**机器学习多因子策略**

**前言**

在二级市场的量化策略中，多因子策略称得上是最早被创造但同时也是变化最多的投资策略之一，好的因子意味着长期稳定的收入，因此能够发现一个好的因子是每位宽客的一致愿望。多因子策略理解起来并不复杂，实现起来却可以通过多种不同的渠道，从而带来不同的表现，本文基于经济学家在2005和2012发表的两篇论文，探究市值因子在我国二级市场中的表现。

与传统多元线性回归不同，本文采用了支持向量回归（Support Vector Regression）和随机森林回归（Random Forest Regression）两种机器学习算法得到因子，实证表明，SVR算法下得到的因子带来的收益率显著优于传统多元线性回归，而随机森林算法下得到的因子带来的收益率虽略低于多元线性回归，但在经历回撤时拥有更加稳定的表现。

**多因子策略简介**

所谓因子，通俗来讲就是一个“标准”，这个“标准”决定我们选择哪些股票，决定我们什么时候买入卖出，决定我们交易的份额。而多因子，顾名思义就是根据多个“标准”确定我们的投资策略，最为经典的多因子模型之一当属Fama和French提出的三因子模型，而其基本思路为，我们假定股票的超额收益率可以由市场风险，市值风险，账面市值比三个因素决定，因此我们将股票一个时期内的超额收益率对这三个因素进行回归，再将回归得到的参数对本时期内的超额收益率进行预测，比较真实值和预测值。如果预测值大于真实值，也就是说理论上的超额收益率应该大于当前的超额收益率，那么根据有效市场假说，未来的超额收益率应该上升，反之亦然。

除了最简单的三因子模型外，Fama和French还在2013年提出了五因子模型，对个股的超额收益率进行了更加细致的解读。

**市值解释因子简介**

2005年，在哈佛大学Matthew Rhodes-Kropf教授以及杜克大学S. Viwanathan和David T. Robinson教授合著的论文《Valuation Waves and Merger Activity: The Empirical Evidence》中，股票市值被分解为如下的三因子模型：

m=α0IND+α1b+α2ln(NI)++α3I(<0)ln(NI)++α4LEV+ε

m=α0IND+α1b+α2ln(NI)++α3I(<0)ln(NI)++α4LEV+εm=α0IND+α1b+α2ln⁡(NI)++α3I(<0)ln⁡(NI)++α4LEV+ε

其中，INDINDIND为行业虚拟变量矩阵（若该股属于某行业，则将这一行业虚拟变量的值设为1，其他行业虚拟变量的值设为0），mmm为个股的对数市值，bbb为个股的对数净资产，NININI为公司净利润，这里注意，为了区分净利润的正负，我们增加了一个代表正负的虚拟变量，当且仅当净利润为负时，这个虚拟变量的值为1，并且我们取净利润绝对值的对数作为自变量。LEVLEVLEV为公司的财务杠杆（负债除以资产）。作者用这个模型对特定美股进行拟合，平均拟合优度超过80%，也就是说，这个三因子模型可以被认为是有效的。

2012年，马里兰大学的Charles R. Hulten教授及其博士生Janet X. Hao发表的论文《The Role Of Intangible Capital in the Transformation and Growth of the Chinese Economy》中，发现上市公司的市值与其净资产、开发支出、组织资本以及市盈率密不可分，选取开发支出力度大的公司进行下述回归，拟合优度可以达到94%。

m=α0+α1b+α2RD+α3O+α4PE+ε

m=α0+α1b+α2RD+α3O+α4PE+εm=α0+α1b+α2RD+α3O+α4PE+ε

其中，α0α0α0为年度虚拟变量矩阵，mmm和bbb的含义与上文相同，RDRDRD为对数开发支出，OOO为对数组织资本，PEPEPE为市盈率。  
综上所述，我们可以认为，股票在某一时间点的市值可以被多个因子解释，我们把这些因子统称为市值解释因子，由于股票价格等于股票总市值除以股票数量，因此在股票数量不变的情况下，我们可以用市值解释因子对股票价格进行预测。

**模型搭建与因子选择**

从上文我们能够看出，某个时间点上，股票的市值可以被多个因素解释。根据多因子策略的基本思路，我们在截面上将市值对多个因子进行回归，得到的残差值越小，说明股票市值向下偏离其理论值越严重，也就意味着该股票未来上涨的可能性越大。  
而在选择因子方面，我们结合两篇文献，选择了以下财务指标作为解释因子的自变量：对数净资产，对数净利润，公司财务杠杆，营业收入增长率，对数开发支出，以及行业虚拟变量。

在回归方法上，我们首先采用了最传统的多元线性回归：

m=α0IND+α1b+α2ln(NI)++α3I(<0)ln(NI)++α4LEV+α5g+α6RD+ε

m=α0IND+α1b+α2ln(NI)++α3I(<0)ln(NI)++α4LEV+α5g+α6RD+εm=α0IND+α1b+α2ln⁡(NI)++α3I(<0)ln⁡(NI)++α4LEV+α5g+α6RD+ε

其中mmm为个股的对数市值，INDINDIND为行业虚拟变量矩阵（若该股属于某行业，则将这一行业虚拟变量的值设为1，其他行业虚拟变量的值设为0），bbb为个股的对数净资产，NININI为公司净利润，这里注意，为了区分净利润的正负，我们增加了一个代表正负的虚拟变量，当且仅当净利润为负时，这个虚拟变量的值为1，并且我们取净利润绝对值的对数作为自变量。LEVLEVLEV为公司的财务杠杆（负债除以资产），ggg为营业收入的增长率（年度同比），RDRDRD为对数开发支出（若没有则为0）。

除了多元线性回归，我们尝试了包含随机网络、Adaboost、支持向量机回归的多种机器学习回归算法，发现随机森林回归和支持向量机回归表现较好，接下来我们对这两种机器学习算法进行简要介绍。

**随机森林回归**

随机森林是一种比较新的机器学习模型，它的基本思路是：多次有放回地在样本集中抽取样本，从全体特征中选择部分特征对模型进行训练，每次训练的结果对新实例进行预测时，随机森林需要整合其各个决策树的预测结果。解决回归问题时，每个树得到的预测结果为实数，最终的预测结果为各个树预测结果的平均值。

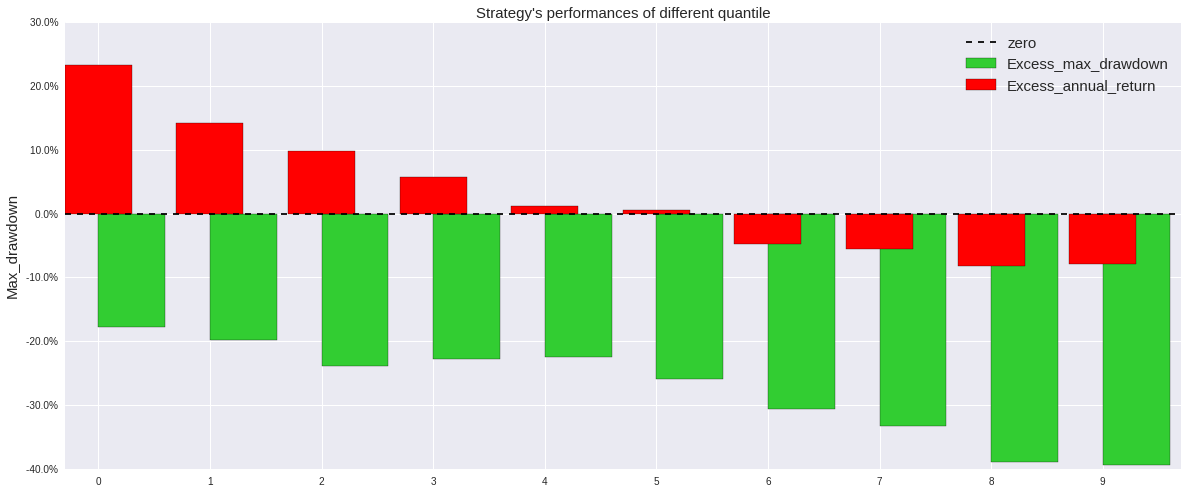
**支持向量回归**

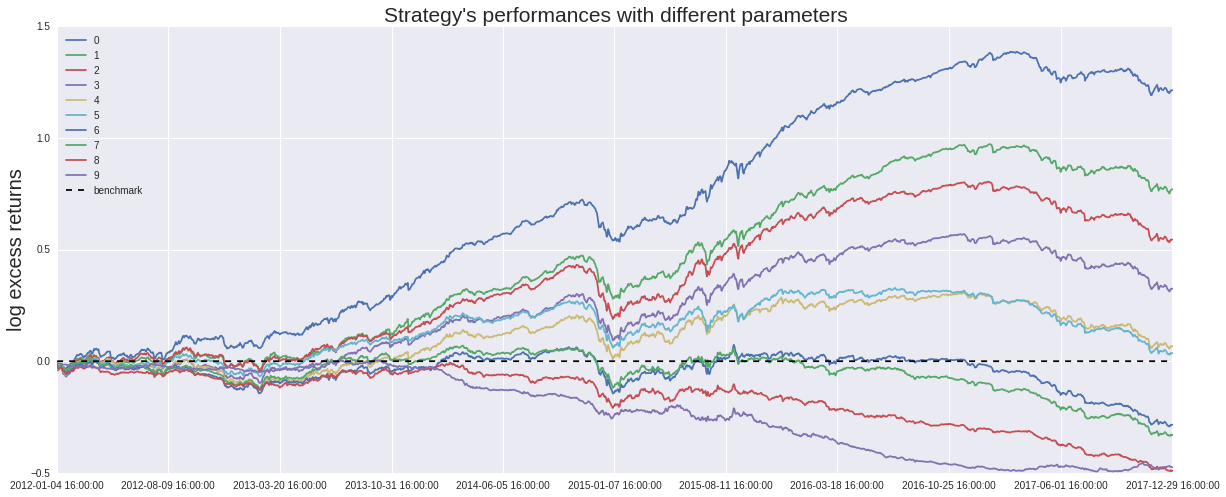
支持向量回归（SVR）是支持向量机在回归方面的应用，基本思路大致相同，关于SVM算法详细解释可以关注聚宽量化课堂的文章：  
两个模型的数学原理都有一定难度，所幸python的sklearn库提供了两种算法的程序包，对算法本身没有兴趣进行深究的读者，可以直接应用sklearn中的函数。我们将上述用来线性回归的自变量作为特征输入模型进行回归，并对回归结果进行分析。

**因子有效性验证及策略设计**

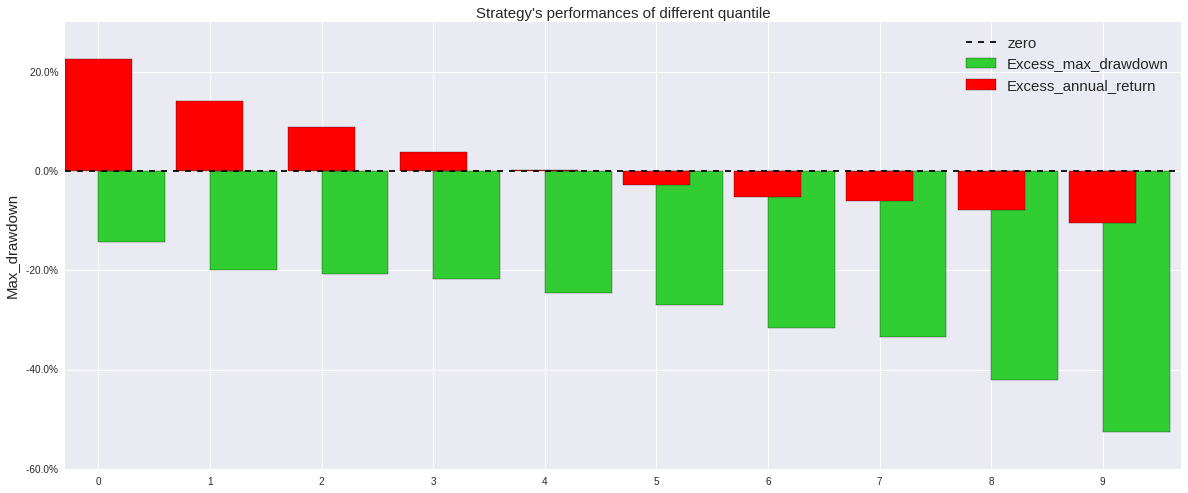
我们采用上述三种方法，从聚宽数据库中调取相应数据进行回归，其中行业虚拟变量矩阵采用申万一级行业划分，股票池选择中证全指中的所有股票，回测时间为2012年1月1日至2017年12月31日。我们每十个交易日进行一次回归，并将回归得到的结果用到当日的股票上，计算预测值与真实值之间的差距（残差），按照残差从小到大的顺序对股票进行排序，依据不同分位数，分十档买入，每十个交易日进行一次调仓，将不在该档的股票卖出，得到如下分档超额收益率，最大回撤比率以及对数收益曲线。

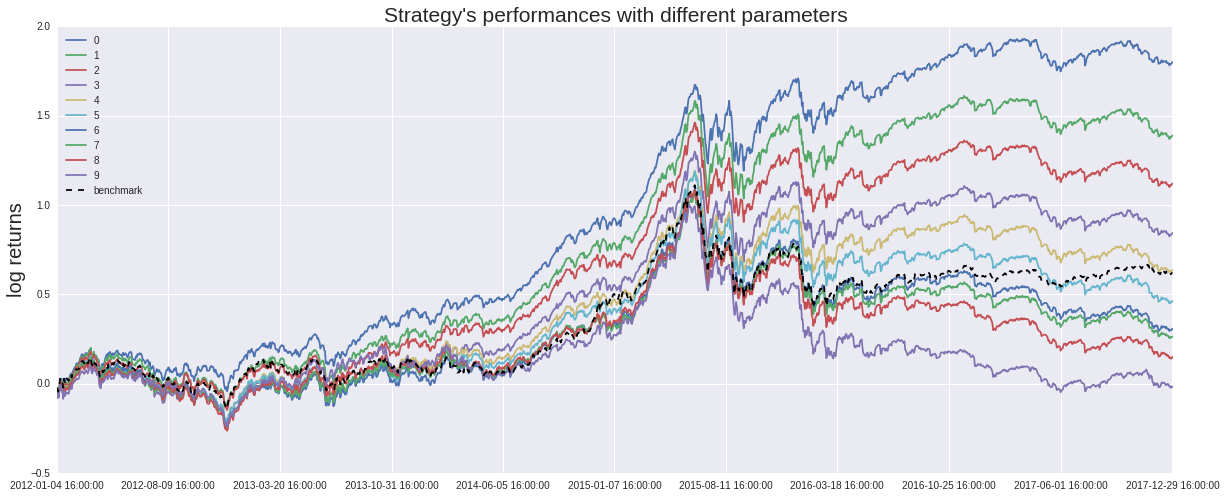
**线性回归**





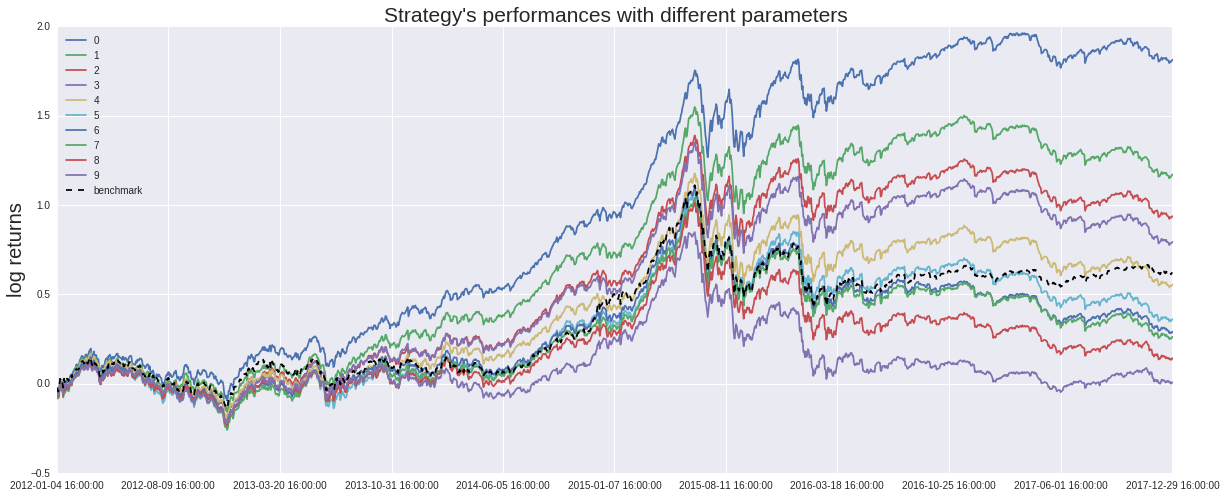
**随机森林回归**





**支持向量回归**





从图像中我们可以清楚地看到，无论基于哪种回归算法，我们都能得到单调的分档超额收益率，而其对数收益曲线界限也非常明朗，这证明了该因子在该段时间内是显著有效的，这也就意味着如果我们采用纯多头策略，买入残差排名更靠前的少数股票，将会拥有更大的套利空间，本着这样的想法，我们修改策略为，对中证全指股池中的全部股票进行回归，调仓周期为十天，每次调仓时买入（或继续持有）因子排名在前十位的股票，若股票的排名偏离这个区间，则卖出。

对这一策略，利用三种不同的回归算法进行回测，得到2012-2017年的收益状况如下：

**线性回归**



**随机森林回归**



**支持向量回归**



**总结**

从策略的回测结果我们可以清楚地看到，长期来看，市值因子的年化收益率均在50%左右，表现良好，其中，以SVR算法下表现最为良好。同时，线性回归在2017年初遭遇较大回撤，随机森林回归近来表现也略有下滑，这与2017年众多因子失效的市场环境也是吻合的。我们在此提供详细的SVR策略的回测结果，有兴趣的读者可以深入学习。