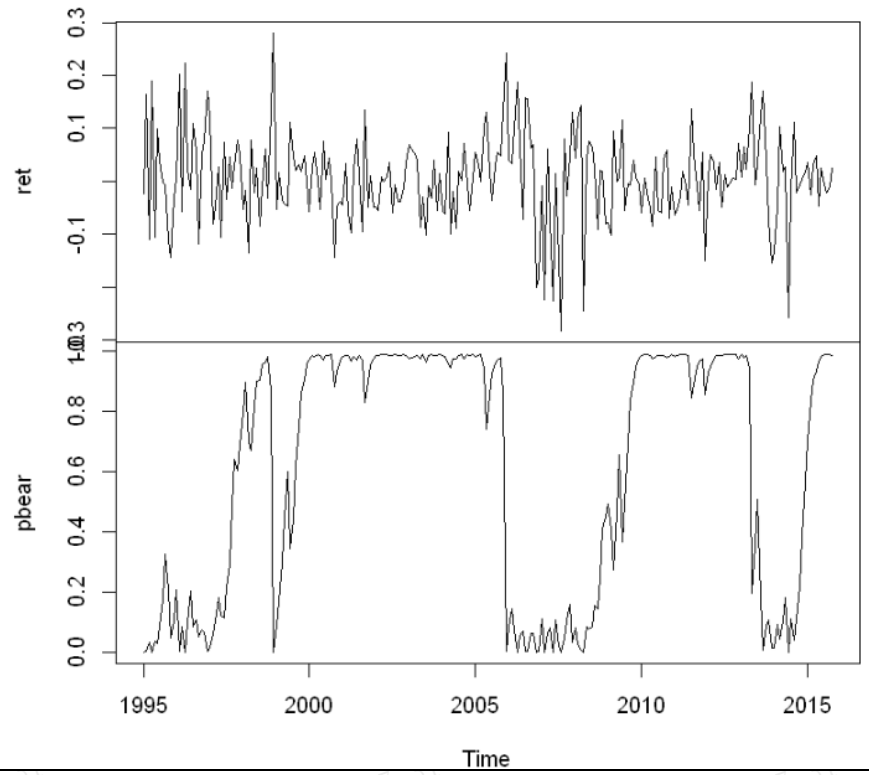
隐马尔科夫模型的应用实例

基于 HMMs 来判断上证综指的上涨和下跌趋势，依据趋势判断结果来验证 HMMs 的状态划分方式对于国内 A 股市场是否有效。我们选择从 1995 年 1 月份至今的日度收益率数据，分别按照日度、周度及月度进行状态划分。在模型判断市场上涨时持有上证综指，市场下跌的时候持有现金。一般在进行 HMMs 预测时，观察变量数据量越大越好，在其足够大的情况下才能够确保每种状态出现足够多的频次。

我们选择从 1995 年 1 月份开始到现在的上证综指日度数据，收益率为正并且波动率较高的周期定义为上涨趋势；收益率为负并且波动率较低的周期定义为下跌趋势。

图 14 是以月度收益率的频率根据这段时间区间计算出来的处于状态 1 下的概率，状态 1 对应的是低收益率低波动率。

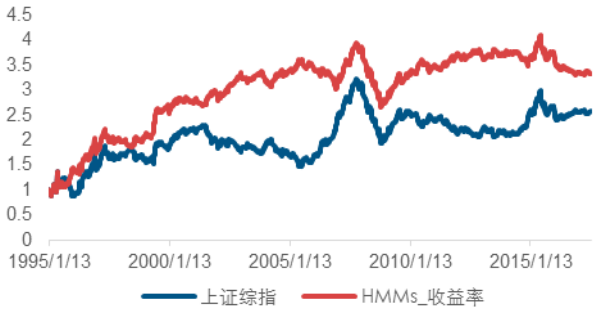
图 15：市场下跌趋势状态下的后验概率



资料来源：Wind，长江证券研究所

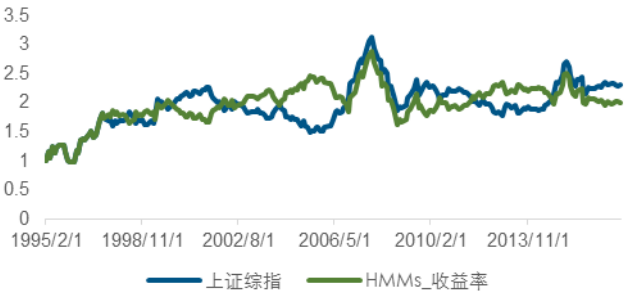
从隐马尔科夫模型判断的状态与市场实际所处的状态的契合度来看，日度状态判断效果较差，对于收益率和波动率无法起到显著区分作用。整体比较而言，周度效果最佳。相对于上证综指的择时效果见下图。

图 16：HMMs 模型的周度择时效果



资料来源：Wind，长江证券研究所

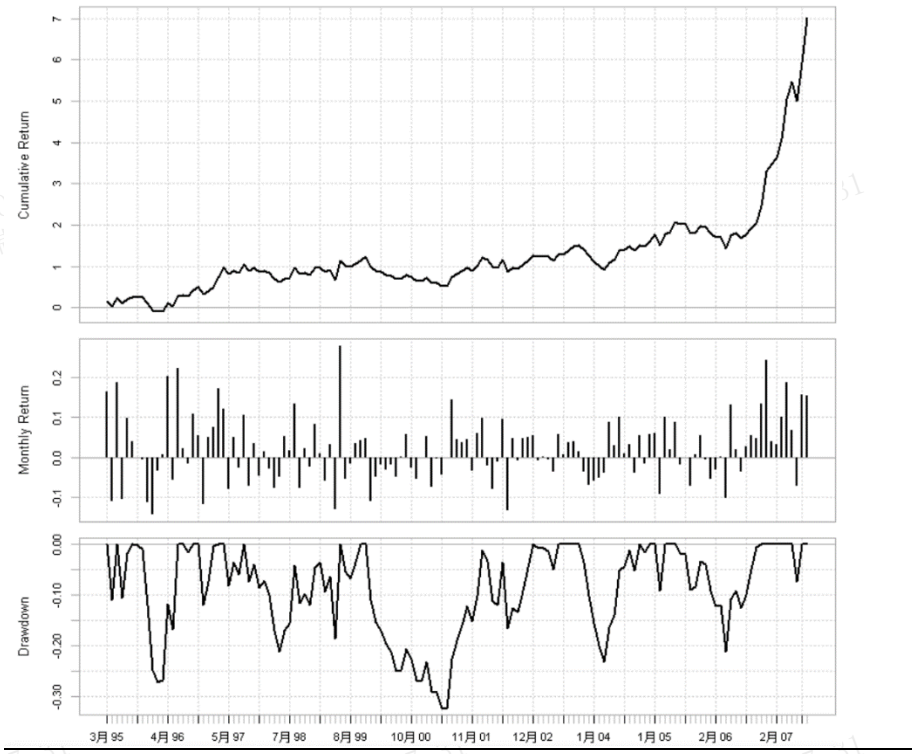
图 17：HMMs 模型的月度择时效果



资料来源：Wind，长江证券研究所

从时间段上来看，无论是日度、周度还是月度模型，HMMs 的划分在 95 年到 07 年这段时间区间效果较好，相对基准有较高的超额收益，但是 07 年之后相对于上证综指判断效果较差，无超额收益。

图 18：1995 年至 2007 年月度判断择时效果展示（累计收益率、月度收益率及回撤）



资料来源：Wind，长江证券研究所

从模型中的五个估计值也可以判断出，在 95 年至 07 年期间，模型估计出的状态 1 和状态 2 分界清晰，状态 1 下收益率均值为正，方差较大，状态 2 则正好相反；07 年底至今，进行划分时，两种状态下收益率均值均为负，说明划分效果并不好。

图 19：2007 年年末至 2017 年 7 月月度判断择时效果展示（累计收益率、月度收益率及回撤）

资料来源：Wind，长江证券研究所

实际应用中也可以将隐马尔科夫模型用到频率更高的预测之上，并且采数据集滚动的方式不断的更新和增加数据量，但是其相比上面利用随机森林模型进行择时，效果较差。

总结

作为机器学习方法系统性介绍的第一篇，本文主要是尝试用一些简单易懂金融实例来展示这些复杂算法的潜在应用情景。我们将机器学习模型分成了监督学习、非监督学习、深度学习及其他，本篇就是针对监督学习模型。

总结下来，涉及到的主要模型有：惩罚回归模型、K 最近邻、卡尔曼滤波、逻辑回归、支持向量机、随机森林及隐马尔科夫模型。惩罚回归模型在金融领域的应用较为成熟和普遍，主要是进行变量的去线性和去干扰。K 最近邻、支持向量机、逻辑回归等近些年来也被普遍用于选股和择时，形成了较为清晰的应用逻辑，在结合大数据的层面上，可以有效增强传统量化投资。卡尔曼滤波和隐马尔科夫模型较为成熟的应用均是在其他行业领域，其在投研上有应用潜力，但是目前并未发现在投研的应用情景上较常规方式更为突出的表现。

本文中所有实例均是为了方便具体化各种模型应用环境，并且熟悉模型的操作过程，实际投资中可以结合最前面表格中的应用情景做更为深度的延展和探索。

投资评级说明

行业评级	报告发布日后的 12 个月内行业股票指数的涨跌幅度相对同期沪深 300 指数的涨跌幅为基准，投资建议的评级标准为：	
	看　　好：	相对表现优于市场
	中　　性：	相对表现与市场持平
	看　　淡：	相对表现弱于市场
公司评级	报告发布日后的 12 个月内公司的涨跌幅度相对同期沪深 300 指数的涨跌幅为基准，投资建议的评级标准为：	
	买　　入：	相对大盘涨幅大于 10%
	增　　持：	相对大盘涨幅在 5%～10%之间
	中　　性：	相对大盘涨幅在-5%～5%之间
	减　　持：	相对大盘涨幅小于-5%
	无投资评级：	由于我们无法获取必要的资料，或者公司面临无法预见结果的重大不确定性事件，或者其他原因，致使我们无法给出明确的投资评级。

联系我们

上海
浦东新区世纪大道 1589 号长泰国际金融大厦 21 楼（200122）

武汉
武汉市新华路特 8 号长江证券大厦 11 楼（430015）

北京
西城区金融街 33 号通泰大厦 15 层（ 100032）

深圳
深圳市福田区福华一路 6 号免税商务大厦 18 楼（518000）

重要声明