DigQuant

题目：通过机器学习优化股票多因子模型

——泰迪杯数据挖掘挑战赛

DigQuant 点宽

什么是因子

什么是因子

仅在BP因子库中，我们就有500多个因子，12个大类因子

基础科目与衍生类

1. 净营运资本（NetWorkingCapital）
2. 净债务（NetDebt）
3. 留存收益（RetainedEarnings）
4. 毛利（GrossProfit）
5. 企业自由现金流量（FCFF）

什么是因子

仅在BP因子库中，我们就有500多个因子，12个大类因子

基础科目与衍生类

1. 净营运资本（NetWorkingCapital）
2. 净债务（NetDebt）
3. 留存收益（RetainedEarnings）
4. 毛利（GrossProfit）
5. 企业自由现金流量（FCFF）

收益风险类

1. 120日方差（Variance120）
2. 20日方差（Variance20）
3. 60日方差（Variance60）
4. 个股收益的120日峰度（Kurtosis120）
5. 个股收益的20日峰度（Kurtosis20）

为什么要用多因子模型

什么是定价因子

定价因子

CAPM模型

简化的CAPM模型

$$\mathbf{E}(R) = E(R_m) * \boldsymbol{\beta}$$

$\mathbf{E}(R)$为股票或者投资组合的期望收益率

$E(R_m)$为市场组合的收益率

β是股票或投资组合的系统风险测度

定价因子

CAPM模型

APT模型

APT套利定价模型

$$r_j = \sum_{k=1}^{n} X_{jk} * f_k + u_j$$

X_{jk}: 股票j在因子k上的因子暴露（因子载荷）

f_k: 因子k的因子收益

u_j: 股票j的残差收益率

Fama-French三因子模型

Fama-French三因子模型概述

Beta　市值　估值

四因子

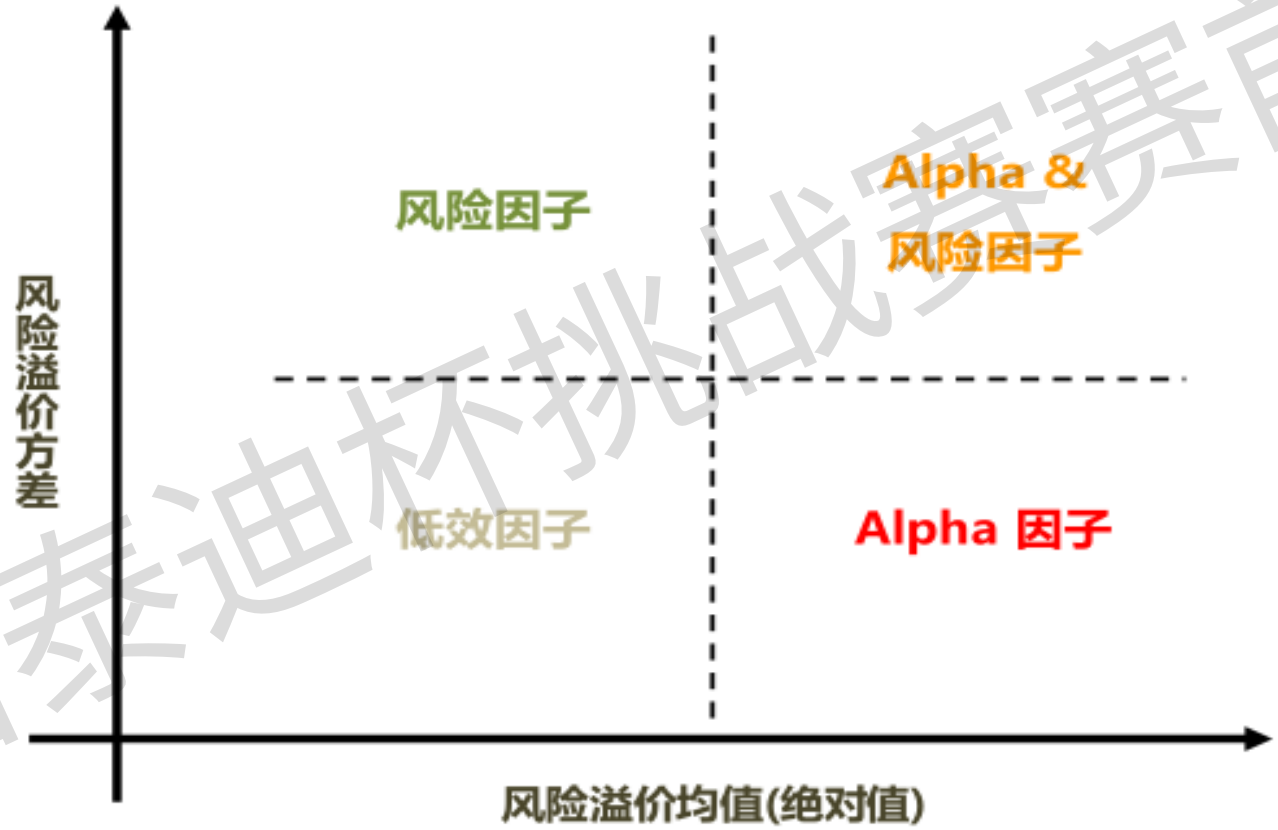
五因子模型

三因子模型概述

$$r_{i,t} = \alpha_{i,t} + \beta_i MKT_t + s_i SMB_t + h_i HML_t + \varepsilon_{i,t}$$

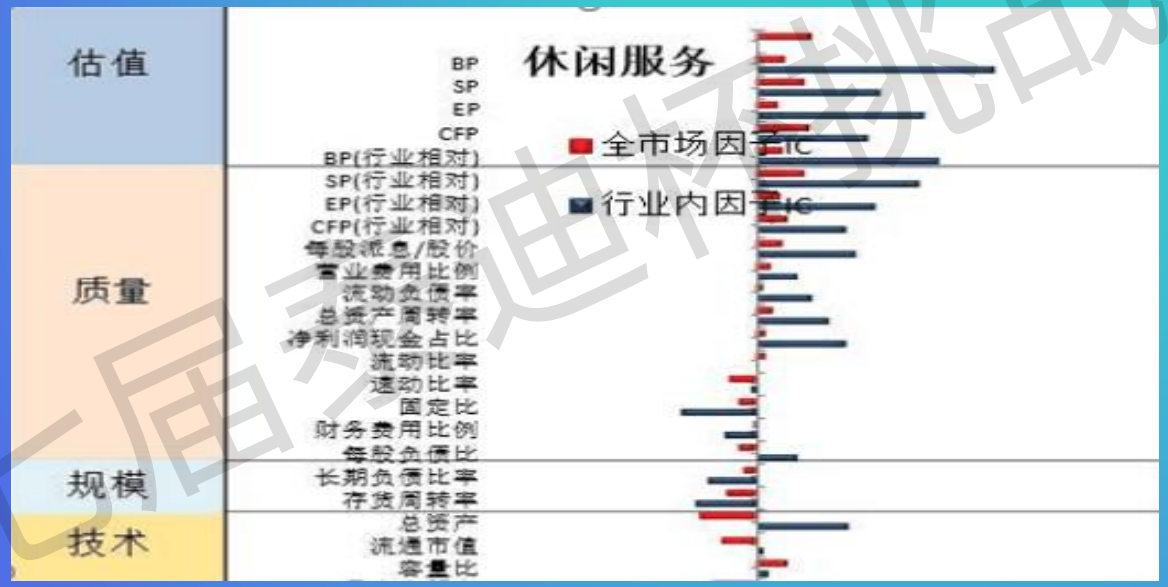
什么是Alpha因子

Alpha 因子 与 风险因子



多因子模型的现实背景

- 对国内来说，基于财务因子（比如市盈率，市值等）及长周期的量价因子（比如月度反转，月度成交量等）为主要因子的传统多因子模型在 A 股过去获得过较为稳健的超额收益。



多因子模型面临的挑战

- 由于A 股市场存在的明显的风格切换（比如2017年下半年从传统的小市值风格切换到只有极少数大市值股票上涨，而绝大部分股票下跌），导致传统多因子模型的稳定性及有效性受到较大考验。

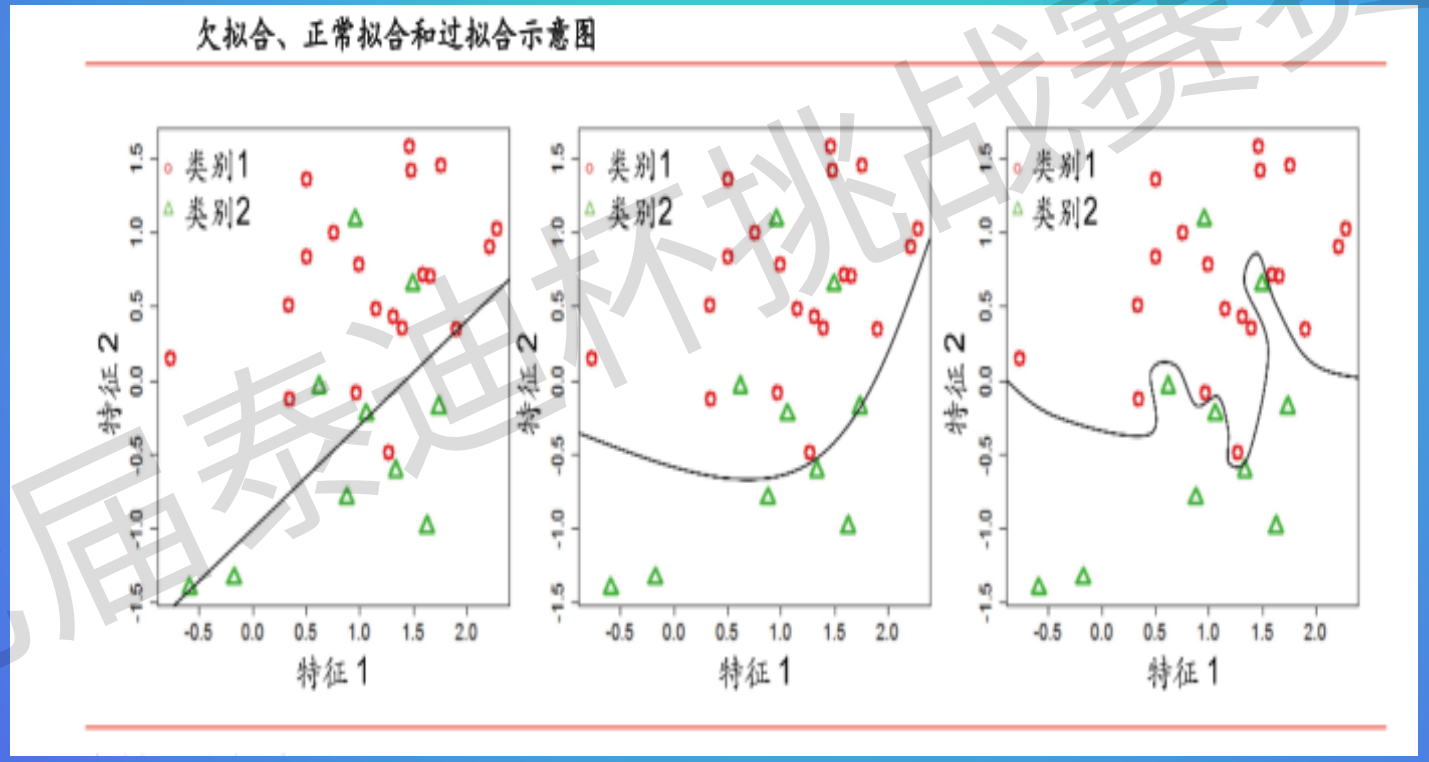
机器学习方法的直观印象

线性回归

$$r_{i,t} = \alpha_{i,t} + \beta_i MKT_t + s_i SMB_t + h_i HML_t + \varepsilon_{i,t}$$

机器学习方法

Tom M.Mitchell(1997)给出机器学习引用最多的定义：对于任务T和性能度量P，如果一个程序在任务T上以P衡量的性能随着经验E而自我完善，那么称这个程序在从经验E中学习。



提升多因子模型效果的新方法

相比传统线性多因子模型（通常是线性加权因子的得分），机器学习算法，能够通过其非线性的表达，更加精细地捕捉因子的信号，从而发现市场局部无效性，获取超额收益。因此越来越多人开始尝试利用机器学习技术，挖掘因子信号，试图获得稳定的alpha收益。

提供数据及交易平台

这里提供Auto-Trader策略回测交易引擎

Auto-Trader， 提供A股历史交易价格及各大类因子数据

支持语言： python / matlab

问题综述

使用2016年1月1日至2018年9月30日的中国A股股票历史数据，挖掘给定的大类股票因子中较优的因子。

然后，通过机器学习算法，提升这些因子的表现，并分析不同机器学习算法提升因子表现的优劣。

使用Auto-Trader策略研究回测引擎对2016年1月1日至2018年9月30日的数据进行策略回测。

问题一：挖掘大类因子中最优因子

利用Auto-Trader中的情绪类、成长类和动量类等12大类因子的日频数据，任意选择N个大类因子，对选股空间（可以是全A、沪深300、中证500等）分别做单因子策略研究和绩效分析， 找出20160101-20180930大类因子中，年化Sharpe最高的因子。

问题二：使用机器学习算法提升因子表现

基于机器学习算法（ 随机森林，支持向量机，Adaboost等 ）对问题一中挑选的因子进行增强。

对比基准：将问题一中获取的因子等权重线性加权得到基准效果。

比较运用机器学习算法后的选股策略与对比基准的改进。

回测时间：20160101-20180930

问题三：控制机器学习模型风险

在解决问题二的基础上，对获得的机器学习增强方法，进一步风险控制，控制组合相对基准对冲指数的最大回撤

基准对冲指数：沪深300指数，上证50指数，中证500指数中任意一只

风险控制方法：包括但不限于使用风险模型（均值方差，或者barra），择时模型，行业中性等方法控制最大回撤。

回测时间：20160101-20180930

Thank you!