Машинное обучение в экономике

Введение в машинное обучение

Потанин Богдан Станиславович

доцент, кандидат экономических наук

2024–2025

Машинное обучение

Определение

- Машинное обучение возникает во второй половине 20-го века на стыке множества дисциплин, таких как математическая статистика, искуственный интеллект и оптимизация.
- **Проблема определения** – как и множество других дисциплин, таких как эконометрика и интеллектуальный анализ данных (data mining), машинное обучение используется для анализа данных.
- **Ключевое отличие** – машинное обучение концентрируется, преимущественно, на получении точных прогнозов, в отличие, например, от эконометрики, в которой основной упор делается на получении точных оценок параметров модели.
- Например, в машинном обучении мы можем приследовать цель спрогнозировать заработную плату, а в эконометрике – точно оценить параметры, характеризующие влияние образования на зараплату. В качестве таких параметров могут рассматриваться, например, регрессионный коэффициент при переменной на образование или средний эффект воздействия.

История анализа данных

Ранние этапы анализа данных

- С древних времен люди собирали, систематизировали и анализировали данные.
- Например, в древнем Египте и древнем Риме собирались данные об урожае и налогах, которые впоследствии использовались в административных целях.
- В Вавилоне собирались астрономически данные, которые применялись для прогнозирования движения планет.
- В 17 веке возникают методы анализа социо-демографических данных. Исследователи начинают искать закономерности в этих данных и делать содержательные выводы.
- Пионером в данной области стал Джон Граунт, который проанализировал таблицы смертности в Лондоне, регулярно публиковавшиеся с 1603 года и содержавшие информацию о количестве смертей и их причинах, а также возрасте и поле умерших.
- В частности, Джон Граунт обратил внимание на то, что мужчин рождается и умирает больше чем женщин, а также оценил вероятность дожить до определенного возраста как долю оставшихся в живых к соответствующему возрасту.

История анализа данных

Теория вероятностей и математическая статистика

- **Проблема** – на ранних этапах анализ данных представлял собой множество разрозненных, преимущественно эвристических методов, решавших частные задачи и не объединенных общей теоретической базой.
- Без теоретической базы сложно развивать методы, изучать их эффективность, сравнивать эти методы между собой и обосновывать их применимость к той или иной задаче.
- **Решение** – использование теории вероятностей, а впоследствии и сформировавшейся на ее базе математической статистики, как теоретической базы для методов анализа данных.
- Математическая статистика стала активно развиваться со второй половины 17 века, в особенности, после публикации доказательства закона больших чисел для бернуллиевских случайных величин Яковом Бернулли в 1713 году.
- Например, в начале 19 века Адриен Мари Лежандр (1805 год) и Карл Фридрих Гаусс (1795 или 1809 год) предложили метод наименьших квадратов (МНК) и обосновали его с точки зрения теории вероятностей. Кроме того, в 1810 году Пьер-Симон Лаплас использовал доказанную им центральную предельную теорему для обоснования асимптотических свойств МНК оценок.
- С начала 20-го века активно развиваются теории построения доверительных интервалов и тестирования гипотез. В результате обоснование метода анализа данных с точки зрения теории вероятностей позволяет «подключить» к нему эти полезные инструменты.

- В 20 веке математическая статистика начинает активно применяться для анализа экономических данных, в частности, оценивания параметров экономических моделей, в результате чего формируется эконометрика.
- Одними из ключевых основателей эконометрики были Ян Тинберген и Рагнар Фриш, получившие в 1969 году первую в истории нобелевскую премию по экономике, за создание прикладных динамических моделей анализа экономических процессов.
- Существенный вклад в развитие эконометрики внес и Евгений Евгеньевич Слуцкий, применявший теорию случайных процессов для анализа макроэкономических данных. Также, он внес серьезный вклад в микроэкономику (уравнение Слуцкого: эффект замещения и эффект дохода) и теорию вероятностей (теорема Слуцкого).

История анализа данных

Развитие эконометрики

- Экономисты традиционно развивают эконометрику не только за счет применения математической статистики, но и внося **вклад** в ее развитие. Так, например, нобелевский лауреат по экономике Ларс Петер Хансен предложил обобщенный метод моментов (GMM) и изучил его свойства. Также, экономист Филип Грин Райт предложил метод инструментальных переменных.
- **Особенность эконометрики** – часто направление вклада в математическую статистику мотивируется содержательными экономическими соображениями и теоретическими моделями. Например, модель Роя мотивировала развитие статистических методов анализа усеченных и цензурированных данных. Наиболее известными из этих методов являются метод Тобина и метод Хекмана.
- Популярность эконометрики и тесная связь ее развития с математической статистикой, вероятно, обусловили то, что применение математической статистики в других социальных науках, таких как политология, психология и социология, также часто именовалось эконометрикой.

История анализа данных

Искусственный интеллект

- Исследования в области искусственного интеллекта (ИИ / AI) приследовали цель создания компьютеров, способных решать присущие человеку задачи, такие как распознование изображений, игра в шахматы и общение.
- На ранних этапах искусственный интеллект часто опирался на заранее запрограммированные человеком правила. Например, многие программы для игры в шахматы использовали различные алгоритмы, опиравшиеся на теорию игр.
- **Проблема** – искуственный интеллект опирается на заранее детерминированные человеком правила, что ограничивает его способность адаптироваться к новой информации и может потребовать больших временных затрат на разработку новых алгоритмов.
- **Решение** – обучить компьютер решать задачи с помощью аналализа данных. Например, на основании анализа тысяч игр можно обучить компьютер играть в шахматы.
- Машинное обучение занимается созданием исскуственного интеллекта, обучающегося на данных.

Сферы применения машинного обучения

Искусственный интеллект

- Сфера применения машинного обучения в задачах искусственного интеллекта крайне обширна, поэтому, перечислим лишь некоторые примеры.
- **Компьютерные игры** – поведение неигровых персонажей (NPC), процедурная генерация локаций (пещер, планет и т.д.), адаптация условий игры, например, параметров сложности в зависимости от результативности игрока.
- **Музыка** – замена голоса одного певца на другого (AI cover), генерация музыки.
- **Видео** – замена изображения одного человека на другого (deepfake), генерация видео, анализ видео, например, для выявления преступлений.
- **Текст** – генерация текстов и оценивание их тональности (веселый / грустный, позитивный / негативный).
- **Изображения** – генерация изображений, анализ изображений, например, для определения факта наличия на них определенных объектов (людей, животных и т.д.).

Сферы применения машинного обучения

Прогнозирования в экономических задачах

- С использованием машинного обучения можно прогнозировать дефолты по кредиту, продажи, уход клиента, переход по ссылке, цены акций и т.д.
- Например, прогноз вероятности дефолта можно использовать при принятии решения о выдаче кредита или о процентной ставке по нему, а также для того, чтобы заранее выявить клиентов, находящихся в группе риска, и принять соответствующие меры, например, по реструктуризации кредита.
- Особую роль машинное обучение играет в построении рекомендательных систем (фильмы, музыка, товары и т.д.). С учетом того, что эти системы охватывают крайне большое число пользователей, даже небольшое повышение точности прогнозов рекомендательных систем может привести к крайне существенному (в абсолютной величине) росту прибыли фирмы.
- Например, если ваша система за день показывает 1 миллиард рекомендаций и каждая успешная рекомендая приносит вам 1 цент, то, при прочих равных, повышение вероятности успешной рекомендации даже лишь на 0.01 принесет фирме дополнительные $10^9 * 0.01 * 0.01 = 10^5 = 100000$ долларов.

Сферы применения машинного обучения

Эконометрика

- В эконометрике машинное обучение часто используется для повышения точности оценок вспомогательных параметров (nuisance parameters), которые сами по себе не подлежат содержательной интерпретации, но используются для оцениваниях других, имеющих экономический смысл параметров.
- Например, в задач оценивания эффектов воздействия (ATE, ATET, CATE и т.д.) машинное обучение используется для оценивания условных математических ожиданий (обычно не интерпретируются), которые, впоследствии, используются для оценивания эффектов воздействия (интерпретируются).
- Также, машинное обучение активно применяется для получения вспомогательных переменных для эконометрического анализа.
- Например, исследователь может проанализировать тональность новостей и, сформировав соответствующую переменную, использовать ее для прогнозирования цен акций или инфляции.

Информация о курсе

Цели и структура

- **Цель** – научиться применять методы машинного обучения для:
  - прогнозирования социально-экономических показателей.
  - эконометрического анализа, в частности, оценивания эффектов воздействия и создания вспомогательных переменных.
- Цели курса обуславливают следующие акценты:
  - объяснение принципов работы методов машинного обучения с точки зрения теории вероятностей и математической статистики.
  - акцент на методы обучение с учителем в задачах классификации и регрессии.
- Основная информация, включая презентации лекций и материалы семинаров, публикуется на викистраничке курса:

http://wiki.cs.hse.ru/index.php?title=Машинное_обучение_в_экономических_-исследованиях_2024-2025

Информация о курсе

Оценка

- Формула оценки за курс:

$$\text{Оценка} = 0.01 \times \text{ДЗ1} + 0.29 \times \text{ДЗ2} + 0.70 \times \text{Экзамен}$$

- **Домашнее задание 1** (ДЗ1) – разбиться на группы до 2-х человек и определиться с темой второго домашнего задания. Например, оценивание влияния образования на зарплату.
- **Домашнее задание 2** (ДЗ2) – выполняется в группах до 2-х человек и концентрируется на оценивании эффектов воздействия с помощью методов машинного обучения.
- **Экзамен** – задачи и открытые вопросы по материалам лекций.
- **Творческий акцент** – домашнее задание выполняется на самостоятельно симулированных (созданных) данных, генерация которых совпровождается содержательным обоснованием предполагаемого процесса генерации данных.