Causal Inference: прозрение и практика

Лекция 1. Основные понятия Causal Inference

Юрашку Иван Вячеславович

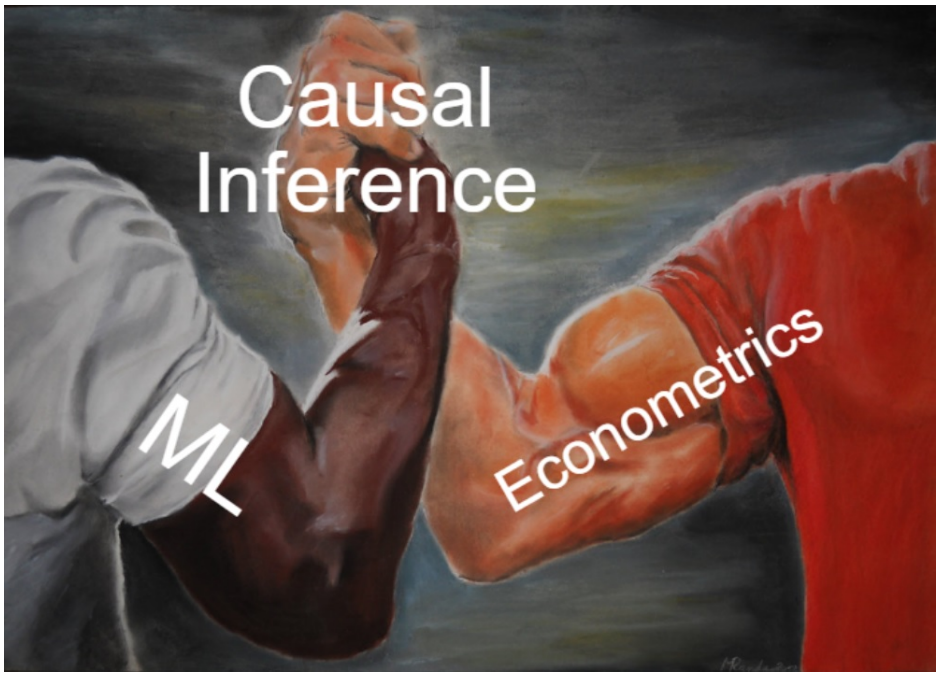
June 26, 2024

Введение

В нашем веке центральное место в анализе и использовании данных занимает Data Science. Однако часто данное понятие сводят к одним лишь алгоритмам машинного обучения или даже искусственному интеллекту, преуменьшая другие важные аспекты этой области знаний.

История формирования современной науки о данных началась со сближения двух могущественных инструментов - эконометрики и машинного обучения. В разные времена они казались двумя противоположностями в анализе данных. Машинное обучение было ориентировано на высокую точность прогнозов, порой жертвуя понятностью моделей. Эконометрика же делала акцент на интерпретируемости, понимании причинно-следственных связей, иногда оставаясь в тени из-за ограниченности моделей.

Однако со временем стало ясно, что для полного понимания данных необходимо научиться объединять эти два подхода. Здесь на сцену выходит причинно-следственный вывод (Causal Inference). Эта область Data Science помогает раскрыть причины явлений, объединяя преимущества как машинного обучения, так и эконометрики. Judea Pearl в своей статье 2021 года подчеркивает важность причинно-следственного вывода как “ключевого элемента для достижения баланса между радикальным эмпиризмом ML и интерпретационным подходом эконометрики”.



> 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.

Возможность с помощью Causal Inference не только прогнозировать события, но и понимать их причины, делает его неотъемлемой частью современной науки о данных, значительно повышающей качество принимаемых решений и эффективность использования доступной информации.

В этом цикле статей мы рассмотрим базовые и продвинутые инструменты Causal Inference в рамках задачи выявления «причинности» и исследования «эффекта». Кратко коснемся теоретических аспектов и научимся применять некоторые алгоритмы на практике. Будем использовать язык программирования python, библиотеку HypEx с реализацией алгоритмов. В качестве примеров будут использованы как синтетические, так и реальные наборы данных.

А ML-то голый!

Безусловно, ML - мощный инструмент для обработки информации. И с его помощью 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 предскажет наличие лица в определённой 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 сенсорных данных, полученных из окружающей среды автомобиля. 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".

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

Машинное обучение известно своей недостаточной способностью решать проблемы обратной причинности. Решение таких проблем требует ответа на вопросы типа "а что, если", которые экономисты называют контрфактуальными. Что произойдет, если я использую другую цену вместо той, которую я запрашиваю за свой товар в настоящее время? Что произойдет, если я буду придерживаться диеты с низким содержанием сахара вместо диеты с низким содержанием жиров, на которой я нахожусь? Если вы работаете в банке, выдаете кредиты, вам придется разобраться, как смена клиентской линии влияет на вашу выручку. Или, если вы работаете в органах местного самоуправления, вас могут попросить придумать, как улучшить систему школьного образования. Стоит ли давать планшеты каждому ребенку, потому что эпоха цифровых знаний велит вам это сделать? Или стоит построить старомодную библиотеку?

Как отмечается в книге Matheus Facure Alves "Causal Inference for The Brave and True",

> "в основе этих вопросов лежит причинно-следственная связь ... и к сожалению для машинного обучения, мы не можем полагаться на прогнозы корреляционного типа, чтобы с ними справиться"

В качестве наглядной иллюстрации такой истории из "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.

Если мы владеем интернет-магазином и хотим понять, какие элементы дизайна сайта и маркетинговые кампании влияют на продажи, то с помощью методов причинно-следственной связи мы сможем определить, какие из них действительно приносят наибольший доход, и направить ресурсы в нужное русло.

Порой, в ходе изучения данных случается выявить совершенно неожиданные причинно-следственные связи. Как, например, исследование статистики заболевания оспой среди деревенских жителей, фермеров и доярок привело к изобретению первой в мире вакцины от оспы. Иногда эксперимент проходит с допущением ошибок. Тогда могут возникать заблуждения и неясные выводы. Например, знаменитый «зефирный эксперимент Уолтера Мишеля», его опровержение и опровержение опровержения.

Сперва определимся с тем, что конкретно мы хотим научиться делать.

Постановка задачи и обозначения

Формализуем задачу следующим образом.

Пусть имеется множество объектов, которые представляют интерес в рамках исследования (назовём его $\mathcal{U}$, от английского слова "universe"). Этими объектами могут быть пациенты, потенциальные клиенты коммерческой компании, города – всё, что угодно. Значение произвольного параметра X у объекта i будем обозначать с добавлением верхнего индекса, а вектор значений этого параметра – без индекса: X^i, $\mathbf{X}$.

Рассмотрим возможность воздействия на объект. Оно может принимать различные формы, включая лечение пациента, проведение рекламной кампании по привлечению клиентов или введение административных ограничений в определенных городах. Это лишь некоторые из бесчисленных вариантов воздействия, которые могут быть применены. В контексте причинно-следственного вывода, воздействие представляет собой разделение группы объектов на две части на основе бинарного признака. Однако на практике воздействие может быть совершенно несущественным или вовсе отсутствовать, и в таком случае мы имеем дело с фиктивным воздействием, что также является распространенным явлением.

В терминологии причинно-следственного вывода, воздействие, которое исследуется, называется "treatment" (лечение). Этот термин происходит из медицинских испытаний, где он используется для обозначения лечебного метода или медикамента, применяемого к пациентам. Однако в контексте причинно-следственного вывода "treatment" может обозначать любое воздействие на часть исследуемой системы. Это может быть рекламная кампания, изменение политики или любое другое вмешательство.

Мы будем представлять влияние в виде двоичного признака T^i, который может принимать значения 0 или 1, без учета промежуточной интенсивности. Обозначим сравниваемые множества объектов, разделенных по значению T, как $\mathcal{A}$ и $\mathcal{B}$ соответственно.

Целевую переменную изучаемого объекта обозначим как "target" или Y. Это обычно вещественная величина, измеряемая в определенный момент времени, часто в будущем. Важно отметить, что ее значение фиксируется заранее и не зависит от времени. Например, при исследовании влияния мартовских SMS-оповещений на клиентов нас может интересовать, как это отразится на количестве их покупок в июне. В этом случае количество покупок в июне будет нашей целевой переменной Y. В то же время количество покупок, скажем, в мае – это совершенно другая величина, называемая лаговым значением целевой переменной. Для определенности будем обозначать такие лаговые значения отдельным символом, например $Y_{\texttt{lag 2 month}}$

Представим, что для каждого изучаемого объекта существуют две параллельные вселенные, различающиеся только наличием воздействия на этот объект. Пусть мы можем узнать значения целевой переменной как при $T = 0$, так и при $T = 1$. Обозначим эти величины как $Y^i_{(0)}$ и $Y^i_{(1)}$. Их разность называется "treatment effect" ($TE^i = Y^i_{(1)} - Y^i_{(0)}$), которая представляет собой реальное отражение эффекта воздействия на объект i.

Кроме того, одна из этих вымышленных вселенных совпадает с реальной. Реальные значения называются factual, а параллельные им – counterfactual. Например, если на объект i_1 в реальности воздействовали и он принадлежит множеству $\mathcal{B}$, то его значения $Y^{i_1}_{(1)}$ и $Y^{i_1}_{(0)}$ будут являться $Y^{i_1}_{factual}$ и $Y^{i_1}_{counterfactual}$ соответственно.

Для каждого отдельного объекта существует свой TE^i. В рамках решаемой задачи наш истинный интерес заключается в том, чтобы оценить так называемую величину **среднего эффекта от воздействия** на множество $\mathcal{M}$, или average treatment effect ($ATE_{\mathcal{M}}$).

$$ATE_{\mathcal{M}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} TE^i = \mathbb{E}\, TE = \mathbb{E}\,(Y_{(1)} - Y_{(0)}).$$

Поскольку множество $\mathcal{M}$ (matter objects), по которому мы усредняем, не всегда совпадает с множеством $\mathcal{U}$ всех изучаемых объектов, будем уточнять множество усреднения с помощью нижнего индекса, когда это необходимо. Кроме того, существуют некоторые прижившиеся понятия и обозначения. Для удобства приведем список обозначений:

- $\mathcal{U}$ — множество всех объектов исследования.
- T — бинарная переменная, индикатор принадлежности объекта к целевой группе ($T = 1$) или

контрольной группе ($T = 0$).

- Y — целевая метрика (исход).
- Y^i_{factual} — фактическое, наблюдаемое значение исхода Y для объекта i.
- $Y^i_{\text{counterfactual}}$ — контрфактическое значение исхода Y для объекта i.
- $\mathcal{A}$ — контрольная группа, для которой $T = 0$.
- $\mathcal{B}$ — целевая группа, для которой $T = 1$.
- $Y^i_{(0)}$ — потенциальный исход объекта i в случае **не** подвергания воздействию.
- $Y^i_{(1)}$ — потенциальный исход объекта i в случае подвергания воздействию.
- $ATE = ATE_{\mathcal{U}}$ — средний причинный эффект (Average Treatment Effect) для всего множества объектов.
- $ATC = ATE_{\mathcal{A}}$ — средний причинный эффект на контрольную группу.
- $ATT = ATE_{\mathcal{B}}$ — средний причинный эффект на целевую группу.

Если мы знаем counterfactual

Рассмотрим синтетический пример.

```
import pandas as pd
import hypex
from hypex.utils import datasets


df = datasets.gen_oracle_df(
    factual_only=False,
    random_state=145
).loc[:, ['Treatment', 'Target_untreated', 'Target_treated']]
df
```

df

	Treatment	Target_untreated	Target_treated
0	0	600	650
1	0	500	550
2	1	500	550
3	0	600	650
4	1	500	650
5	1	700	850
6	1	300	450
7	0	600	750

```
Y_0 = df['Target_untreated']
Y_1 = df['Target_treated']

Y_factual = Y_0 * (1 - df['Treatment']) + Y_1 * df['Treatment']
Y_counterfactual = Y_0 * df['Treatment'] + Y_1 * (1 - df['Treatment'])

TE = Y_1 - Y_0

TE[df['Treatment'] == 0].mean(), TE[df['Treatment'] == 1].mean(), TE.mean()
```

Как мы можем здесь увидеть,

$$ATE_{\mathcal{A}} = 75$$

$$ATE_{\mathcal{B}} = 125$$

$$ATE_{\mathcal{A}\cup\mathcal{B}} = 100$$

Нам повезло: имея доступ к counterfactual значениям, мы легко и точно определили ATE. Однако в реальности мы не можем измерить величины из параллельных вселенных. Поэтому были разработаны методы аппроксимации ATE на основе доступных данных. Рассмотрим самый наивный из этих методов.

Simple mean difference

Рассмотрим классический пример, иллюстрирующий, что иногда очевидные выводы оказываются ошибочными. Возьмём два госпиталя: один действовал уже много лет, когда был построен второй. Новый госпиталь был оснащён передовыми технологиями и привлёк лучших специалистов. Однако в процессе времени выяснилось, что средний уровень смертности во втором госпитале значительно превысил показатели первого.

```
df = datasets.gen_special_medicine_df(
    data_size=1000,
    dependent_division=True,
    random_state=None,
)

df.head(8)
```

Sample data from pandas

disease_degree	experimental_treatment	residual_lifetime
3	1	10.25
1	1	0.80
2	1	20.32
3	1	4.89
1	1	5.92
1	0	4.78
1	0	11.42
2	1	2.06

```
(
    df.loc[df['experimental_treatment']==1, 'residual_lifetime'].mean() -
    df.loc[df['experimental_treatment']==0, 'residual_lifetime'].mean()
)
```

Output: -1.67

Мы получили совершенно контринтуитивный результат, даже по знаку противоположный естественным ожиданиям. Причиной этого стало то, что новый медицинский центр привлекал преимущественно пациентов с более тяжёлыми формами заболевания.

Это статистическое смещение (bias) в распределении пациентов искажало общую картину. Из-за чего нельзя было делать выводы на основании прямого сравнения смертностей в медицинских центрах.

Давайте попробуем устранить эти различия в распределениях, сделав данные однородными по тяжести заболеваний, и повторим наш эксперимент.

```
df = datasets.gen_special_medicine_df(
    data_size=1000,
```

```
        dependent_division=False,
        random_state=None,
    )

    df.head(8)
```

Sample data from pandas

disease_degree	experimental_treatment	residual_lifetime
3	1	10.25
1	1	0.80
2	1	20.32
3	0	4.14
1	1	5.92
1	0	4.78
1	0	11.42
2	1	2.06

```
(
    df.loc[df['experimental_treatment']==1, 'residual_lifetime'].mean() -
    df.loc[df['experimental_treatment']==0, 'residual_lifetime'].mean()
)
```

Output: 1.95

Видим, что значение ATE стало более правдоподобным.

Попробуем разобраться. Распишем для произвольного множества $\mathcal{M}$ значение, полученное при помощи simple difference method ($SD_{\mathcal{M}}$). Для краткости обозначим за $\mathcal{M}_{(0)}$ и $\mathcal{M}_{(1)}$ подмножества множества $\mathcal{M}$, обусловленные значениями T.

$$\begin{aligned}
SD_{\mathcal{M}} &= \frac{1}{|\mathcal{M}_{(1)}|}\sum_{i\in\mathcal{M}_{(1)}} Y^i_{\text{factual}} - \frac{1}{|\mathcal{M}_{(0)}|}\sum_{i\in\mathcal{M}_{(0)}} Y^i_{\text{factual}} \\
&= \frac{1}{|\mathcal{M}_{(1)}|}\sum_{i\in\mathcal{M}_{(1)}} Y^i_{(1)} - \frac{1}{|\mathcal{M}_{(0)}|}\sum_{i\in\mathcal{M}_{(0)}} Y^i_{(0)} \\
&= \frac{1}{|\mathcal{M}_{(1)}|}\sum_{i\in\mathcal{M}_{(1)}} \left(Y^i_{(1)} - Y^i_{(0)}\right) + \left(\frac{1}{|\mathcal{M}_{(1)}|}\sum_{i\in\mathcal{M}_{(1)}} Y^i_{(0)} - \frac{1}{|\mathcal{M}_{(0)}|}\sum_{i\in\mathcal{M}_{(0)}} Y^i_{(0)}\right) \\
&= ATE_{\mathcal{M}_{(1)}} + BIAS(\mathcal{M}, 0)
\end{aligned}$$

Следовательно,

$$ATE_{\mathcal{M}_{(1)}} = SD_{\mathcal{M}} - BIAS(\mathcal{M}, 0)$$

где выражение, обозначенное как $BIAS(\mathcal{M}, 0)$ можно интуитивно интерпретировать следующим способом. Если все бы изучаемые объекты находились во вселенной без воздействия на них, то эта величина описывала бы среднюю разницу Y. Иными словами, это показатель различия групп $\mathcal{M}_{(0)}$ и $\mathcal{M}_{(1)}$, которая не зависит от treatment.

Аналогично определяется

$$BIAS(\mathcal{M}, 1) = \frac{1}{|\mathcal{M}_{(1)}|}\sum_{i\in\mathcal{M}_{(1)}} Y^i_{(1)} - \frac{1}{|\mathcal{M}_{(0)}|}\sum_{i\in\mathcal{M}_{(0)}} Y^i_{(1)}$$

и доказывается

$$ATE_{\mathcal{M}_{(1)}} = SD_{\mathcal{M}} - BIAS(\mathcal{M}, 1)$$

Кроме того, из определения ATE и, вообще говоря, для любой пары непересекающихся множеств ($\mathcal{M} =$

$\mathcal{M}_{(0)} \amalg \mathcal{M}_{(1)}$) легко нетрудно видеть, что выполняется тождество

$$ATE_{\mathcal{M}} = \frac{|\mathcal{M}_{(1)}|}{|\mathcal{M}|} ATE_{\mathcal{M}_{(1)}} + \frac{|\mathcal{M}_{(0)}|}{|\mathcal{M}|} ATE_{\mathcal{M}_{(0)}}$$

Совмещая это и предыдущее выражение, получаем

$$ATE_{\mathcal{M}} = SimpleDiff_{\mathcal{M}} + \frac{|\mathcal{M}_{(1)}|}{|\mathcal{M}|} BIAS(\mathcal{M}, 1) + \frac{|\mathcal{M}_{(0)}|}{|\mathcal{M}|} BIAS(\mathcal{M}, 1) = BIAS(\mathcal{M})$$

Скажем простыми словами. BIAS - наш главный недруг. На этой теории построена бóльшая часть методов, которые будут описаны в нашем курсе.

Когда разделенные группы между собой достаточно похожи, то $BIAS(\mathcal{M})$ близок к нулю. В таком случае метод SimpleDiff вернет результат, довольно хорошо приближающий $ATE_{\mathcal{M}}$.

В следующей статье мы более детально рассмотрим класс методов, в которых используется данный подход.