《因子分析与机器学习策略》简介

这门课面向的对象是专业的量化研究员、或者打算向这个方向转岗求职、或者尽管是其它职业，但决心以专业、严谨的态度探索量化研究的学习者。

学完这门课程并完全掌握其内容，你将具有熟练的因子分析能力、掌握领先的机器学习策略构建方法，成为有创新研究能力和比较竞争优势的量化研究员。

1. 课程目标

在学完本课程之后，你将会获得以下能力（或工具）：

1. 掌握 Alphalens 因子分析框架，并在工作中运用。
2. 懂得如何阅读 Alphalens 的分析报表，根据报表判定因子有效性。
3. 懂得如何运用 Alphalens 挖掘因子的价值。
4. 介绍六类400+因子，带走Alpha101 等 350+独立因子（同一算法、不同参数和周期只算一种因子）。
5. 掌握因子挖掘方法论，有能力挖掘新因子、改进旧因子。
6. 带走Pair Trading中性策略，通过机器学习模型寻找配对资产。
7. 精通 XGBoost 模型，带走基于XGBoost的资价格预测模型、和趋势交易模型。这将在一段时间内，成为你工作中的法宝。

2. 先修要求

在学习本课程之前，学员需要掌握的 Python 编程基础，包括：

1. Python 基础语法和常用库，包括时间日期、字符串、文件 IO、列表、字典、模块导入、typing 等等。
2. 统计学知识。需要有大学基础的统计学知识，对一些基础概念有初步了解，这部分内容在《量化 24 课中有详细讲解》。
3. Jupyter Notebook。我们提供了《Notebook 入门》和《Notebook 高级技巧》供大家学习。
4. Numpy 和 Pandas。需要有入门级的知识。课程中使用了大量 Numpy 和 Pandas 高级技巧，如果没有事先掌握，会增加阅读示例代码（包括听课）的难度。建议在上本课时，同时学习我们的《量化交易中的 Numpy 和 Pandas》课程（免费）。我们也会在讲课中讲解一些语法知识，但不是课程重点。
5. 对机器学习、神经网络有一定的认识。在我们讲解机器学习核心理论时，这会帮助你跟上进度。

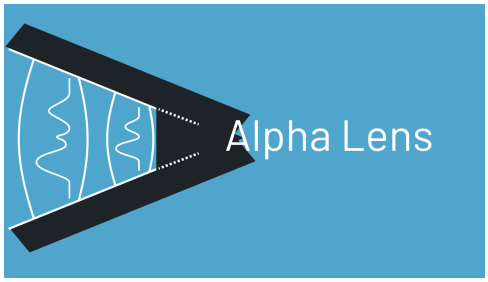
如果不满足前两个条件，学习本课程可能会有一定困难。如果不满足后面三个条件，你仍然可以学习本课程，但需要多花一些时间来熟悉这些内容。

3. 课程内容

课程内容涵盖了因子挖掘、因子检验和构建机器学习模型三大模块。

3.1. 因子检验方法

只有掌握了因子检验的方法，我们才能判断挖掘出的因子是否有效。因此，因子检验方法是本课程的起点，从第 2 章开始，到第 7 章结束，共 6 个章节。



Alphalens Logo

我们将从介绍因子检验的原理入手，手动实现因子检验的各个流程；再介绍开源因子分析框架 Alphalens。我们不仅会介绍如何把 Alphalens 用起来，还会重点介绍如何解读它的报表、如何进行参数调优和排错。这部分包含了大量业界经验、正反例对比，内容之深、之新，是你目前在网络上无法看到的。

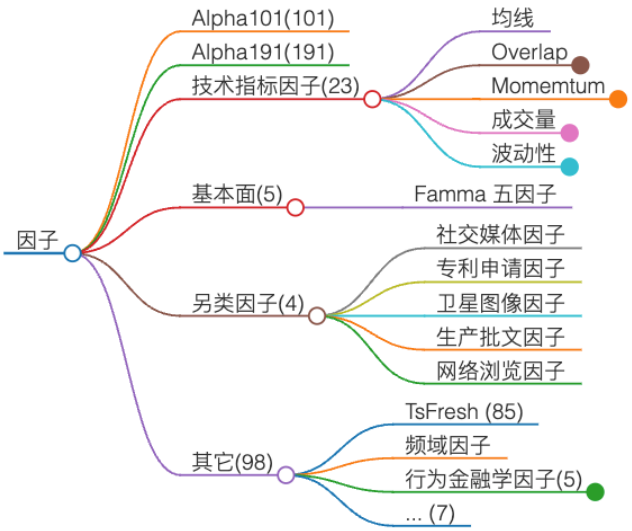
当你懂得如何通过报表来判断因子的好坏、如何灵活运用 Alphalens 以揭露隐藏在大量数据之下的因子与收益的关系的时候，你就真正成长为因子分析高手。

3.2. 因子挖掘

第 8 章到第 12 章构成了课程的第二部分。

因子挖掘，或者说特征工程，是构建交易策略的重要一环，也是量化研究员最日常的工作项。我们将介绍 Alpha 101 因子、Ta-lib 和技术指标因子、行为金融学因子、基本面因子、另类因子。

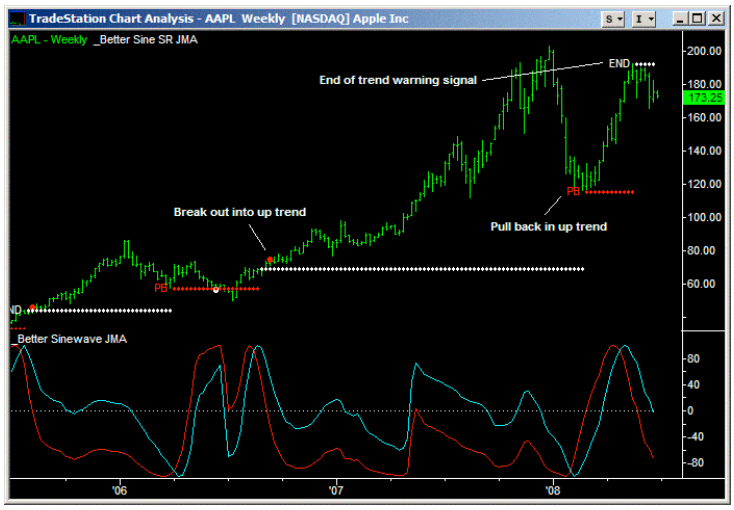
如果掌握这些因子还嫌不够，我们还将在第 12 章介绍因子挖掘方法论。你可能是从我们发表在网上的各种因子与策略挖掘的精彩文章吸引，从而找到我们的，在这里，我们将把掌握的资源和方法论全部教授给你。



各类因子

在介绍 Alpha 101 因子时，我们把重点放在如何理解它的数据格式和实现算子上。这是理解 Alpha 101 的基础，掌握了这些算子，你就完全能够读懂和实现全部 Alpha 101 因子。然后，我们会介绍其中的几个因子。我们会告诉你如何阅读它复杂的表达式，如何理解作者的思路和意图。

在实现 Alpha 101 因子上，由于已经有许多开源的实现存在，因此，我们不打算重新发明轮子，而是向你介绍一个我们认为实现最完整、正确率最高的一个开源库，并在我们的附录中可以直接上手使用它。此后，你可以把它纳入你的量化兵器库。

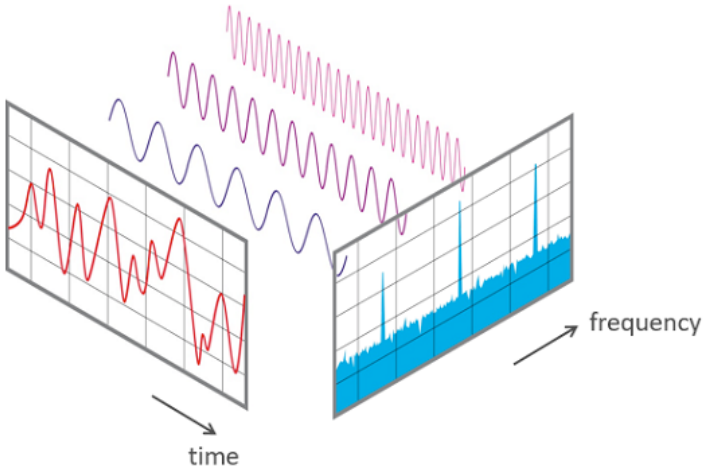


Hilbert Sine Wave

在第 9 章，我们将介绍 Ta-lib 库以及它实现的一些技术指标。我们将介绍均线、Overlap、Momemtum、成交量和波动性等 20 个左右的指标。有一些你可能已经比较熟悉了，比如均线，也有一些你可能不太熟悉，比如基于希尔伯特变换的趋势线和 Sine Wave Indicator（如 [](#Hilbert Sine Wave) 所示）。和其它章节一样，我们仍然会保持足够的研究深度，会介绍像冷启动期、如何将老的技术指标翻新应用（以 RSI 为例）等等。

在第 10 章，我们将介绍基本面因子和另类因子。由于数据获取的难度和质量问题，我们将以介绍原理为主，不一定都给出代码实现。

在第 11 章，我们将介绍不属于任何归类，但仍然很重要的因子，比如小概率事件（黑天鹅）因子；我们会引入导数概念，介绍两个有效的一阶导、二阶导动量因子；时频变换构造频域因子；我们还将介绍一些行为金融学因子，这是当前金融学的热门概念，在短线交易中非常有用。



通过 FFT 提取频域因子

3.3. 构建基于机器学习的交易策略

这一部分我们先快速介绍机器学习的核心概念（第 14 章）。我们会介绍损失函数、目标函数、度量函数、距离函数、偏差、方差、过拟合与正则化惩罚等核心概念。这些概念是机器学习核心概念中偏应用层面一些的概念，是我们不得不与之打交道的概念。

> **Tip**
>
> 如果你需要深入理解机器学习和神经网络、自己能发明新的网络模型和机器学习算法，那么你还需要补充线性代数、梯度优化、反向传播和激活函数等知识。不过，对掌握我们这门课程，达到熟练运用已知的算法模型并会调参调优，掌握我们介绍的概念就足够了。

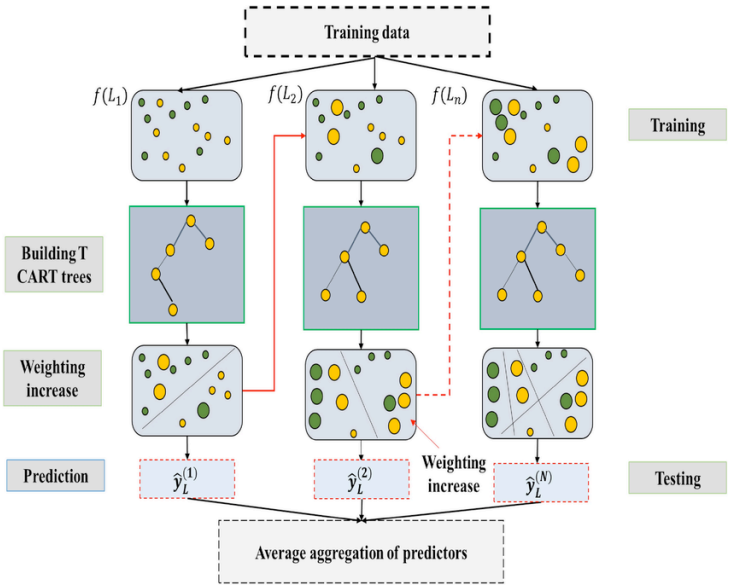
本课程选择的机器学习库是 sklearn。sklearn 是一个非常强大的机器学习库，以丰富的模型和简单易用的接口赢得大家的喜爱。在第 15 章，我们先向大家介绍 sklearn 的通用工具包 -- 用来处理无论我们采用什么样的算法模型，都要遇到的那些共同问题，比如数据预处理、模型评估、模型解释与可视化和内置数据集。

第 16 章我们会介绍模型优化方法，这是多数从事机器学习与人工智能的人所能掌握的核心技能，也是我们做出一个优秀的机器学习交易模型的关键之一。我们将演示如何使交叉验证、如何使用网格搜索 (GridSearch)、随机搜索 (RandomizedSearch) 等方法。

量化领域的机器学习有它自己的特殊性，比如在交叉验证方面，我们实际上要使用的是一种称为 Rolling Forecasting（也称为 Walk-Forward Optimization 的方法）。我们将在第 16 章的最后部分，详细介绍这种方法以及它的实现。

接下来我们介绍一个聚类算法（第 17 章）。在量化交易中，Pair Trading 是一类重要的套利策略，它的先决条件是找出能够配对的两个标的。这一章我们将介绍先进的 HDBSCAN 聚类方法，演示如何通过它来实现聚类，然后通过 statsmodels 中的相关方法来执行协整对检验，找到能够配对的标的。最后，我们还将演示如何将这一切组成一个完整的交易策略。

在第 18 章，我们将介绍 XGBoost，这是一种梯度提升决策树模型。由于金融数据高噪声的特性、以及难以获得大量有效标注数据原原因，使得梯度提升决策树模型目前仍然是在量化交易领域应用最广泛、效果最好的机器学习模型。



我们会先从决策树模型讲起，介绍 XGBoost 一路走来的优化历程。然后以一个详尽的例子，介绍如何使用 XGBoost，训练一个模型并深入到它的内部：我们将可视化这个模型的重要特征、绘制出它的模型树。最后我们以介绍在 XGBoost 如何进行交叉验证和调优结束。

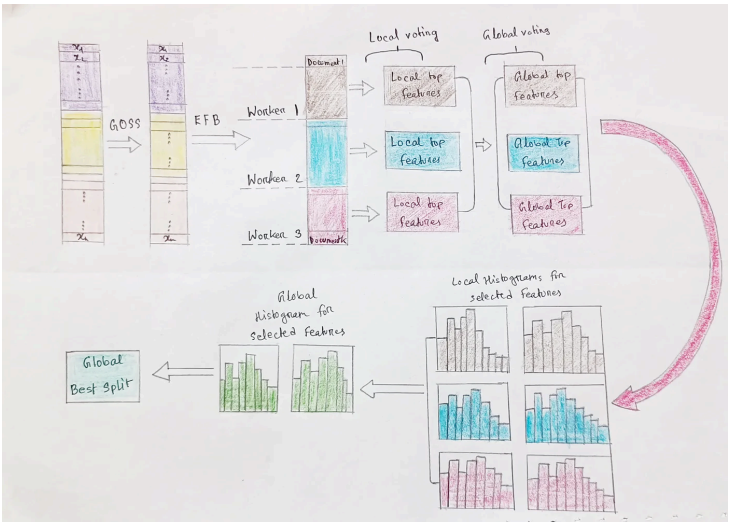
在做了大量理论与实操的学习之后，我们已经完成了所有铺垫，现在是时候学习如何构建基于 XGBoost 的量化交易策略了。我们将抛弃几乎是无效的端到端训练的方式（即输入价格，希望能预测出下一个价格），改用一些探索性、但更加有效的策略示例。

我们将在第 19 章里，介绍基于 XGBoost 回归模型，如何构建一个价格预测模型。我们会介绍策略原理、实现和优化方案。尽管我们构建的是一个价格预测模型，但它决非你在网上看到的那种端到端的 toy 类型的模型！

另一个模型将在第 20 章里介绍，它是基于 XGBoost 的分类模型构建的一个交易模型。换句话说，它不负责预测价格，但能告诉你应该买入、还是卖出。在本章中，我们还要介绍如何制作标注工具

这两个示例指出了在目前条件下，应该如何使用机器学习构建策略的根本方法 -- 由于金融数据饱含噪声，所以我们不能指望端到端的模型能够工作。但如果我们能清晰地定义问题，找出有效特征，机器学习就会非常强大！这将是你在市场上战胜他人的利器。

第 21 章我们会对 XGBoost 模型进行更深入一些的拷问。我们介绍另一个梯度提升决策树模型的实现，即由微软开发的 LightGBM。一般认为，它在性能上要强于 XGBoost，内存占用更小，在其它指标上与 XGBoost 相比各有千秋。



LightGBM, by Hossain@medim

我们已经介绍了三个非常实用的例子，涵盖了套利交易、价格预测和交易模型。但是，资产管理中还有一个重要的课题，就是组合管理。基于机器学习，我们如何实现组合管理？我们也将在这一章回答这个问题。

我们的课程将结束于第 22 课。我们将介绍深度学习的先驱--CNN 网络在 K 线模型识别上的应用。我不认为 CNN 网络在 K 线模型识别上有任何优势，我会详细地介绍为什么。但是，也许你有兴趣解决这个问题，所以，我还是会介绍 CNN 在 k 线识别上的一般做法。

比起深度学习，我更看好强化学习在交易方面的应用。在加密货币、商品期货上，重要的不是选择投资品种，而是根据市场的变化决定交易时机、仓位、杠杆，这天然就是一个强化学习问题。我将介绍强化学习的基本概念、相关学习资源。

最后，还有两个无法归入到上面所有这些类别 -- 无论是机器学习、深度学习还是强化学习，但仍然非常重要的智能算法 -- Kalman Filter 和 Genentic Algorithm。

整个课程的大纲可以在 这里 查阅。

4. 课程编排

课程内容由正文、习题和补充材料三部分组成。

课程正文内容以应用为主，对机器学习的核心理论，只讲到在应用中必须接触的部分。在几乎每一章都提供了大量拓展阅读材料和注释，供希望在具体细节或者底层体系上深入研究的同学。对这部分内容，没有时间的同学可以跳过，不影响课程学习效果。

课程附有大量习题。习题的目的是：

1. 部分示例中涉及一些编程技巧，在量化研究中比较常用，所以编入习题强化记忆。
2. 部分话题的结论是开放性的、探索性的，不适合作为正式内容讲授。
   我们为本课程精心准备了大量的习题。你应该充分利用这些习题来巩固和拓展自己学到的知识与技能。这些习题多数是自动批改的，因此你可以及时了解到自己知识掌握的程度。

习题分发、提交作业和获取老师批阅结果流程都是自动化的，通过 Nbgrader 来实现。如果你之前没有接触过 Nbgrader，可以在 课程须知 的关于作业一节中掌握它的使用方法。

视频、教材和习题内容相互补充，相当于报一门课，得三门课！

本课程只涵盖了量化交易中的部分知识。如果要独立从事交易或者做完量化全栈工作，建议补充学习《量化 24 课》。本课程与《量化 24 课》的区别是，本课程内容更为专精，《量化 24 课》内容更广泛，涵盖更全面。