Предсказание курса акций на основе методов машинного обучения.

Kovrizhnukh Dmitrii Daniil Merkulov
kovrizhnykh.diu@phystech.edu daniil.merkulov@sckoltech.ru

Project Proposal

Задачей проекта является нахождение значений акций при помощи алгоритмов машинного обучения и вычислительной математики, а также составление оптимальных портфелей на их основе с целью получения максимальной прибыли на заданном временном интервале.

1 Idea

Человечество всегда хотело предсказывать будущее, понимать, что будет впереди. Как говорил мой учитель физики: "Если бы я знал курс доллара на завтра, я был бы миллионером". Действительно, зная стоимость акций на завтра, можно заранее продать или купить их, чтобы заработать. Более того, в мире существуют алгоритмы предсказания значений функций по её куску и некоторым параметрам, которые различаются по своей вычислительной сложности и эффективности. Идея этого проекта состоит в том, чтобы взять самые известные из них и сравнить на разных типах ценных бумаг.

Для каждой кампании дана статистика в виде csv файла, содержащая стоимость в зависимости от времени, а также мультипликаторы : P/EP/S, EV/EBITDA, ROE, ROA, Current ratio и D/E. Для чистоты эксперимента возьмём бумаги трёх разных типов: Фонды, волатильные акции и валюту. Возьмём три алгоритма предсказания: нейронную сеть на основе библиотеки tensorflow, алгоритм случайного леса, и алгоритм адаптивной фильтрации. Задача состоит в том, чтобы применить каждый алгоритм к каждому типу бумаг, получить предсказания на следующие сутки и сравнить их с реальным результатом за этот день.

1.1 Problem

Теперь представим математические формулировки наших задач.

Алгоритм Марковица для данной доходности:

Задача оптимизации портфеля активов с вектором средней доходности r ковариационной матрицей V может быть сформулирована следующим образом

$$\begin{cases} \sigma_p^2 = d^T V d \to \min \\ d^T r = r_p \\ d^T e = 1 \end{cases}$$

К этим условиям в задаче оптимизации портфеля активов следует добавить условие положительности портфеля (долей). Однако, в общем случае финансовых инструментов предполагается возможность открытия коротких позиций (отрицательных долей инструментов в портфеле). Тогда можно найти общее аналитическое решение задачи. Если обозначить,

$$A = \begin{pmatrix} r^T \\ e^T \end{pmatrix} V^{-1}(r, e) = \begin{pmatrix} r^T V^{-1} r & r^T V^{-1} e \\ e^T V^{-1} r & e^T V^{-1} e \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

то решение задачи имеет вид

$$d^* = V^{-1}(r, e) A^{-1} \begin{pmatrix} r_p \\ 1 \end{pmatrix}$$

Тогда зависимость дисперсии оптимизированного (эффективного) портфеля от требуемой доходности будет иметь вид

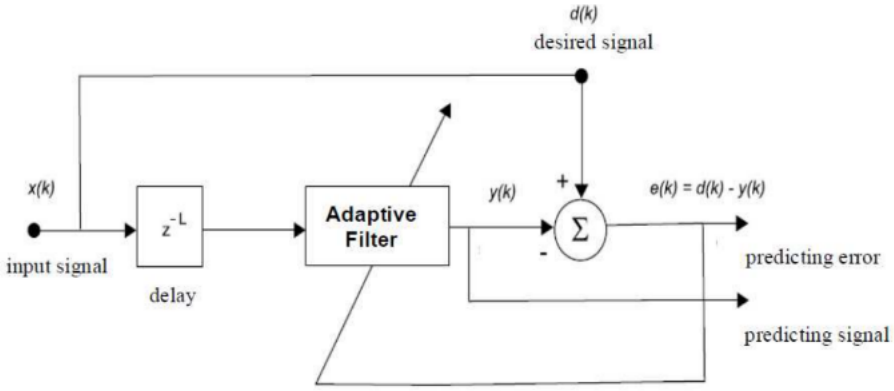
$$\sigma_p^2 = (r_p, 1) A^{-1} \begin{pmatrix} r_p \\ 1 \end{pmatrix} = \frac{a_{22} r_p^2 - 2 a_{12} r_p + a_{11}}{a_{11} a_{22} - a_{12}^2} = \sigma_0^2 \frac{(r_p - r_0)^2}{r_0 (r_1 - r_0)} + \sigma_0^2$$

где $\sigma_0^2 = 1/a_{22}, r_0 = a_{12}/a_{22}$ — минимально возможная дисперсия доходности портфеля и соответствующая ему средняя доходность

$r_1 = a_{11}/a_{12}$ — доходность портфеля, с соотношением риск-доходность таким же как и портфель минимального риска (графически это единственная точка пересечения с параболой прямой, проходящей через начало координат и вершину параболы)

Алгоритм адаптивной фильтрации:

Рекурсивный адаптивный алгоритм наименьших квадратов (RLS, recursive Least-Squares) позволяет добиваться высокой производительности и скорости сходимости при использовании в переменных во времени средах.



Общая блок-диаграмма адаптивного фильтра для предсказания сигналов

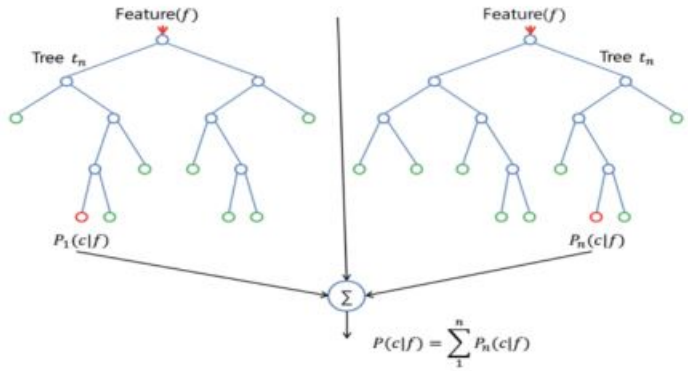
Еще один важный параметр — это так называемое окно наблюдений, то есть временной период анализа, который обычно имеет прямоугольное или экспоненциальное соответствие. Параметр, контролирующий соответствие и длительность окна наблюдений называется коэффициентов забывания (он применяется к памяти алгоритма), 0 « < 1 и объектная функция выглядит так:

$$F(e(k)) = \xi^{d}(k) = \sum_{i=0}^{k} \lambda^{k-i} e^{2}(k) = \sum_{i=0}^{k} \lambda^{k-i} \left[d(i) - x^{T}(i) w(k) \right]^{2}$$

Эта функция выпукла в многомерном пространстве w(k), то есть $\xi^{d}(k)$ имеет лишь один глобальный минимум и ни одного локального минимума. Поиск этого минимума и есть наша задача. Таким образом, мы можем достичь этой точки, установив градиент $\xi^{d}(k)$ равным нулю, что приведет к такой формуле:

$$w(k) = \left[\sum_{i=0}^{k} \lambda^{k-1} x(i) x^{T}(i) \right]^{-1} \sum_{i=0}^{k} \lambda^{k-1} x(i) d(i).$$

Последним по списку, но не по значению будет алгоритм случайного леса:
Случайный лес (Random forest) представим в виде ансамбля деревьев решений, которые обучаются независимо. Каждый из решающих деревьев, обычно, делают простым. Для принятия окончательного решения выбирается значение, за которое проголосовало большинство деревьев. Каждое дерево в отдельности даёт низкое качество, но за счет их большого числа результат улучшается. На рисунке ниже продемонстрирован пример случайного леса из двух деревьев.



В построении деревьев на данных существуют свои особенности. Опишем традиционный подход в построении дерева. Пусть N - количество примеров на обучении. Тогда: 1. выбирается подвыборка обучающей выборки размера N (возможно, с возвращением) 2. строится дерево. При поиске разбиения при построении каждого узла рассматриваются не все признаки, а только часть из них (sklearn реализация - корень из количества признаков) 3. дерево строится до полного исчерпания подвыборки, либо на основе каких-либо других критериев Каждое из деревьев стараются сделать как можно проще. При принятии решения выбирается класс, за который проголосовало большинство деревьев решений.

2 Outcomes

В результате будут получены:

1)Интерактивный python notebook, который будет давать рекомендации по ребалансировке существующего портфеля в соответствии с выбранной стратегией, а также демонстрировать результат работы всех трёх алгоритмов на портфеле с одинаковой начальной суммой в течение некоторого времени.

2) Презентация, в ходе которой будут сравнены, объяснены полученные результаты и сделаны выводы о целесообразности использования данных алгоритмов

3)* По возможности будет сделан интерактивный сайт, который будет хоститься на гит-хабе и иметь функционал notebook.

3 Литературный обзор

3.1 Предисловие

На данную тему написано не так много исследований, как ожидалось, во многом потому, что алгоритмы машинного обучения весьма новы, а финансы требуют достаточно сложной математики, поэтому обе области в объединении запрашивают серьёзное владение материалом. К счастью, прогресс не стоит на месте и некоторые исследования можно найти в открытом доступе.

3.2 Применённые материалы

Базой является [1], именно на этой теории основывается данное исследование. На ней лежит вся та теоретическая основа, на которой будет строиться остальные математика и вычисления. Именно теория к случайным процессам объясняет теоретические принципы работы всех алгоритмов машинного обучения и является необходимой базой для понимания работы,

Основным будет являться алгоритм Марковица, как наиболее известный и с которым в дальнейшем будет проводиться сравнение. Его математическое обоснование содержится [2] и эта статья является больше теоретических минимумом для ряда фактов. В другой работе [3] содержится решение задачи для портфельной теории Марковица с некоторыми улучшениями для понижения риска. В самой статье описываются: теория, методология, постановка задачи и практические результаты. Её ценность состоит в том, что она даёт именно общее понимание данного алгоритма и его применения.

Несмотря на то, что в работе [4] рассматриваются алгоритмы адаптивной фильтрации для интегральных микросхем, там содержатся схемы и алгоритмы, объясняющие их работу и математическая база, упрощающая понимание других статей на данную тему особенно с точки зрения интерпретации алгоритмов.Проще говоря, там понятным языком написано, что и как работает. В работе [5] исследователи подстроили фильтр под работу с акциями и а также написали полноценный адаптивный фильтр, посчитали функции корреляции и выяснили наиболее оптимальное время предсказания.

Последним по списку, но не по значению будем алгоритм случайного леса: в работе [6] даётся небольшое описание алгоритма и основной упор идёт именно на написание кода и объяснение его составляющих, а также конкретных шагов работы с данными и анализ результатов.

За примерами кода: https://habr.com/ru/post/516236/ и https://colab.research.google.com/github/MerkulovDaniil/mipt21/blob/main/notebooks/Portfolio_optimization.ipynb - для алгоритма Марковица, https://russianblogs.com/article/3220274772/, [7] - для случайного леса; [8] - для алгоритма адаптивной фильтрации.

Отдельно хочется отметить курс Роланда В.И. по вычислительным финансам [9] и [10], а также [11]. С одной стороны это методы вычислительной математики в предсказаниях, с другой стороны там разъясняются современные методы в экономике, один из которых я хочу применить в моём исследовании. Также оттуда будут взяты примера кода, приведённых выше алгоритмов и некоторые другие фишки, которые упростят написание статьи.

В качестве данных выступят csv файлы со статистиками котировок - графиков цены акций от даты торгов. Их можно найти на гит-хабах соответствующих курсов или специализированных сайтах. Приложим некоторые их них. https://www.finam.ru/profile/moex-akcii/gazprom/export/ https://www.moex.com/ru/index/MRSV/archive/?from=2022-03-15&till=2022-04-14&sort=TRADEDATE&order=desc
Однако для сбора массовой статистики на несколько сотен тикеров воспользуемся данной функцией https://habr.com/ru/post/332700/

Наконец, скажем про курс оптимизации, который ведут Меркулов Д. и Гасников А. [12], к теории которого планируется прибегать постоянно. Без него просто невозможно было бы написать данную статью.

4 Метрики качества

Результатом работы программы будет суммарная стоимость акций в портфеле. За базовое значение возьмём алгоритм Марковица, как 100 процентов и относительно него будем считать эффективность остальных стратегий. Также необходимо будет сравнить, на каком сроке, какой алгоритм показал лучший результат.

Второй метрикой будет количество реализованных алгоритмов, чем больше реализую, тем лучше. Планируется написать 3 алгоритма, однако если успею, добавлю четвёртый по выбору.

Финальной метрикой будет понятность презентации для аудитории, а также ответы на вопросы по результатам, для чего всё это и затевается. Считается, что чем меньше "простых" вопросов - тем понятнее презентация и чем больше "сложных" - тем она интереснее.

5 Примерный план

- 17 апреля - Сбор csv статистик по акциям
- 21 апреля - Написание алгоритма для нейронной сети
- 24 апреля - Написание алгоритма для случайного леса
- 28 апреля - Написание алгоритма для адаптивной фильтрации
- 3 мая - Сбор графиков и метрик
- 5 мая - Создание презентации и стэнда
- 10 мая - Актуальная дата начала работы
- 11-12-13 мая -https://www.youtube.com/watch?v=u--c7PXp8qA

References

[1] С. А. Гуз Е. О. Черноусова М. Г. Широбоков Е. В. Шульгин А. В. Гасников, Э. А. Горбунов. https://arxiv.org/pdf/1907.01060.pdf. *MIPT*, 21(5):31–213, 2005.

[2] В.Ю. Попов В.А. Бабайцев, А.В. Браилов. БЫСТРЫЙ АЛГОРИТМ МЕТОДА КРИТИЧЕСКИХ ЛИНИЙ МАРКОВИЦА. *Финансовый университет при правительстве РФ*, 21(20), 2011.

[3] Юрий Дорн. Сравнительный анализ долгосрочных стратегий формирования портфеля ценных бумаг. *https://www.hse.ru/data/2014/06/10/1324138096/%D0%94%D0%B8%D0%BF%D0%BB%D0%BE%D0%BC%20%D0%94%D0%BE%D1%80%D0%BD.pdf*, 21(19):03–23, 2013.

[4] В.И. Джиган. Алгоритмы адаптивной фильтрации нестационарных сигналов на базе параллельных вычислений. *Государственное унитарное предприятие г. Москвы Научно-производственный центр «Электронные вычислительно-информационные системы»*, 21(14):330–333, 2023.

[5] H. M. de Oliveira J. E. Wesen, V. Vermehren V. Adaptive filter design for stock market prediction using a correlation-based criterion. *PALAVRAS CHAVE. Filtragem adaptativa, Bolsa de Valores, correlação*, 21(16):330–333, 2017.

[6] Шульга Валентин Александрович. Сравнительный анализ алгоритмов машинного обучения в задачах исследования фондового рынка. *https://dspace.spbu.ru/bitstream/11701/26368/1/Diplomnaa_rabota_.pdf*, 21(18):30 –51, 2020.

[7] https://evilinside.ru/kak-predskazat-cenu-akcij-algoritm-adaptivnoj-filtracii/. -, 21(8):1–1, 2020.

[8] ITI Capital. https://habr.com/ru/company/iticapital/blog/274821/. *Habr.com*, 21(7):1–1, 2016.

[9] LechGrzelak Роланд В.И. https://github.com/LechGrzelak/Computational-Finance-Course. *Git-hub Repository*, 21(10):1–1, 2021.

[10] R. Seydel. Class notes on computational finance. *http://www.mi.uni-koeln.de/ seydel/classnotes.html*, 21(11):1–1, 2016.

[11] L.A. Grzelak C.W. Oosterlee. Mathematical modeling and computation in finance: With exercises and python and matlab computer codes. *World Scientific Publishing, 2019,* https://www.youtube.com/watch?v=IRMn6JQuU8A&list=PL6zzGYGhbWrPaI-op1UfNl0uDglxdkaOB , 21(12):1–11, 2019.

[12] Merkulov Daniil. https://mipt21.fmin.xyz/. *MIPT*, 21(12):330–333, 2023.