零基础算法交易入门

施贺卿

2020
十月

Contents

个人网页: http://shiheqing.github.io

1 第一节：算法交易简介

1.1 算法交易常用算法、概念和计算机实现

量化交易从业者和学术界的一大研究课题就是在金融市场中寻找 α（超额收益）。不过 α 的寻找并非一个容易的过程，我们需要在成千上万的信息当中筛选出统计学意义上显著能获得 α 的信息，而金融数据当中的信噪比往往是十分低的。且找到了能带来 α 的信息后也不是一劳永逸的，有效信号的收益表现通常都是一个关于时间的递减函数。如果我们把金融市场当中的可观测数据看作一个 $\sigma-$ 代数流，即我们在当下时刻 t 能够掌握 $[0,t]$ 区间里的所有信息，那么从数学的角度，任何在 $t+1$ 时刻我们所做出的决策，都可以表示为基于 $\sigma-$ 代数流上的变量的函数，即 $S(A_1, A_2, ...A_i|\mathcal{F}_t)$，$i \in \mathbb{N}^+$, $0 \leq t$, A_i 表示我们用于制定决策函数 $S(\cdot)$ 的第 i 个金融随机变量。通常一个简单的决策函数 $S(\cdot)$ 会包含一些简单的时间序列，例如 A_1 表示某个股票的时间序列，A_2 表示无风险利率的时间序列。

算法交易作为量化金融的一个偏应用的方向，主要是靠数学、统计学的一些理论工具来对可能用到的时间序列 A_i 进行建模，并且利用计算机编程语言将其实现，进而通过观测数据找到我们的策略 $S(\cdot)$ 来获得 α。课程中我们只谈一些 P-Quant 领域的一些算法，暂时不去管风险中性测度下的定价问题。任何算法的数学理论对于 Quant 来说都是应用的基础，我们接下来将一起探讨贝叶斯方法、时间序列分析和机器学习等数学统计工具在 P-Quant 领域内的应用。

课程中提到的算法主要用 Python3.7 来实现，有些算法在 R 中有直接可用的 package 我们也会用 R 来做阐释。作为一门零基础的算法交易入门课程，我们在解释推到数学模型之余，更重要的是用数据来实践我们所学的模型。用到的数据基本上都是获取自 Yahoo Finance。课程涉及的数据和代码都可以在我的这个Github 项目中找到。

希望这门小课的内容可以让金数专业的同学在已有的数学统计基础上了解一下数学在量化交易实践中的应用，也帮助想在量化资产管理、量化对冲基金等量化相关行业寻求职位的同学大致地介绍一下 P-Quant 分支的基础知识框架，以备面试之用。当然，最重要的是希望大家可以体会到数学、统计学等量化工具在解释和预测金融时间序列中简洁且优美的力量。

1.1.1 贝叶斯方法概述

金融市场的本质是一个反映投资者对金融随机变量的观点的市场。因为有了不同投资者对某个金融资产（随机变量）的不同态度，有的看好有的不看好，这些不同的态度会体现在投资者的决策函数 $S(\cdot)$ 上。这样的特质让贝叶斯学派的统计方法在金融系统的量化中有了很大的用武之地，因为贝叶斯方法的核心也是通过在概率模型中不断根据新的观测结果来更新模型对某一事件发生概率的估计，本质上也是一种对主观态度的建模。

贝叶斯方法不同于传统的频率学派（Frequentist），它对一个随机事件发生概率的认知是随着客观世界的数据改变而改变的，而频率学派对概率的认知是完全通过大量重复随机事件而来的。贝叶斯方法首先需要一个先验态度（prior belief），这一态度是贝叶斯类算法在没有接收任何数据之前的对要建模的随机事件的发生概率的初始认知，我们将这个先验概率记作 $\mathbb{P}(\theta)$。有了先验概率之后，贝叶斯方法根据观测到的数据 D，从而更新我们的先验概率 $\mathbb{P}(\theta)$ 以得到一个后验条件概率 $\mathbb{P}(\theta|D)$。

从 $\mathbb{P}(\theta)$ 到 $\mathbb{P}(\theta|D)$ 的更新算法是通过贝叶斯法则实现，

$$\mathbb{P}(\theta|D) = \frac{\mathbb{P}(D|\theta)\,\mathbb{P}(\theta)}{\mathbb{P}(D)}。 \tag{1}$$

相较于频率学派，贝叶斯方法明显的优点在于它可以根据随机变量的观测值来不断更新后验概率，观测数据越多后验概率就越接近真实值。而频率学派为了估计一个随机事件的概率，需要进行大量的重复试验，试验次数越多则概率估计越准确。但是，很多现实问题并无法进行完全独立同分布的重复试验，或者重复试验的成本很高，特别是在分析金融随机变量时。因此，贝叶斯方法是许多量化模型的基础。

1.1.2 时间序列分析概述

时间序列分析在量化交易算法中的主要作用就是将序列相关性量化。许多交易策略都需要用到金融时序的序列相关性，因为序列相关性可以定量地描述 t 时刻的资产价格在多大程度上是与 $t-1$ 时刻的资产价格相关的。利用时间序列分析这一统计学工具，我们可以系统性地揭示资产价格在不同时间点上的相关性结构，并且通过相关性结构来解释资产价格的运动轨迹和预测未来资产价格。

时间序列分析的主要目的其实就是尝试从理解过去当中预测未来。时间序列也是一种特殊的带时间标记的随机变量，随机变量在一个时间点 t 上有一个取值，而时间点往往是离散的（一日、一周、一年、...），因此时间序列也称为离散随机过程。在算法交易实践中，我们往往会给一个已经观测到的在 $t \in [0,t]$ 上的金融时间序列拟合一个时序模型，以此来挖掘时间序列的内在函数关系，并预测未来的时序走势。

时间序列通常有以下几种特性：

- 趋势（Trend）趋势可以分为确定性和随机性趋势。确定性趋势可以用确定的函数表达来解释，而随机性趋势是一个随机游走，不能用函数表达和解释。金融时序当中通常都会含有确定性趋势，尤其是大宗商品（Commodity）类的价格时序，这为使用时序分析来开发可盈利的算法交易策略提供了前提条件。例如，国内基金常用的 CTA（Commodity Trading Advisor）策略，就是主要利用时序分析技术来识别确定性趋势；
- 季节性（Seasonal Variation） 许多金融时间序列的轨迹也是有季节性规律的，如天然气相关的衍生品价格；
- 序列相关性（Serial Dependence） 金融时间序列最重要的特性之一。金融时序通常都会出现高度自相关的阶段，且这些自相关的时间点大多都是聚集的。例如，资产收益波动率就有明显的聚集特征，且有较高的自相关性。根据这一特点，利用时间序列分析我们可以开发波动率交易策略。

1.1.3 机器学习方法概述

机器学习是另一个算法交易常常需要用到的统计学习工具。简单来说，机器学习就是将一些统计模型拟合到观测数据集上。拟合地好坏程度往往决定了机器学习所学到的模型的表现，我们用损失函数（Loss function）来评估拟合的好坏。常用的损失函数有 Mean Squared Error（MSE）、Log-loss 等。粗略地讲，机器学习算法又可以分类为有监督学习（Supervised Learning）、无监督学习（Unsupervised Learning）和强化学习（Reinforcement Learning）。我们的课程内容将主要涉及有监督和无监督学习，因为这两类的模型发展的历史更久且也有比较好的可解释行，适合用在量化金融这

个对模型可解释性较高的领域。

在量化金融行业中，机器学习算法常用来预测未来资产价格，优化交易策略所需要的参数，风险管理，以及从含有噪音的市场数据中提取交易信号。作为第一节课，我们先大致过一遍有监督和无监督学习的一些常用算法。

Supervised Learning

分类问题（Classification）：分类问题的目标是预测一个类别变量（Categorical Variable）的取值。类别是一个有限元素的集合 K，且类别的取值是数字而不是类别本身的名字。数学地描述分类问题就是我们要估计一个概率 $p(y = k|\underline{x})$。我们要分类的一个数据 y 通过如下模型来分类，

$$\hat{y} = \hat{f}(\underline{x}) = argmax_{k\in K}\, p\,(y = k|\underline{x})\,。 \tag{2}$$

逻辑回归、朴素贝叶斯、支持向量机和深度卷机神经网络等都是常用的分类算法；

回归问题（Regression）：回归问题与分类问题最大的不同在于回归问题尝试预测的变量 y，有 $y \in \mathbb{R}$，而不是像分类问题那样 $y \in \{K\}$。但是本质上还是构建一个概率分布模型

$$\hat{y} = \hat{f}(\underline{x}) = argmax_{z\in\mathbb{R}}\, p\,(y = k|\underline{x})\,, \tag{3}$$

被预测变量 y 的预测值 $\hat{y}$ 也是通过最大化（3）中的概率 p 实现。线性回归、支持向量回归和随机森林等都是常用的回归算法。

Unsupervised Learning

降维问题（Dimensionality Reduction）：在有监督学习中，模型的构建需要输入特征变量（Features）来预测目标变量（Target Variable）。而无监督学习中只有特征变量及其观测值，并没有标签集。但是无监督学习可以根据已观测到的特征变量的数值，来对特征的维度进行缩减。因为有些特征对于数据的描述作用并不大，利用一些降维算法提取出对数据描述作用最大的变量，可以为进一步搭建模型降低计算成本，也降低了噪音的含量。开发交易算法时，我们手中有的金融变量往往是很多的。拿一个简单的股票策略举个例子，假如我们的目的是构建一个含有不同行业因子的股票组合，但是我们手里的数据往往是没有经过处理的，有大量相同或者相似行业的股票在数据集中。为了更简洁更低成本地构建一个多因子股票策略，降维算法就可以拿来将相关性高的重复股票删除，留下我们需要的因子间相关性不高的多因子策略。常见的降维算法有 Principal Component Analysis (PCA)。

聚类问题（Clustering）：解决聚类问题是无监督学习的另一个应用场景。因为我们的数据集中并没有包含标签，所以算法只能从特征矩阵入手，将数据划分为不同的种类。在算法交易当中，聚类算法会拿来聚类特性相似的金融资产。另外，市场条件也可以用聚类算法来分析，我们的特征矩阵内可以包含各种各样的金融时序变量，聚类算法会将相似趋势的输入数据聚类在一起，而这些聚类在一起的相似数据往往代表着一段时间内的市场条件。

1.2 QSTrader **回测环境的搭建与基本使用**

QSTrader 回测环境是基于 Python3.7 搭建的，我们在自己的 Python 环境中 pip install 一下 QSTrader 包即可使用。