第七届“泰迪杯”数据挖掘挑战赛——

A 题：通过机器学习优化股票多因子模型

Fama 通过分析美国市场几十年的数据发现，美国股市绝大部分可以被市值、估值以及市场收益 3 个因子解释，并因此获得了 2013 年诺贝尔经济学奖。Fama 的工作开启了通过因子化分析股市获取超额收益的先河，此后学术界及业界不断地寻找其他能获取超额收益的因子及其组合和风险控制的方式。

在我国，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 股市场曾经获得过较为稳健的超额收益，但是由于 A 股市场存在明显的风格切换（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的风格），传统多因子模型的稳定性及有效性受到了较大的考验。

相比传统的线性多因子模型，机器学习算法能够通过对因子的非线性表达，捕捉到更加精细的市场信号，获取较为稳健的超额收益。

根据 2016 年 1 月 1 日至 2018 年 9 月 30 日我国 A 股市场的数据（数据提取方式见附录 2），筛选出各大类股票因子中较优的子因子。在此基础上，分析不同的机器学习算法对提升这些因子的等权重线性模型表现的优劣，并使用“Auto-Trader 策略研究回测引擎”进行策略回测（初始资金为 1000 万元整，手续费为双边千分之 3，每月月初调仓）。

可以从以下角度入手进行分析：

(1) 利用 Auto-Trader 中各大类因子（见附录 3）的日频数据（数据提取方式见附录 4），分别做单因子策略研究和绩效分析，挑选出使得年化夏普比率（Sharpe ratio）最优的各个大类的因子。

(2) 基于机器学习算法对 (1) 中挑选的因子，进行增强，利用 2016 年 1 月 1 日至 2018 年 9 月 30 日的数据进行选股和回测，比较不同机器学习算法选股策略与等权重线性模型选股策略之间年化夏普比率的优劣。

(3) 对选股策略进行风险控制，要求将最大回撤控制在 10%以内，重新完成 (2)。

注：除提交论文外，参赛队还须提交策略的回测报告，提交方式详见附录 7。

参考文献

[1] Aurélien Géron，机器学习实战：基于 Scikit-Learn 和 TensorFlow，机械工业出版社，2018.9.

[2] 李航，统计学习方法，清华大学出版社，2012.3.

[3] 林晓明，陈烨，华泰多因子模型体系初探—华泰多因子系列之一，2016.9.21.

[4] 林晓明，陈烨，人工智能 1：人工智能选股框架及经典算法简介，2017.6.1.

[5] 罗军，胡海涛，大浪淘金，Alpha 因子何处寻，2011.8.15.

附录 1　名词解释

夏普比率：用来衡量产品风险收益的相对表现。夏普比率为正值，说明在衡量期内产品的平均收益率超过了无风险利率。同时夏普比率越大，说明产品单位风险所获得的风险回报越高。当夏普比率为负数时，按大小排序没有意义。一般采用一年期定存利率作为无风险收

益率。计算公式为

$$S_m = \frac{\overline{REX}}{\sigma_{REX}}$$

其中 S_m 表示单位时间产品的夏普比率，$\overline{REX}$ 表示单位时间的超额收益率平均值，σ_{REX} 表示单位时间超额收益率的标准差。

超额收益率平均值计算公式为

$$\overline{REX} = \frac{1}{n}\sum_{t=1}^{n}(R_t - R_f)$$

其中 R_t 表示产品收益率的序列数据，R_f 表示市场无风险收益率，n 表示根据时间频度决定的收益率个数。

年化夏普比率计算公式为

$$S_A = S_m\sqrt{N}$$

其中 S_A 表示产品收益率的年化夏普，S_m 表示单位时间产品的夏普比率，N 表示不同周期的年化因子，计算周期为日，相应的 N 为 250。

风险控制：控制组合相对基准（基准可以是 hs300，sz50，zz500 指数）的回撤，包括但不限于使用风险模型、择时模型、行业中性等方法控制最大回撤。

附录 2　获取 A 股代码及价格

这里取数据示例以 python 为例。

获取 A 股交易代码：traderGetCodeList

获取 2018 年 9 月 30 日沪深 300 指数的成分股及权重：

A=get_code_list('hs300',date='2018-09-30')

设置 date 参数可以获取指定日期的成分股和权重；不设置 date 参数，默认取最近一个交易日的成分股和权重。

```
In [31]: A
```

Out[31]:

	code	name	block_name	weight
0	szse.000001	平安银行	payh	0.976
1	szse.000002	万科A	wkA	1.228
2	szse.000060	中金岭南	zjln	0.111
3	szse.000063	中兴通讯	zxtx	0.420
4	szse.000069	华侨城A	hqcA	0.197
5	szse.000100	TCL集团	TCLjt	0.279
6	szse.000157	中联重科	zlzk	0.156

获取 2017 年 6 月 25 日至 7 月 26 日平安银行和万科 A 的 bar 数据：

day_data=get_kdata(target_list=['SZSE.000001','SZSE.000002'],frequency='day',fre_num=1, begin_date='2017-06-25',end_date='2017-07-26',fill_up=False,df=True,fq=1, sort_by_date=False)

```
In [33]: day_data
```

Out[33]:

	time	code	open	high	low	close	volume	amount	open_interest
0	2017-06-26 15:00:00	SZSE.000001	8.97944	9.11520	8.97944	9.01823	710769.0	6.63763e+08	NaN
1	2017-06-27 15:00:00	SZSE.000001	9.01823	9.10550	8.98914	9.07641	546016.0	5.09162e+08	NaN
2	2017-06-28 15:00:00	SZSE.000001	9.06671	9.20247	9.04732	9.14429	1168796.0	1.10244e+09	NaN
3	2017-06-29 15:00:00	SZSE.000001	9.14429	9.16368	9.08611	9.14429	488804.0	4.59810e+08	NaN
4	2017-06-30 15:00:00	SZSE.000001	9.11520	9.14429	9.02793	9.10550	499633.0	4.68004e+08	NaN
5	2017-07-03 15:00:00	SZSE.000001	9.11520	9.14429	9.05702	9.11520	388349.0	3.64466e+08	NaN

更多获取标的及价格的用法参见：**点宽网-AT 相关-Matlab API** 或者**点宽网-AT 相关-Python API**

附录 3 因子分类说明

常见因子分类可采用如下分类。

基础科目衍生类	净营运资本、净债务、留存收益、毛利等
质量类因子	产权比率、超速动比率、非流动资产比率、股东权益比率等
收益风险类因子	120 日方差、股价偏度、历史贝塔、历史波动等
情绪类因子	20 日成交量标准差、20 日平均换手率、20 日换手率与 120 日换手率之比等
成长类因子	营业收入增长率、总资产增长率、5 年收益增长率等
常用技术指标因子	MA10、MA120、MTM、DBCD 等
动量类因子	BIAS20、CMO、PVT 等
价值类因子	PE、PB、PCF、PS 等
每股指标类因子	EPS、EBIPTS、TORPS 等
模式识别类因子	十字暮星、吞噬形态、刺透形态、倒锤头等
行业与分析师类因子	12 月相对强势、分析师盈利预期、投资回报率预测等
特色技术指标	绝对价格振荡器、平均价格、均势指标等

以上为因子分类数据，具体可以在**点宽网—数据字典—BP 因子**中找到，参赛者也可以根据 **BP 因子**或**数据字典**中的其他数据来构建因子分类进行分析，但应明确说明构建因子分类的数据来源。

附录 4 获取股票因子数据

获取 2017 年 6 月 25 日至 7 月 2 日平安银行的 PE 和 PB 的因子数据：

Factor_code=get_factor_by_code(factor_list=['PE','PB'],target='SZSE.000001',begin_date='2017-06-25', end_date='2017-07-02')

```
In [3]: Factor_code
```

Out[3]:

	date	PE	PB
0	2017-06-26	7.0262	0.8504
1	2017-06-27	7.0715	0.8558
2	2017-06-28	7.1244	0.8622
3	2017-06-29	7.1244	0.8622
4	2017-06-30	7.0942	0.8586

获取 2017 年 6 月 30 日的因子数据：

Factor_day=get_factor_by_day(factor_list=['PE','PB'],target_list=['SZSE.000001','SZSE.000002'], date='2017-06-30')

```
In [5]: Factor_day
```

Out[5]:

	code	PE	PB
0	SZSE.000001	7.0942	0.8586
1	SZSE.000002	13.198	2.4186

附录 5　Auto-Trader 软件说明

编程语言：Auto-Trader 平台须运用 Matlab 软件或者 Python 软件

(1) Matlab

Matlab 版本：2013a 及以上版本

安装环境：Windows 7 或 Windows 10，64 位

建议配置：i5 8G 及以上内存

(2) Python

Auto-Trader 仅支持 64 位版本(x64)的 Python 3.5x，Python 3.6x 和 Python 3.7x

Auto-Trader 已经成为 PyPI 中的项目，安装好 Python 后，可以通过在 cmd 中输入 pip install atrader 安装 atrader。atrader 完全使用 SDK 的方式，用户使用时可以使用任意的 IDE，包括 Pycharm、NoteBook、VScode、Sublime 等等常用的 Python IDE 进行 Python 策略的编写。当使用 atrader 时，需要保持 Auto-Trader 客户端打开且正常登陆。

附录 6　赛题说明

1. 关于赛题数据说明

(1) 示例数据：2019 年 3 月 16 日公布 2016 年 1 月 1 日至 2016 年 12 月 31 日 A 股市场股票因子数据。

(2) 全部数据：2019 年 4 月 13 日公布 2016 年 1 月 1 日至 2018 年 9 年 30 日 A 股市场股票因子数据。

(3) 数据下载请登录 DigQuant 量化社区（网址：www.digquant.com.cn，进入【泰迪杯】栏目，建议使用 Google 浏览器），初始登录账号密码见竞赛组委会发出的竞赛报名成功邮件，社区和 Auto-Trader 客户端的登录名和密码一致。

(4) 数据的使用方法请查看 DigQuant 社区【泰迪杯】专栏中的《比赛数据包》。

2. 数据包使用方法的补充说明

数据包 SingleBPFactor_2016.zip 为 Zip 格式压缩文件，下载成功后，使用 Zip 解压软件解压后形成数据文件夹。进入文件二级文件夹（文件夹 I）中，将会看到有一个或多个文件夹（如：sse、szse 等）。创建 C:\Users\Public\Documents\Bitpower\AT\SingleBPFactor 文件夹（文件夹 II），将文件夹 I 中的所有子文件夹及其中的文件复制到文件夹 II。登陆 Auto-Trader 客户端后就可以使用这些数据文件。

注：答题过程中遇到数据或软件使用问题，请拨打 0755-86952080 客服热线。

附录 7　赛题结果提交说明

第一部分：作品提交

(1) 论文命名为“A 题”，附件命名为“作品附件”。

(2) 论文及附件内不得出现队号、学校、学院、队员以及指导教师姓名等任何相关信息，否则该作品视为无效作品。

(3) 请参赛队于 2019 年 4 月 26 日 16:00 之前在竞赛官网【提交作品】处提交论文（PDF 版，大小不超过 50M）及附件（论文正文（Word 版）、源数据（组委会提供的源数据除外）、过程数据、程序的压缩包，大小不超过 200M）。

第二部分：策略提交

策略的回测报告、Matlab 或 Python 源代码、执行脚本说明、调用的外部的代码包（若没有调用，则无需上传）通过 Auto-Trader 内置的“私有云策略池”上传提交。操作步骤如下：

(1) 策略在 Auto-Trader 上完成后，点击【同步到私有云】。

(2) 登录 www.digquant.com.cn，进入泰迪杯栏目页面，点击【提交策略】。

(3) 点击【从私有云选择策略】，选择拟提交的策略，若有使用外部函数和数据包，点击【上传附件】一并上传，最后点击【提交策略】。

(4) 每个参赛队只有第 1 位队员有权限提交策略，在 2019 年 4 月 26 日 16:00 之前可重新选择作品覆盖提交，系统默认以提交的最后一份作品为准；提交的策略需与论文一致，否则会影响成绩。

答案提交操作演示请登录 www.digquant.com.cn【泰迪杯栏目】中的赛前培训版块查看。