

Causal Inference: прозрение и практика

Лекция 1. Основные понятия Causal Inference

Юрашку Иван Вячеславович

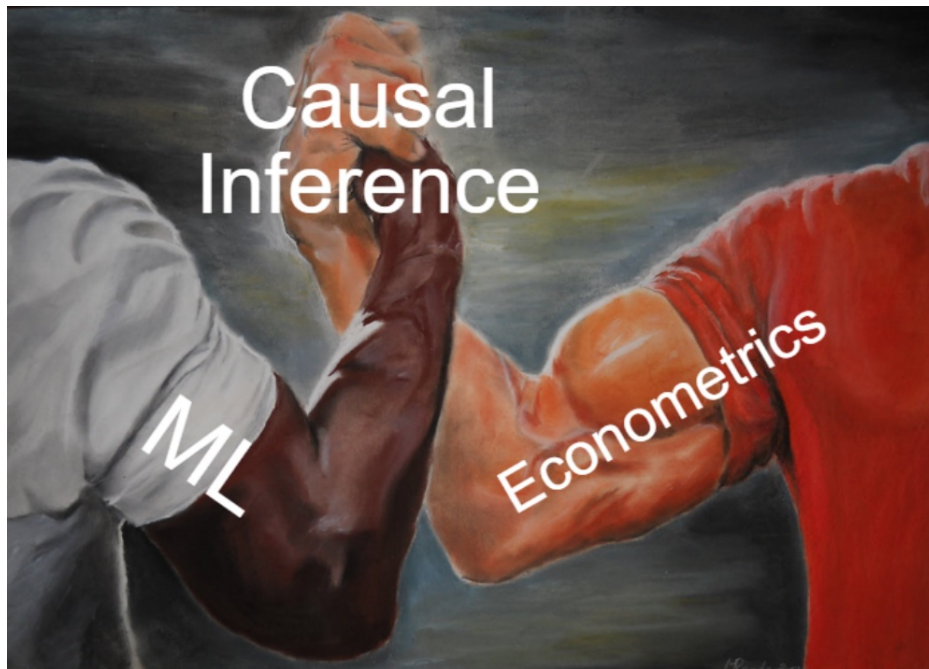
May 29, 2024

Causal Inference - это МЛ?

В современном мире Data Science играет ключевую роль в анализе и использовании данных. Мы живем в эпоху, где информация - это золото, а умение извлекать из нее ценные знания - наша сила. Однако часто понятие Data Science ограничивается лишь алгоритмами машинного обучения или даже искусственным интеллектом, умаляя другие важные аспекты этой дисциплины.

Вот где начинается история о сближении двух мощных инструментов: эконометрики и Machine Learning. В разных эпохах они казались как бы двумя противоположными полярностями в анализе данных. Машинное обучение стремилось к высокой точности прогнозов, зачастую уступая интерпретируемости моделей. С другой стороны, эконометрика ставила акцент на интерпретируемость, понимание причинно-следственных связей, иногда уходя в тень из-за ограниченности моделей.

Однако со временем стало понятно, что для полного понимания данных нам нужно объединить эти подходы. И здесь на сцену выходит Causal Inference, или причинно-следственная связь. Этот инструмент помогает нам разгадывать причины за явлениями, объединяя преимущества как машинного обучения, так и эконометрики. Так, Judea Pearl в своей статье 2021 года подчеркивает важность CI как ключевого элемента для достижения баланса между эмпирическим и интерпретируемым.



Таким образом, Causal Inference — это область статистики и научных исследований, направленная на выявление и измерение причинно-следственных связей между переменными. Она помогает определить, какое воздействие оказывает изменение одной переменной на другую, отличая это воздействие от простых корреляций.

Погружение в мир причинно-следственной связи и машинного обучения не только расширит ваш кругозор, но и даст вам ключ к разгадке сложных и важных вопросов, стоящих перед современным обществом.

Допустим, вы владеете интернет-магазином и хотите понять, какие факторы влияют на продажи. С помощью методов причинно-следственной связи вы сможете определить, какие из ваших маркетинговых кампаний действительно приносят наибольший доход, и направить свои усилия и ресурсы в нужное русло.

Еще один пример - медицинская сфера. С помощью анализа причинно-следственных связей можно выявить, какие лечебные методы наиболее эффективны для конкретного заболевания, что позволит разрабатывать более точные и эффективные методики лечения.

Этот курс - не просто набор теории, он предлагает вам практические инструменты для анализа данных и принятия обоснованных решений. С его помощью вы сможете выйти за рамки обычных аналитических методов и раскрыть потенциал данных, лежащих у вас под рукой. Полученные знания не только помогут читателю в работе, но и дадут возможность вносить реальные изменения в мир вокруг нас.

И все же Causal Inference - это не ML

Машинное обучение в настоящее время успешно решает задачи прогнозирования. Как подчеркивают Ajay Agrawal, Joshua Gans и Avi Goldfarb в книге "Prediction Machines":

"Новая волна искусственного интеллекта на самом деле приносит нам не интеллект, а важный компонент интеллекта - прогнозирование".

С машинным обучением мы можем совершать самые разнообразные и впечатляющие вещи. Главное требование заключается в том, чтобы сформулировать наши задачи как задачи прогнозирования. Хотите перевести текст с английского на португальский? Тогда создайте модель машинного обучения, которая предсказывает португальские предложения по английским. Хотите распознавать лица? Тогда разработайте модель машинного обучения, которая определяет наличие лица в определенной области изображения. Хотите создать автомобиль с автоматическим управлением? Тогда создайте модель машинного обучения, которая предсказывает направление поворота руля, а также давление на тормоза и акселератор при предоставлении изображений и данных с сенсоров, полученных из окружающей среды автомобиля.

Однако ML - не панацея. Он может производить чудеса в рамках строгих условий, но при этом может потерпеть крах, если данные немного отличаются от того, что модель привыкла видеть.

Машинное обучение известно своей неспособностью решать проблемы обратной причинности. Оно требует ответа на вопросы типа "а что, если", которые экономисты называют контрфактуальными. Как отмечается в "Prediction Machines", ML не справляется с такими задачами. Оно может предсказывать на основе данных, но не может оценить воздействие изменений. В качестве примера из книги "Prediction Machines":

"Во многих отраслях низкая цена ассоциируется с низкими продажами. Например, в гостиничной индустрии цены низки вне туристического сезона, а в период пикового спроса цены высоки и гостиницы полностью заполнены. Исходя из этих данных, наивное предположение может подсказать, что повышение цены приведет к увеличению числа проданных номеров".

По сути, ответ на вопросы о причинности является более сложной задачей, чем многие могут подумать. Это то, чему посвящен курс "Causal Inference: прозрение и практика". В нем мы исследуем, как использовать данные для изучения причинно-следственных связей и оценки воздействия вмешательств на результаты. Сперва определимся с тем, что конкретно мы хотим научиться делать.

Постановка задачи и обозначения

Формализуем задачу следующим образом.

Пусть существует множество объектов, которые нас интересуют (обозначим его \mathcal{U} от слова universe). Изучаемым объектом может быть пациент, потенциальный клиент коммерческой компании, город — что угодно. Значение произвольного параметра X одного конкретного объекта i будем обозначать с добавлением верхнего индекса, а вектор значений, соответствующий этому параметру - будет употребляться без индекса: X^i , \mathbf{X} .

Пусть существует возможность подействовать на объект. В реальном мире это может быть, например, лечение пациента, рекламная кампания для потенциальных клиентов, введение администрацией города юридических ограничений на поведение в людных местах. Ограничений практически нет. Оказанное влияние будем рассматривать как двоичный признак T_i со значениями из $\{0, 1\}$, без учета возможной интенсивности. А сравниваемые множества объектов обозначим буквами \mathcal{A} и \mathcal{B} .

В терминологии causal inference воздействие, которое исследуется или рассматривается, называется "treatment" (лечение) из-за того, что этот термин часто используется в контексте медицинских испытаний, где исследуют воздействие определенного лечебного метода или медикамента на пациентов. Однако в более широком смысле "treatment" может означать абсолютно любое воздействие на часть исследуемой системы объектов. Следуя прижившимся нормам, обозначим этот параметр словом treatment, или, кое-где в формулах для краткости просто буквой T . Формально, когда мы говорим о воздействии - это всего лишь разделение группы объектов на две части по некоторому бинарному признаку. Фактически, какого-либо значимого воздействия может и не быть. В таком случае мы имеем дело с фиктивным воздействием - часто встречаемая ситуация.

Целевую переменную изучаемого объекта будем обозначать как "target" или Y . Обычно это вещественная величина. Договоримся сразу, дабы избежать путаницы, что Y - это показатель, измеряемый в конкретный промежуток времени, часто - в еще не наступивший. Например, при исследовании влияния на клиентов сентябрьских смс-оповещений, нас может интересовать, каким образом это отразится на количестве их покупок в декабре. Для нас это будет целевой переменной Y . При этом показатель количества покупок в сентябре - это уже совсем другая величина. Для наглядности назовем ее лаговым значением целевой переменной. Когда она будет нам попадаться - будем обозначать ее отдельным символом, например $Y_{\text{lag } 3 \text{ month}}$.

Теперь представим, что для каждого изучаемого объекта существует две псевдовселенных, отличающихся одним только наличием воздействия на этот объект. Пусть также мы могли бы узнать значения целевой переменной как при $T = 1$, так и при $T = 0$. Обозначим эти величины за $Y_{i0} = Y_i|_{T=0}$ и $Y_{i1} = Y_i|_{T=1}$. Их разность обозначается как "treatment effect" ($TE_i = Y_{i1} - Y_{i0}$). Эта величина - реальное отражение эффекта воздействия на объект i .

Кроме того, одна из двух вымышленных нами вселенных совпадает с реальной. Все реальные значения называются factual, а параллельные им - contrfactual. Например, если на какой-то объект i_1 подействовали в реальности, то он принадлежит множеству \mathcal{A} . Также $Y_{i1} = Y^{\text{factual}}$, $Y_{i0} = Y^{\text{contrfactual}}$.

Для каждого объекта существует свой TE_i . В рамках решаемой задачи наш истинный интерес заключается в том, чтобы оценить так называемую величину **среднего эффекта от воздействия**, или average treatment effect (ATE).

$$ATE_M = \frac{1}{n} \sum_{i=1}^n TE_i = \mathbb{E} TE = \mathbb{E} (Y_1 - Y_0).$$

Поскольку множество \mathcal{M} (matter objects), по которому мы усредняем, не всегда будет совпадать с множеством \mathcal{U} всевозможных изучаемых объектов, будем уточнять множество усреднения посредством нижнего индекса.

Кроме того, существуют прижившиеся понятия и обозначения: • $ATE = ATE_U AT T = ATE_A AT = ATE_B$, :

$$\mathbb{E} (Y_1 - Y_0) = \mathbb{E} (Y|_{T=1} - Y|_{T=0}) - \text{средняя разность} \dots$$

$$\mathbb{E} (Y|T = 1) - \mathbb{E} (Y|T = 0)$$

Для более точного описания процессов АВ тестирования и причинных выводов введем следующие математические обозначения:

- \mathcal{U} — множество всех объектов нашего исследования.
- Y — метрика отклика, например, конверсия или доход.
- T_i — бинарная переменная — индикатор принадлежности объекта i к тестовой группе:
 - $T = 1$ — контрольная группа (\mathcal{A}).
 - $T = 0$ — тестовая группа (\mathcal{B}).
- $Y_i^{(1)}$ — отклик объекта i в тестовой группе.
- $Y_i^{(0)}$ — отклик объекта i в контрольной группе.

Наглядно

*** пример в табличном виде с примером импорта ***

Нам повезло. Имея *contrfactual* значения мы разобрались, что к чему. Легко и идеально точно решили задачу. К сожалению, в действительности нам недоступен замер величин из параллельных вселенных. Поэтому были изобретены методы аппроксимации АТЕ по имеющимся данным. Чтобы получше разобраться, рассматриваем после каждого пункта приведенный из хайпекса пример.

Correlation is not causation

"Все события вымышлены, любые совпадения случайны" - приблизительно так обязывают нас правила этики начать данный параграф.

Итак, в одной стране, в одном крупном городе, существовал госпиталь для лечения определенного рода больных. Через дорогу от него решили возвести абсолютно новый медицинский центр, специализирующийся на том же смертельном недуге. Он был оснащен всеми передовыми технологиями и туда пригласили работать самых лучших докторов этой области медицины. Целью было внедрить новые методы хирургии. Поскольку любой такой новый дорогостоящий проект хотелось бы обмерять со всех сторон, было принято решение некоторое время позволить функционировать обеим организациям для того, чтобы выявить разницу в подходах к лечению на практике.

Информация о новом кампусе разлетелась со скоростью звука, выдыхаемого друзьями или родственниками тяжело заболевших (это несколько другая скорость, заметно превышающая обычную скорость звука), и вскоре люди хлынули за возможностью получить необыкновенный шанс.

Прошли месяцы, и результаты нового медицинского центра начали анализироваться. По первоначальным отчетам казалось, что он показывает более высокую смертность среди пациентов по сравнению с соседним госпиталем. Однако, при более детальном рассмотрении данных стало ясно, что существовал статистический *bias* в распределении пациентов между этими двумя учреждениями.

Оказалось, что вновь открывшийся медицинский центр привлекал в основном пациентов с более тяжелыми формами заболевания, так как люди их семьи или родственники стремились обратиться туда за получением новейших методов лечения. В то время как госпиталь продолжал принимать пациентов с менее серьезными случаями.

Этот *bias* в распределении пациентов искажал общую картину, делая исходные данные о смертности в медицинском центре недостоверными для сравнения с госпиталем. Таким образом, необходимо быть внимательным к подобным факторам при анализе результатов исследований, чтобы избежать искажений и сделать более точные выводы о качестве медицинского обслуживания.