武汉大学《金融科技》课程作业 2

本次项目作业为智能量化投资，即利用人工智能技术开展金融市场投资。我们将搭建一套完整的智能投资框架，并运用中国股票市场数据进行回溯测试。

① 因子的含义及计算方法；

② 检验单因子；

③ 构建基于机器学习的多因子投资模型。

你可以按照每步指导逐步完成本作业，亦可自行设计框架完成。

0. 作业要求

- 提交报告一份（明确每个题目的答案），格式自拟，保持简洁；提交代码（不需要提交数据）。
- 本次作业涉及到大量数据，自行构建较为繁琐和耗时。步骤 1 中要求大家构建 2 个因子（亦可参照要求自行选取），其余可通过国泰安数据库下载，也可以从 GitHub 下载：https://github.com/WHUFT/ML-Quantamental。内有论文、数据和代码等，请根据自己的情况下载。
- 采用坚果云链接提交作业，按提示输入学号姓名后上传文件即可，或用地址：https://workspace.jianguoyun.com/inbox/collect/3f6de3b95dec4cc4be0639ff811a09ce/submit
- 截止时间：4 月 27 日 12:15（根据最后提交版本的时间戳）
- 迟交规则：迟交一周扣当次作业的 25%

1. 数据收集

根据基本面数据进行交易采用月度交易频率。首先需要准备所有个股的超额收益数据。国泰安数据中的收益来源为：

首页->数据中心->单表查询->股票市场系列->个股交易数据->月个股回报率文件->考虑现金红利再投资的月个股回报率

其次需要准备基本面因子数据。需要从国泰安(或万德等金融数据库)下载财务数据和量价数据，计算可得到月度因子数据(其他频率的数据可自行探索)。样本为所有 A 股股票，样本区间为 199612-201911，收益区间为 199701-201912。

作业需要同学们收集两个常见因子：因子动量 mom6 和资产市值比 AM，其他因子可从 GitHub 下载。Github 数据区间截止 2018 年底，可以将两个因子截至相同时间（2018 年底）。

两个因子的计算方式如下。

① 动量因子 mom6=使用过去 6 个月的收益计算，但剔除最近的一个月收益，要求数据不少于五个月：为了计算动量因子 mom6，需要准备股票过去 6 个月的收益数据，计算累积收益即可。

② 资产市值比 AM=季末总资产/季末流通市值：计算得到资产市值比，需要计算得到上市公司的季末总资产和季末流通市值。

查阅国泰安数据库说明可知，计算上述两个因子所需的基础数据如下：

① 过去 6 个月的收益：首页->数据中心->单表查询->股票市场系列->个股交易数据->月个股回报率文件->考虑现金红利再投资的月个股回报率

② 资产总计：首页->数据中心->单表查询->公司研究系列->财务报表->资产负债表->资产总计

③ 流通市值：首页->数据中心->单表查询->股票市场系列->个股交易数据->月个股回报率文件->月个股流通市值

问题 1：列出万科公司 2019 年 1 月的收益率和 2018 年 12 月的流通市值?

问题 2：查看 2019 年 12 月所有上市公司"资产总计"分布(列出均值、方差、最小值、 25、 50、 75 分位数、最大值、样本数、缺失值数等)。

观察上述问题 2 可以发现，部分数据存在着缺失值，请思考：如何处理缺失值？为什么？

问题 3：请将“资产总计”中的缺失值填充为横截面中值(或均值)，并列出处理缺失值之后的"资产总计"分布。

填充缺失值之后，就可以开始按照上述的公式计算因子值。其中，动量至少过去 5 个月有数据，否则不予计算，赋 NA 值。

请给出万科公司 2019 年 1 月动量、AM 的值。

问题 4：查看万科公司 1996 年 12 月-2019 年 11 月的 AM 分布，并作图列出该因子的时序图(x 轴为时间，y 轴为因子)。

因子数据推荐整理成宽表形式，也可以整理成长表形式便于保存，节省存储空间。

表 1　宽表格式示范

时间/股票	000001	000002	……	688399
199612				
199701				
199702				
……				
201911				

表 2　长表格式示范

时间	股票	因子(AM)
199612	000001	
199701	000001	
……		
201911	000001	
199612	000002	
……	……	
201911	688399	

回溯到过去(比如 2006 年 12 月末)，我们只能得到在 2006 年 12 月末可用的数据。某些财务报表数据由于财报的公布时滞而不可用。在我国，t 年年报总是在 t+1 年的 4 月公布； t+1 年一季报与 t 年年报公布时间类似； t+1 年半年报则在 t+1 年的 8 月公布； t+1 年三季报则在 t+1 年的 10 月公布。考虑到这一点，我们需要将计算得到的因子做填充处理：

1. 利用 t 年年报因子数据填充 t+1 年 4-7 月份数据（在 t 年年报数据和 t+1 年一季报数据中，年报数据质量更高，所以使用年报数据）；

2. 利用 t+1 年半年报因子数据填充 t+1 年 8-9 月份数据；

3. 利用 t+1 年三季报因子数据填充 t+1 年 10 月~t+2 年 3 月份数据。

2. 单因子选股(基于投资组合排序，可自行拓展)

单因子选股的基本思想是在 t 月根据因子值大小对所有样本股票进行排序，并分成 10 等份，构建 10 个投资组合；并计算 t+1 月的收益。主要分为如下的步骤（请联系现实投资步骤来思考）。

步骤 1：t 月排序。在每个月的截面上，按照因子大小对该月的样本股票进行排序。比如在 1996 年 12 月末按照所有股票的 AM 从小到大进行排序。

步骤 2：排序后构建投资组合。通常分成 10 等份，构建 10 个等权投资组合。

步骤 3：持有投资组合一个月，计算各组合 t+1 月的收益。比如，1996 年 12 月末构建的股票投资组合，计算持有投资组合至 1997 年 1 月末的(月度)投资组合收益。

步骤 4：同时构建多空头寸。每月做空因子值最小的投资组合(1 分位)，做多因子值最大的投资组合(10 分位)，计算多空收益。

步骤 5：获得时序收益。从 1997 年 1 月开始投资，月度调仓，直到 2019 年 12 月。可以获得每个组合的月度收益时序。

步骤 6：投资绩效评估。针对过去投资收益时序进行绩效评估(参考投资学中“绩效评估”内容)。

问题 5：列出 2018 年 12 月 AM 升序排序的第一组中前十只股票的 AM 值，并输出这十只股票在 2019 年 1 月的收益。

问题 6：列出 2018 年 12 月 AM 升序排序的 10 组投资组合和多空组合在 2019 年 1 月的月度收益，亦可用柱形图展示（参考投资学教科书）。

问题 7：绘制 1997 年 1 月至 2019 年 12 月的 10 个分位组合的投资收益时序。x 轴为月，y 轴为 10 个分位组合的收益。计算投资组合收益可以按照等权重（或市值加权）计算。

问题 8：列出整个投资期内，10 个分位数投资组合和多空组合的平均收益、经过 Newey-West 调整的 t 值、未经调整的 t 值、夏普比率、经过 CAPM 和 FF3 模型调整后的 alpha 及 NW 调整的 t 值。

问题 9：为什么用 Newey-West 调整的 t 值？请查阅资料。并将其与非 NW 调整的 t 值进行比较。

问题 10：观察动量因子和 AM 因子的分组收益和多空组合收益，是否随着因子的增大而收益增大？可以参考《投资学》绘制柱形图，或其他可视化方式。

此时你已经掌握了单因子选股的回溯测试方法。利用课程提供的 95 个因子数据，可比较不同因子所得的投资绩效。

问题 11：列出等权(和市值加权)情况下均能取得显著多空组合收益的因子，列出这些因子以及采用这些因子进行投资的绩效评估结果。

3. 基于机器学习的多因子量化投资系统

单因子模型采用单个因子来选择股票并检验其对于股票收益的预测能力。现实中，股票收益可能受到多个因子的影响。理论部分参考《投资学》的多因子模型部分。通常可以采用打分法和回归法等方式构建一个综合性的因子来筛选股票，课程作业要求同学们采用机器学习来进行多因子投资。基本的思路是**拟合多个因子与未来月收益之间的关系，预测未来收益；投资者根据预测收益进行排序和构建投资组合；最后进行绩效评估**。可以发现第二部分和单因子部分类似，只是因子变成了人工智能算法的预测收益或综合因子值。

我们采用问题 11 中的显著因子作为输入变量(大概 8 个，根据自己的结果进行调整，或采用全量因子数，全量因子的计算时间稍长。8 个参考因子为：01size、12std_turn、13volumned、14std_vol、15illq、16LM、17retnamx、80rd_mve)。我们将构建的模型为：

$$ret_{i,t} = f(x_{i,t-1})$$

其中$ret_{i,t}$为股票 i 第 t 月的超额收益=原始收益-无风险收益率，$x_{i,t-1}$为股票 i 在 $t-1$ 月的因子向量。分析模型特性可以发现，该问题是一个典型的回归任务。

机器学习模型包括超参数与模型参数，其中模型参数根据数据所学得，而超参数则由用户指定或通过历史数据交叉验证得到。目前不考虑超参数的选择，采用程序默认设置或手工指定。这里采用 RandomForest 算法（可自行选择）。默认参数：树的个数 n_estimator=100；最大深度 max_depth=6；学习率 learning_rate=0.05；n_jobs = -1（此处请开启多线程，否则速度较慢）。

假设我们位于 2018 年 12 月末，我们将运用智能投资模型基于 2018 年 12 月可用的因子预测 2019 年 1 月的超额收益。那么，站在 2018 年 12 月末，可用的数据为 1996 年 12 月-2018 年 12 月的因子数据和 1997 年 1 月-2018 年 12 月 (t) 的超额收益数据。如果时间到了 2019 年 1 月末，我们看到的数据多了 2019 年 1 月的因子和收益数据。请联系现实中投资的流程思考该过程，做出投资的流程图。

步骤 1：因子-超额收益数据集构建。构建如下的因子-超额收益数据

表 3 因子-超额收益数据格式示范

时间(t)	股票	因子 1(t-1)	因子 2(t-1)	.....	因子 n(t-1)	超额收益(t)
199701	000001					
199701	000002					
.....						
199701	688399					
199702	000001					
.....	.....					
201912	688399					

步骤 2： RandomForest 模型的训练。为了拟合得到 2018 年 12 月末可用的模型，准备 2018 年 1 月-12 月可用的因子-超额收益数据作为训练数据。采用该训练数据拟合 RandomForest 模型。

步骤 3：预测 2019 年 1 月的超额收益。准备 2018 年 12 月末的因子，运用训练所得的模型可得 2019 年 1 月所有股票的预期超额收益。

步骤 4：根据该预期收益进行排序，构建投资组合并持有一个月，计算各组合在 2019 年 1 月的持仓收益(等权和市值加权)。

步骤 5：窗口（12 个月）向前滚动到 2019 年 1 月末，重复步骤 2-4，训练并预测 2019 年 2 月的收益，构建投资组合并进行收益等。

步骤 6：该过程从 1997 年 1 月滚动至 2019 年 12 月(或 2018 年 12 月，如果采用 Github 的数据)，可以得到每月的投资组合时序，并进行时序上的绩效评估等。

问题 12：统计第一次训练样本数和测试样本数（因子时间区间为 199612-199711，收益率时间区间为 199701-199712）。

问题 13：列出 2019 年 1 月预期收益最高和最低的 10 只股票，并列出它们的预期收益。

问题 14：计算各组在 2019 年 1 月的持仓收益(等权和市值加权)。

问题 15：按照单因子中绩效评估的方法对运用机器学习方法的多因子模型的时序收益进行绩效评估。

问题 16：比较单因子模型的收益与多因子模型所得收益。可运用表格或绘图展示。

4. 附加题

4.0 说明

前 3 部分做完即可，做得好即可得满分。有余力可针对上述系统进行一定的扩展。我们给出几个可扩展的地方供同学们参考。

4.1 Fama-MacBeth 检验

除了构建投资多空投资组合以外，也可以运用 Fama-MacBeth 进行回归检验。可自行探索。

4.2 机器学习模型的超参数选择

模型对数据的适用性需要通过“训练-验证-测试”来实现。本部分是对第三部分的拓展，将第三部分中的单纯的“训练-测试”改变为“训练-验证-测试”。具体而言，可采用“滑动窗口”法：我们可以将“训练集-验证集”的长度固定，如“11 个月-1 个月”，因为原则上“训练集-验证集”的存在是为了获得最适合当下的模型，测试集才是真正的预测，我们将测试集设定为未来一个月。亦可采用“扩展窗口”法。

4.3 不同机器学习模型的比较

机器学习作为数据分析的一种有力工具，种类繁多。你可以尝试利用其他算法对未来收益做出预测，并将之与 RandomForest 的结果做对比分析。

4.4 更多绩效评估方法

参考《投资学》绩效评估，或参考文献。

参考文献

李斌，邵新月，李玥阳. 2019. 机器学习驱动的基本面量化投资研究. 中国工业经济, (8): 61-79.

Gu, Shihao, Bryan T. Kelly, and Dacheng Xiu. 2020.“Empirical Asset Pricing via Machine Learning.” Review of Financial Studies, 33 (5): 2223–73.