Машинное обучение в финансах

Лекция 1: Введение

Роман В. Литвинов*

*CRO
Финансовая Группа БКС

Высшая школа экономики, Январь 2021

Содержание

Содержание

Машинное обучение

«Джентльмены, когда я был молод это было статистикой, сейчас это называется машинное обучение» (c)

Две основные задачи:

- регрессия (regression)
- классификация (classification)

Машинное обучение (machine learning, ML)

Машинное обучение – набор методов, позволяющих компьютерам обучаться на имеющихся в нашем распоряжении данных для того, чтобы делать (и улучшать сделанные) предсказания/прогнозы (predictions).

Машинное обучение

Ключевые компоненты процесса обучения:

- некоторая решаемая задача/проблема
- мера оценки производительности (performance measure)
- опыт

Типы данных:

- размеченные данные
- неразмеченные данные

Виды машинного обучения:

- обучение на размеченной выборке (supervised learning)
- обучение на неразмеченной выборке (unsupervised learning)
- обучение с подкреплением (reinforcement learning)

Машинное обучение

Анатомия обучающегося алгоритма:

- целевая функция - loss/cost function (квадрат отклонений, напр. – squared error loss)
- критерий оптимальности/оптимизации (optimization criterion for loss function (cost function)) (MSE, как вариант)
- оптимизационная процедура/метод (optimization routine), которая с использованием тренировочных данных найдет необходимое решение для выбранного нами критерия

Основные алгоритмы:

- линейная регрессия, логистическая регрессия
- деревья решений
- метод опорных векторов, k-ближайших соседей
- нейронные сети, DNN и тп.

Машинное обучение. Эволюция

Основные события:

- 1940е-1960е. Кибернетика, Алан Тьюринг, первая нейронная сеть (М. Минский), первая модель персептрона (Розенблатт)
- 1970е. Первая «зима искусственного интеллекта»
- 1980е. Коннекционизм/ параллельная распределенная обработка (Румельхарт), back propagation , нейронные сети с двумя слоями
- 1987-1993. Вторая «зима искусственного интеллекта»: коллапс рынка LISP машин/ экспертных систем
- 1990е-2006. Сдвиг исследований от систем базирующихся на знаниях (knowledge-driven) к базирующимся на данных (data-driven), SVMs, RNNs, Каспаров и Deep Blue
- 2006-2019. AlphaGo, Deep Learning, развитие reinforcement learning, широкое распространение R и Python, Tensorflow, PyTorch, NLP и компьютерное зрение, quantum machine learning

Машинное обучение. Эволюция

Смена парадигм:

- Человек + Данные = Модель (Ньютон, Галилей и другие классические примеры)
- Человек + Данные = Модель, реализованная в виде Алгоритма
- Машина + Данные + Человек = Модель (все это еще требует настройки и руководства со стороны человека)
- Человек?

Смена парадигм:

Иоганн Кеплер, Стив Юрветсон, Третий закон Кеплера (гармонический закон) и “the biggest advance in engineering since the scientific method.” (действительно?)

Машинное обучение. Эволюция

Почему "на этот раз все будет иначе"?:

- Наличие большого количества данных: интернет, мобильные телефоны, носимые устройства/трекеры, сенсоры, «цифровые» города и тп. Глобальная смена точки зрения – накапливать данные «модно»/ руда для data science.
- Дешевые компьютерные мощности: один из ключевых компонентов (наряду с данными) для таких ресурсоемких, с точки зрения вычислительных мощностей, технологий как Deep Learning и DNNs. + появление специализированных чипов.
- Низкие барьеры для входа с т.з. знаний: age of discovery vs age of implementation, достаточно «крепких» инженеров (не элитных учетных), распространение Python/R, открытые библиотеки: Sci-kit Learn, TensorFlow, PyTorch, AI's open research culture.
- Смена основной парадигмы: не конструируем «универсального солдата» (general AI), а концентрируемся на применении в узкой области/решаем конкретную задачу (narrow AI).

Машинное обучение. Эволюция. Влияние на финансы

Несколько примеров передового (cutting edge) использования таких технологий в релевантной именно нам сфере (новостные сводки на эту тему действительно вдохновляют и пугают одновременно):

- Marty Chavez, CFO, Goldman Sachs, 2017 "In 2000, the U.S. cash equities trading desk at Goldman Sachs's New York employed 600 traders. Today there are just two equity traders left. Automated trading programs have taken over the rest of the work."
- Jack Ma, Founder, Mybank, 2019 "Real-time payments data and a risk-management system that analyzes more than 3,000 variables, MYbank has lent 2 trillion yuan (290 billion USD) to 16 million small companies. Process takes 3 mins and involves zero human bankers."
- Laurence Douglas Fink, CEO, BlackRock, 2018 "BlackRock released seven AI-powered sector ETFs in March 2018. ML techniques are being used to parse language in public filings to determine how to weight each company within the sector ETF."

Машинное обучение. Эволюция. Влияние на финансы

- Розничное кредитование: оценка рисков (в тч с помощью новых источников данных (социальные сети, транзакционные данные, активность в сети интернет, данные сотовых операторов и проч), автоматическое принятие решений, модель предсказания доходов, выявление мошенничества.
- Корпоративное кредитование: EWS + NLP (новости, транзакции), оценка рисков с помощью дополнительных источников данных в режиме реального времени (транзакции, иски, мониторинг посевов и тд).

Машинное обучение. Эволюция. Влияние на финансы

- Трейдинг: динамическое хеджирование (deep hedging), быстрая калибрация моделей (deep learning volatility), AAD (algorithmic adjoint differentiation), алгоритмическое (и автоматическое) управление рисками электронной торговли.
- Операционные риски: предиктивная аналитика (сбор и расчет KRIs и метрик потенциального риска в режиме реального времени), мониторинг аномальной активности персонала (операционисты, трейдинг, бэк-офис и тп), социальный скоринг (!!!)

Содержание

О чем этот курс

- Мы разберемся в том, как работают основные методы и технологии машинного обучения, как для задач классификации, так и регрессии (от линейной/ логистической регрессии до случайных лесов и глубоких нейронных сетей).
- Отдельное время будет посвящено процессу гипертюнинга параметров моделей, использованию бустинга, кросс-валидации и тп.
- В рамках данного курса вы сможете получить базовые знания программирования на языке на Python и познакомиться со всей необходимой для этого инфраструктурой и библиотеками (Jupyter notebook, Github, Pandas, Numpy, Sklearn, Matplotlib, etc).

О чем этот курс

- Примеры использования данных технологий будут базироваться на задачах, связанных с факторным инвестированием в целом и предсказанием форвардных доходностей акций, в частности.
- Для этого мы разберемся в процессах загрузки и обработки данных, визуализации, расчета факторов, тестирования идей и проч.
- Попутно студенты смогут получить знания о том, что такое факторное инвестирование, какие бывают факторы и в чем их смысл (risk/reversal, momentum, quality, growth, etc).

О чем этот курс. Факторное инвестирование

Небольшое отступление/упражнение: представьте, что вы зашли в супермаркет (автосалон) и выбираете для покупки яблоки (автомобиль).

Что по вашему мнению влияет на ваш выбор? Какие свойства определяют "хорошее"яблоко (машину), по вашему мнению? Какие факторы влияют на ваш выбор?

Факторное инвестирование (factor investing)

Факторное инвестирование – подход, в рамках которого инвестирование осуществляется на основе оценки факторов (специфических драйверов), влияющих на риск/доходность актива.

О чем этот курс. Quantitative/algorithmic trading

Форвардная доходность (forward return)

Форвардная доходность – доходность финансового инструмента (акции, облигации и тп), которая будет наблюдаться в определенное время в будущем.

Пример: Текущая vs Форвардная доходность (недельная) акций Тесла.

Количественный/алгоритмический трейдинг (quantitative/algorithmic trading)

Количественный/алгоритмический трейдинг – торговля финансовыми инструментами на основе количественных (математических) моделей. Эти модели в виде правил/алгоритмов переносятся на определенный язык программирования и исполняются машинами.

О чем этот курс. Quantitative/algorithmic trading

Минусы такого подхода:

- тяжело найти универсальное правило, применимое к любой ситуации на рынке
- сложно учесть в модели специфическую неколичественную информацию (мнение эксперта)

Плюсы:

- значительная диверсификация активов и стратегий
- учет и анализ огромного количества информации при принятии решения
- избегание поведенческих ошибок (behavioral biases) при инвестировании
- возможность объективного бэктестинга стратегий

О чем этот курс. Quantitative/algorithmic trading

Процесс количественного/ алго трейдинга (Quest for the Holy Grail):

1. data: поиск, очистка, хранение, настройка загрузки-выгрузки данных, трансформация и тп.
2. factors/ signals/ feature engineering: поиск и конструирование факторов, построение сигналов на основании этих факторов
3. strategy: разработка торговой стратегии на основании факторов
4. portfolio construction/ bet sizing: конструирование оптимального портфеля и поведения на рынке (market timing) с учетом существующих стратегий (эффекты корреляции, диверсификаци и тп), аллокация капитала
5. execution/ implementation: реализация стратегии в виде алгоритма на нектором языке программирования. Часто низкоуровневом, таком как C++ , например (low-latency)
6. performance attribution and risk-management: оценка и мониторинг производительности стратегии, управление рисками)

Содержание

Чему вы научитесь

В ходе курса мы в основном сосредоточимся на этапах 1 и 2 слайда 19. В конце курса, я надеюсь, вам станет понятно, что технологии ML применимы практически к любому этапу количественного трейдинга и, более того, могут быть объеденены в полноценный конвеер (ML pipeline). Что порождает интересные вопросы. Такие, например, как - где место человека в этой цепочке? Но... не будем забегать вперед.

Задача (для меня) будет выполнена, если в конце курса вы сможете сказать, что:

1. изучили (и поняли) основные методы и технологии машинного обучения, а также научились их успешно применять (на базе языка Python и соответствующих библиотек)
2. получили базовые навыки программирования на языке Python
3. ознакомились с понятием факторного инвестирования и процессом предсказания (prediction) форвардной доходности акций с использованием различных факторов.

Чему вы научитесь. Количество лекций и темы

- Лекция 1: Введение. О чем этот курс. Краткое содержание. Организационные вопросы. Подготовительная работы (установка Anaconda, Jupyter notebooks, etc)
- Лекция 2: Математика часть 1: Функции, производная, интеграл. Линейная алгебра.
- Лекция 3: Введение в программирование на Python. Numpy. Matplotlib.
- Лекция 4: Математика часть 2: Теория вероятностей. Статистика.
- Лекция 5: Введение в программирование на Python часть 2. Pandas.
- Лекция 6: Факторное инвестирование. Разновидности факторов и их смысл.
- Лекция 7: Введение в машинное обучение.
- Лекция 8: Подготовка данных. Feature engineering.

Чему вы научитесь. Количество лекций и темы

- Лекция 9: Линейная регрессия. Лассо. Ридж. Нелинейная регрессия.
- Лекция 10: Логистическая регрессия. Тюнинг гиперпараметров.
- Лекция 11: Методы повторной выборки (resampling). Кросс-валидация. Бустинг.
- Лекция 12: Деревья решений (decision trees). Случайные леса (random forest).
- Лекция 13: Обобщенные линейные модели (generalized additive models). (ОПЦИОНАЛЬНО)
- Лекция 14: Глубокие нейронные сети (deep learning networks).
- Лекция 15: Обучение с подкреплением (reinforcement learning).
- Лекция 16: Заключение. Темы для дальнейшего изучения. Видение: долгосрочное влияние ML/AI на финансы и эволюция профессии.

Чему вы научитесь. Структура курса

- 2 семестра
- Лекции
- Семинары
- Листки
- Проект

Содержание

Философия курса. Ричард Фейнман

"You can know the name of a bird in all the languages of the world, but when you're finished, you'll know absolutely nothing whatever about the bird... So let's look at the bird and see what it's doing — that's what counts. I learned very early the difference between knowing the name of something and knowing something."

Философия курса. Ричард Фейнман

Метод Фейнмана:

- Выберите концепцию, которую вы хотите выучить.
- Выучите ее так, как будто собираетесь объяснить ребенку. ("Test it this way: you say, “Without using the new word which you have just learned, try to rephrase what you have just learned in your own language.” Without using the word “energy,” tell me what you know now about the dog’s motion.” You cannot. So you learned nothing about science.")
- Найдите ошибки/ сложные места в вашем объяснении. Вернитесь к материалу, посмотрите другие источники. Экспериментируйте. Тестируйте. Записывайте свои мысли. Изменяйте ваше объяснение.
- После того как вам удалось "собрать"ваше объяснение проанализируйте его и попробуйте сделать еще проще.

Философия курса. Рави Вакил

"It is tricky to get things out of talks, even after a lot of practice. It is very easy to go to a talk, and at some point have your eyes glaze over. Talks are like horses: once you are thrown off, it is hard to get back on. Especially if the horse is stomping on your face."

Философия курса. Рави Вакил

"Three Things"is an exercise to learn how to get things out of talks. It can be useful if you are in the first few years of going to seminars — I've intended it as practice for graduate students — but I've also found that I got much more out of talks (especially those out of my comfort zone) when using it.

The theory is as follows. If you can get even three small things out of a talk, it is a successful talk. And if you can't get even three small things out of a talk, it was not a successful experience. Note that the things you get out of a talk needn't be the things that your neighbor got out of a talk, or the things the speaker expected you to get out of the talk."

`https://math.stanford.edu/~vakil/threethings.html`

Философия курса. Рави Вакил

"Here is how it works. Take a clean sheet of paper, or an index card. Your goal is to have three things, and only three things, on this sheet at the end of the talk. The "things"can be of many forms:

- a definition you want to remember
- a theorem you want to remember
- a motivating or key example
- a motivating problem
- a question you want to ask the speaker
- a question you want to ask someone else
- anything else of a similar flavor: something specific that made you think"

`https://math.stanford.edu/~vakil/threethings.html`

Содержание

Подготовительная работа. Установка и настройка инфраструктуры

Критически необходимой частью курса будет являться програм - мирование на языке Python.

Мы будем экспериментировать с кодом в интерактивном режиме во время лекции или семинара. Для этого у каждого из вас должна быть установлена интерактивная среда разработки Jupyter notebook.

`https://jupyter.org`

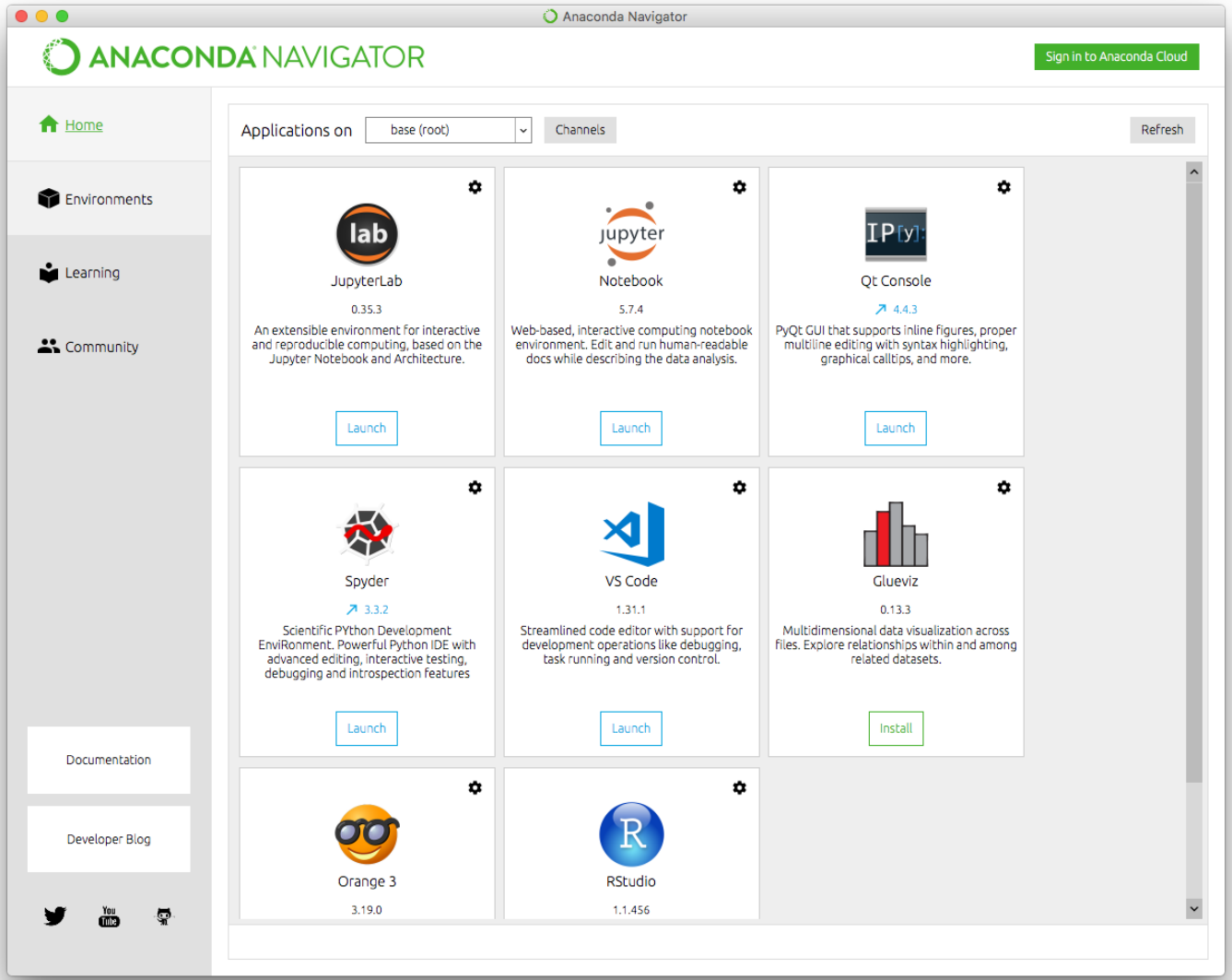
Вам также понадобится интегрированная среда разработки Spyder.

`https://www.spyder-ide.org`

Самый простой и удобный путь - это установить дистрибутив Anaconda на свой компьютер. Вы получите все необходимые средства разработки (Spyder, Jupyter) и библиотеки (Numpy, Matplotlib, Sci-kit learn, etc.)

`https://www.anaconda.com/products/individual`

Подготовительная работа. Установка и настройка инфраструктуры



Содержание

Литература

Рекомендуемая основная литература курса:

- Hastie, T., Tibshirani, R., Friedman, J. (2017). The elements of statistical learning. Springer. (есть на русском)
- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An introduction to statistical learning. Springer. (есть на русском)
- Kuhn, M., Johnson, K., (2016). Applied predictive modeling. Springer. (есть на русском)

Литература

Рекомендуемая дополнительная литература курса:

- McKinney, W. (2018). Python for data science. O'Reily. (есть на русском)
- Scopatz, A., Huff, K. (2015) Effective computation in physics. O'Reily.
- Deisenroth, M. Fisal, A., Ong, C.S. (2020) Mathematics for machine learning. Cambridge.
- Goodfellow, I. Bengio, Y., Courville, A. (2016). Deep learning. MIT. (есть на русском)
- Kroese, D., Botev, Z. (2019) Data Science and Machine Learning. Chapman and Hall/CRC
- Zhou, X., Jain, S. (2014) Active equity management. Xinfeng Zhou.
- Quian, E., Hua, R., Sorensen, E. (2007) Quantitative equity portfolio management. Chapman and Hall/CRC

Литература

Дополнительная литература к сегодняшней лекции:

- Goodfellow, I. Bengio, Y., Courville, A. (2016). Deep learning. MIT. (Ch.1) (есть на русском)
- Ilmanen, A. (2011) Expected returns. Wiley. (Ch.1)
- Murphy, K. (2012) Machine learning. MIT (Ch.1)
- Pedersen, L.J. (2015) Efficiently inefficient. Princeton University Press. (Ch.1-5, Ch.9)
- Фейнман, Р. (2008) Вы, конечно, шутите, мистер Фейнман! Колибри.

Содержание

Вопросы и ответы

Вопросы и ответы