**应用机器学习模型和Alpha策略进行股票选择**

韩昊阳

关键词：机器学习；IC标准；多因子分析；选股；模型选择。

1. **背景**

Alpha策略是指基金经理通过一系列的选股标准，选择市场表现优良，基准收益率能够跑赢大盘的股票组合。通过机器学习的方法让模型自动输出对未来的预测，可以有效地降低人力成本和筛选难度。在分析股票走势过程中，IC（Information Coefficient）、IR（Information Ratio）、Sharpe Ratio、Mean Return、MDD、DDperiod等系数因代表性强而被广泛用于股票分析中。本文通过使用机器学习的方法，应用于多因子股票分析中，达到机器自动选股的目的。

1. **实验数据与方法**

2.1 实验数据：

本实验使用多因子选股模型，对1829只股票，40320分钟的每分钟股票价格进行分析。本研究中使用21种不同因子作为自变量，并通过T分钟前后的价格变化比例构建因变量y（该自变量其实为alpha指数），其构建公式如下：

Y(i+T,j) = log(Price(i+T,j)/Price(i,j))-avg(log(Price(i+T)/Price(i)))（1）

此时构建得到的因变量y应为特定股票在某时间内相对于所有股票的相对收益率大小。其数值越大，反映其alpha因子的数值越强。

2.2 实验方法:

本实验使用Python语言，采用numpy、 pandas等程序包进行数据处理，sklearn等进行机器学习模型导入，matplotlib、seaborn等进行数据可视化操作。本实验运用标准机器学习应用的方法进行数据分析。其流程为：

首先对21种不同因子和对应时间股票价格进行数据导入。考虑到多因为分析的因子种类较多，类型为np.float32，且保存格式为二进制文档，本研究使用numpy.memmap()函数导入数据，并通过数组合并和纬度积压的方法将散乱的自变量数据拼接到一起。

随后进行数据特征判断和输入数据预处理。通过计算自变量种各个因子的均值，方差，标准差，分位数，中位数和分布特征，对因子的分布规律，范围进行可视化。随后通过人工选择确定训练集/验证集的比例，并对自变量，因变量进行切分，得到形状为（34970,1082）和（5000,1082）的训练集和验证集。

接下来进行模型拟合。通过使用以下12种机器学习模型对自变量进行回归分析：

1. Linear Regression
2. Ridge Regression
3. Kernel Ridge Regression
4. Lasso Regression
5. Bayesian Ridge Regression
6. Polynomial Regression
7. Decision Tree
8. Random Forest
9. Bagging Regression
10. Gradient Boosting Decision Tree(GBDT)
11. AdaBoost
12. Support Vector Machine(SVM) Regression

在模型拟合过程中，如果拟合时间大于6000s，我们便终止模型并认为该模型在拟合过程中遇到了模型时间复杂度过高，或者模型不适用与此情况的问题。

接下来使用使用数据集训练模型，并计算各个模型回归得到的因变量预测值与因变量实际值差别的MSE（均方误差），RMSE（均方根误差，标准误差）, 和IC指数，作为其准确率（accuracy）的评价标准。其评价指标的构建方法分别为：

https://gss1.bdstatic.com/-vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D113/sign=55bcf11c4b086e066ea83b4a31087b5a/0d338744ebf81a4cbc2a81e4d02a6059252da687.jpg （2）

RMSE = sqrt[∑di^2/(n-1)] （3）

IC\_i = Pearson(y\_pred\_i, y\_test\_i) （4）

IC = list(IC\_i) （5）

IC\_index = mean(IC)/std(IC) （6）

最后将数据输出，重新整理为float32型数据并导回二进制文档，在平台中分析并得到得到不同模型的下列评判标准：

1. IC
2. IR
3. Sharpe Ratio
4. Mean Return
5. MDD
6. DD period

并以上述标准作为评判模型拟合成功与否的标准。

1. **实验步骤**
   1. 电脑系统配置

安装必须的：

1. Anaconda Navigator(python 3.6); https://www.anaconda.com

2. Numpy, pandas, sklearn, matplotlib, seaborn; pip/conda install xxx

3. Tensorflow, Pytorch, word2vec, autograd, cupy; pip/conda install xxx

4. chrome; https://chrome.google.com

5. wolf(bash simulator for windows);

6. oracle/wind;

7. VPN.

8. Cuda.

3.2 导入数据

注意数据的格式为dtype = float32 的二进制文件，将数据中有效的1082支股票提取出来，并将多音字合并。

* 1. 特征了解

此时应先提取随机的单因子，了解其统计特性并可视化表示。

* 1. 变量合并，训练集/测试集拆分

拆分结果为35000：5000.

* 1. 模型拟合
  2. 精确度计算

使用MSE/RMSE/IC标准进行精确度计算，

* 1. 导出数据，生成报告

1. **结果**

­4.1 模型分析结果

4.1.1 模型选择、拟合时间

在本实验中我们对所使用的12种模型分别进行数据拟合，倘若拟合时间大于6000s（1 hours），说明模型复杂度太高或模型不适用与本种数据类型（特征数量少，但数据量极大）。在实验过程中我们将SVM、Bagging、KRR（Kernel Ridge Regression）和Polynomial模型舍弃，并将拟合所需要时间由少到多排列如下（表1）。需要注意：拟合时间会随机器性能和重复次数改变，但时间相对关系不变。由试验结果可知，拟合效率最高的的是岭回归（Ridge Regression），线性回归（Linear Regression）和决策树（Decision Tree）,而集成学习（Ensemble Method）和核岭回归（Kernel Ridge Regression, KRR）的时间复杂度太高，所需要的资源量太大，故而在本实验中被舍弃。

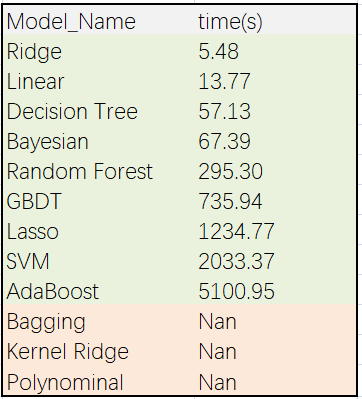


表1：模型选择结果（绿色：可用模型；红色：不可用模型）

4.1.2 模型MSE/RMSE数值

均方误差和均方根误差是判定模型拟合效果是否优秀的重要评价标准，考虑到过大或者过小的原始数据会让MSE数值变得过大或者过小，此时引入RMSE评判标准也是有意义的。由MSE/RMSE的定义可知，MSE/RMSE的数值越小，预测数据离真实值得偏差越小，因而模型描述实验数据有更好的精确度。在本实验中，我们发现随机森林（Random Forest），岭回归和贝叶斯岭回归（Bayesian Ridge Regression）的拟合效果最好，而支持向量机的拟合效果最差，除此以外，线性回归，Lasso回归和Adaptive Boosting（Adaboost）的拟合效果也相对较差。

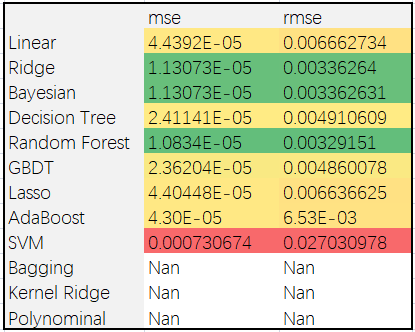


表2：MSE/RMSE模型拟合结果

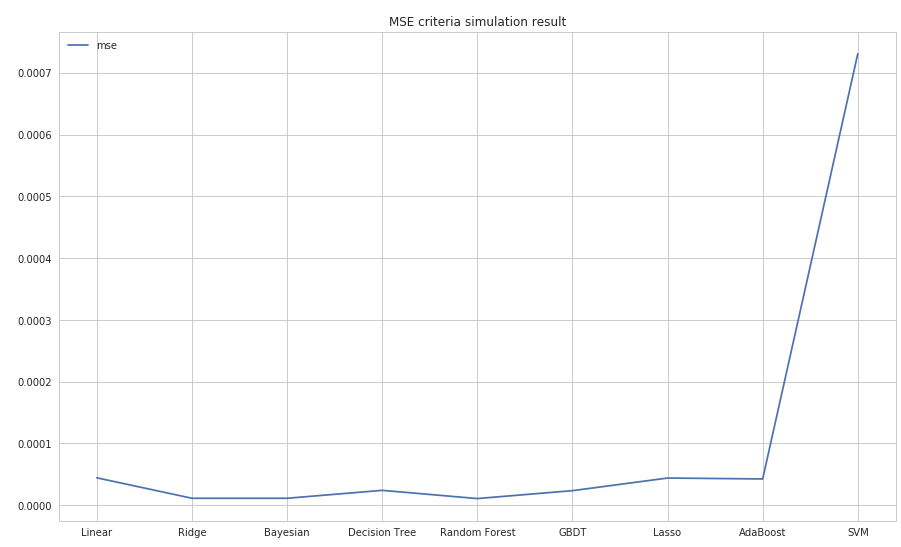


图1：MSE值条形图

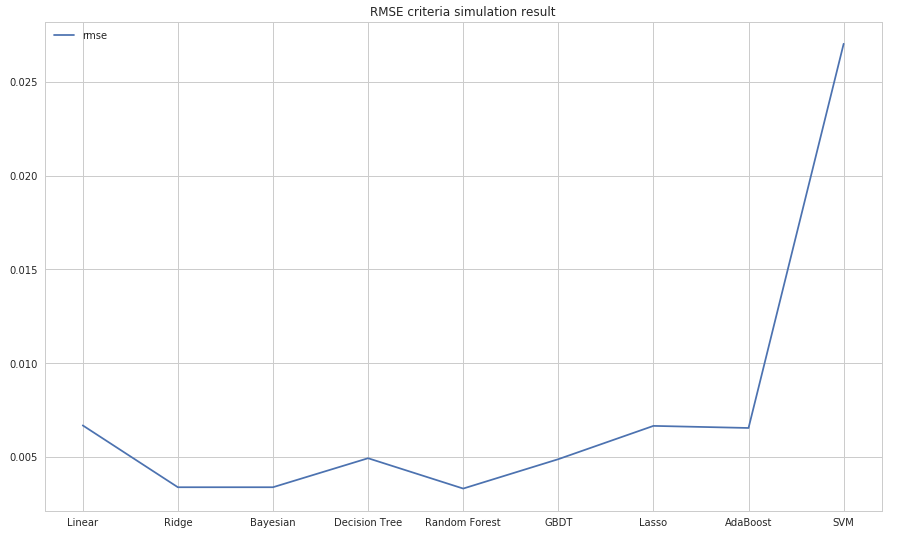


图2：RMSE值条形图

4.1.3 模型IC\_index数值

Information Coefficient和MSE/RMSE的评估方法不同，它是通过计算每分钟拟合值和真实值的Pearson Correlation数值，并对数值进行规范化，来衡量预测结果的优劣。显然，数值越大，拟合值和实际值越贴合，故而数值越大模型拟合效果越好。从这个角度出发我们可以看到，随机森林是毫无疑问最出色的回归器，而线性、Lasso和Adaboost回归的效果均不佳。除此之外我们应当认识到，GBDT的回归效果同样超出其他回归器，而贝叶斯方法下的岭回归和普通的岭回归没有本质区别。

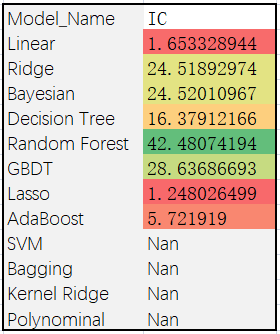


表3：IC\_index模型拟合结果

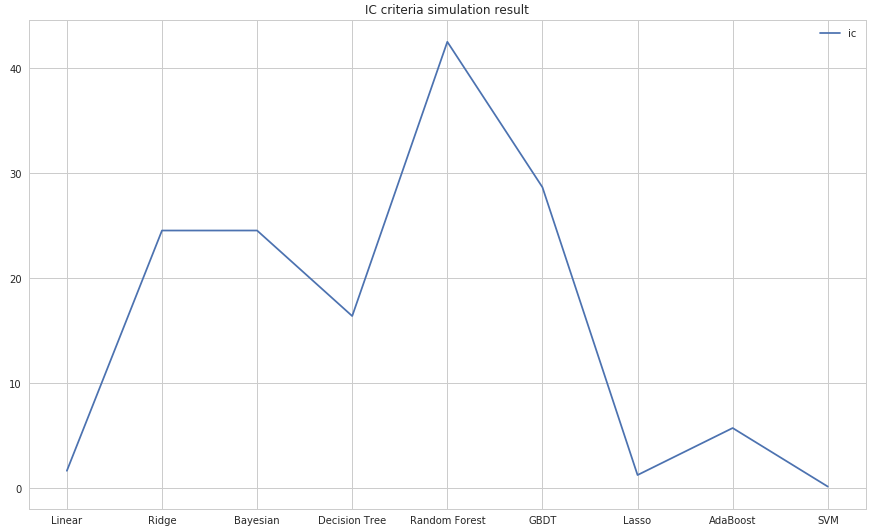


图3：IC值条形图

4.2 拟合数据分析结果

注意：在写报告时SVM还没跑出来，但是它最后跑出来了，故而在数值分析里没有SVM的结果。

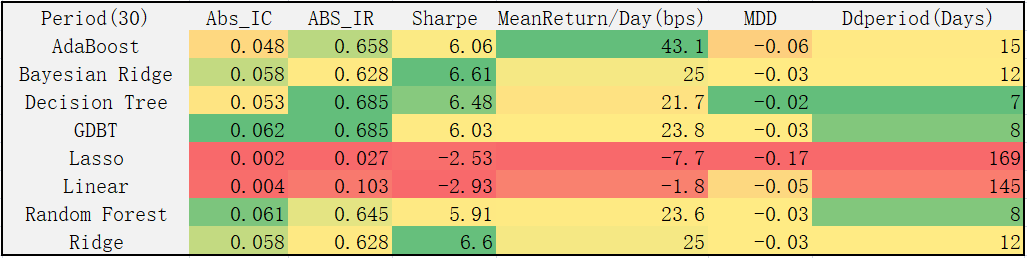


表4.各统计量拟合值

在评估和对比拟合结果时，我们发现不应当将各个因子分开讨论，而应当作为一个整体进行评估。IC和IR值反映信息量的相互关联程度，其数值越大，保留的信息越多；夏普比率是风险调整后的收益率，其大小虽然不能直观反映模型好坏，但可以表示模型对基金总体的期望程度。MDD是最大回撤，其数值越大越好；Mean Return per Day是模型对股票平均收益的预测；而DD period则是回撤所需要的天数，其时间越短越好。

在评价模型时，我们应当将真实的alpha值和模型拟合alpha值同时输入系统，通过比较各个模型和真实情况的统计值差异，来判断模型的优劣，这是本周实验没有考虑到的，在下周的实验中应首先完成这个部分的工作。

我们发现，尽管数据各自有各自的偏向，但毫无疑问Lasso回归和线性回归的信息相关系数不够出色，与此同时随机森林的各个系数都比较平均，应当和真实值相接近。

1. **讨论**
2. 关于几个无法跑出结果模型讨论：

在本次实验中，集成模型（Bagging），核岭回归和多项式回归未能跑出结果。其中多项式回归不是一种回归方法，而是增加自变量中feature数量和维度的一种手段，因而应当与其他模型一并使用。而本实验具有数据集极大，feature较多的特点，人为增加特征数量无益于模型拟合得出结论。

核岭回归是将核函数应用于回归分析的一种技巧，其在数据量为中等大小时速度较快，随后复杂度会急剧增加。因此，对于数据量为上百万的高频交易实验，岭核回归的效果不佳。

Bagging模型拟合速度极慢的原因尚不清楚，下周会进行其他种类的集成模型拟合，判定是bagging模型本身的原因还是集成模型在大规模数据集上的应用效果均不理想。

1. 增加testing set和y\_hat预测比率的方法。

本次试验选择35000:5000的训练/测试数据分离比，因而有效的测试数据和预测y值数量较少。在接下来的模型训练过程中，我们可以通过迭代预测的方法，将每5000个数据输入模型中来预测后5000个数据数值，提升测试集的数量为35000行，并由此进一步得到更明显的各个模型的训练结果差别。

1. 可以增加的实验方法：

在选股过程中，我们暂时采用了传统机器学习模型。在下一周的实验过程中，我们可以增加如下深度学习模型：

1. Multilayer Perception(nn/FFNN/Fully connected network)
2. Simple-RNN
3. GAN-RNN
4. GRU-RNN
5. LSTM-RNN

和机器学习优化方法：

1. Normalized Gradient Descent
2. Stochastic Gradient Descent
3. Momentum Gradient Descent

以及以下降维方式：

1. PCA
2. LDA
3. LLE
4. Laplacian Eigenmaps

以提升模型的准确率和速度。

1. 参数调整的方法

使用sklearn包进行数据拟合的过程中，参数选择是非常重要的一环，正确的参数选择可以有效提升模型运作速度，特别是针对Tree式模型和Random Forest模型。我将其一般的参数设定放在下面:

对于DecisionTreeRegressor:

1. 选择splitter = ‘Random’来让机器更随机而快速地选择分叉节点；
2. Max\_features = ‘log2’来最大程度的阻止决策树生成过多特征；
3. Max\_depth = 10 来抑制子枝的最大生成数量；
4. Min\_samples\_split = 10来降低对每个feature都进行分类的可能性；
5. Min\_samples\_leaf = 10来保证最多生成子枝数量。

额外的，对于RandomForestRegressor：

1. n\_jobs = 10增加并行的cpu工作数量，提升效率。
2. **附录**
3. 各模型平台拟合图像
4. 统计值原始记录
5. 原始代码及jupyter notebook
6. 各模型回归因子结果（较大，可略去）
7. 原始数据（较大，可略去）