Причинно-следственный анализ, часть 1

Из этого текста ты узнаешь:

- *как правильно оценивать влияние эффекта,*
- *как правильно готовить тестовые и контрольные выборки.*

1 Введение

Имя Роналда Фишера наверняка известно всем, кто прошёл курс математической статистики, ведь этот человек оказал огромное влияние на эту науку. Он ввёл всем известное понятие дисперсии, популяризировал применение метода максимального правдоподобия, разработал подход к дисперсионному анализу и, самое главное, создал столь хорошо всем известный датасет ирисы Фишера. Фишер практически всю свою жизнь был во враждебных отношениях с другим известным статистиком этого времени, Карлом Пирсоном, поскольку Фишер, как он считал, нашёл грубую ошибку в труде Пирсона, которую последний так и не признал.

Историю методов, которые мы обсудим сегодня, также можно начать с Фишера. В 1919 году он стал исследователем Ротамстедской сельскохозяйственной опытной станции для анализа большого количества данных и нашёл грубую ошибку в подходе, использовавшемся для сравнения удобрений. В то время для выявления более эффективного вида удобрений применялся следующий подход: выбиралось два разных поля, на каждом из которых рассаживали одну и ту же культуру, но на одном поле использовали одно удобрение, а на втором — другое. В конце сезона с полей собирали урожай и делали вывод о том, какое удобрение оказалось более эффективным. Фишер публично раскритиковал данный подход, ведь, по его мнению, на урожай оказывало влияние большое количество внешних факторов, таких как качество самой почвы. В качестве альтернативного подхода Фишер предложил использовать одно поле, разделённое на большое количество небольших участков. Для каждого участка случайным образом выбиралось удобрение, используемое на нём. Именно такой подход, по мнению Фишера, мог позволить объективно сравнивать качество удобрений. Впоследствии данный метод назвали «принцип рандомизации» и начали повсеместно использовать в A/B-тестировании.

2 Причинно-следственный анализ

Причинно-следственный анализ (causal inference) с определённой натяжкой можно назвать частным случаем дисперсионного анализа (ановы, ANOVA). В дисперсионном анализе рассматривается некоторое целевое значение Y, известное для множества объектов с описаниями X. Требуется выяснить, какие признаки из x влияют на значение y. К примеру, имеется множество фильмов, для которых известен производственный бюджет, режиссёр, актёры, жанр и так далее (так называемые факторы). Также для каждого фильма известны его сборы в первые выходные. Задачей выявления факторов, влияющих на сборы, занимается дисперсионный анализ.

Причинно-следственный анализ решает похожую задачу, но с другой стороны, он пытается избавиться от влияния всех признаков (также называемых факторами или ковариатами) при оценке влияния целевого признака. Подобно тому, что сделал Фишер, когда поставил эксперимент таким образом, чтобы на результат влиял только тип удобрения, но не погода, особенности почвы или географические особенности местности.

Формализуем данную задачу так, как это сделал Дональд Рубин в Rubin (1974). Предположим, что имеется некоторое множество из n объектов. К каждому объекту i мы можем применить или не применить некоторый эффект (лечение, удобрение и так далее), в зависимости от чего значение целевой метрики будет либо $Y_i(0)$ (эффект не был применён), либо $Y_i(1)$ (эффект был применён). Влияние данного эффекта на значения Y определяется как

$$\Delta_i = Y_i(1) - Y_i(0),$$

то есть как разница между применением и неприменением эффекта. Главная проблема, естественно, заключается в том, что лишь одно значение — $Y_i(0)$ либо $Y_i(1)$ — является наблюдаемым, то есть значения Δ_i никогда точно не известны. Однако следуя заветам Фишера, мы можем применить этот эффект к имеющимся объектам случайным образом и оценить среднее влияние эффекта (average treatment effect) как

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)].$$

Обозначим факт влияния эффекта через W_i, где $W_i = 0$ соответствует отсутствию влияния и $W_i = 1$ — его наличию. Тогда представим нашу выборку из n объектов через n пар (Y_i, W_i), где:

1. $Y_i = Y_i(W_i)$ — для каждого объекта известно значение целевой метрики в результате W_i;
2. W_i назначался 0 или 1 случайным образом для каждого объекта.

Тогда значение τ можно легко оценить через разницу в средних (difference in means):

$$\tau_{DM} = \frac{1}{n_0} \sum_{i:W_i=0} Y_i - \frac{1}{n_1} \sum_{i:W_i=1} Y_i,$$

где n_0 — число объектов, к которым не применялся эффект, $n_1 = n - n_0$.

К сожалению, в реальной жизни оценка τ_{DM} редко является наиболее эффективной. На каждый объект i может влиять большое число признаков X_i, влияние которых нельзя исключать. К примеру, применяя лекарство к больным, нужно учитывать пол, вес, возраст и множество других показателей. При случайном распределении эффектов можно принять за искомую разницу влияние других, более сильных параметров.

Для того чтобы сделать данный процесс более справедливым, существует множество методов.

3 Распутывание

Определение спутанности (confounding) было впервые дано в Rosenbaum and Rubin (1983). Спутанностью называется ситуация, когда при оценке влияния x на y существует спутывающая переменная z, которая влияет как на x, так и на y. Спутанность может возникать как из-за случайного распределения объектов, так и при назначении эффектов вручную. К примеру, лечащий доктор почему-то решил, что лекарство лучше действует на пожилых пациентов, оставив молодых в контрольной группе. Соответственно, цель распутывания состоит в поиске этих самых переменных. Математически это можно выразить так:

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[\mathbb{E}[Y_i(1)|X_i] - \mathbb{E}[Y_i(0)|X_i]] =$$
$$= \mathbb{E}[\mathbb{E}[Y_i|X_i, W_i = 1] - \mathbb{E}[Y_i|X_i, W_i = 0]] =$$
$$= \mathbb{E}[\mu_{(1)}(X_i) - \mu_{(0)}(X_i)],$$

где $\mu_{(w)}(X_i) = \mathbb{E}[Y_i|X_i, W_i = w]$. Далее $\mu_{(w)}(X_i)$ оцениваются через $\hat{\mu}_{(w)}(X_i)$, где $\hat{\mu}_{(0)}(X_i)$ обучается на контрольной группе, а $\hat{\mu}_{(1)}(X_i)$ — на тестовой. Далее оценка влияния эффекта оценивается как

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\mu_{(1)}(X_i) - \mu_{(0)}(X_i) \right).$$

Главная проблема данного подхода, очевидно, состоит в обучении $\hat{\mu}_{(w)}(X_i)$. Изначально данный метод был предложен для поиска линейных зависимостей, для которых МНК даёт отличный результат. Однако если данная зависимость является нелинейной, то всё становится намного сложнее: МНК даёт смещённый результат, а более мощные методы машинного обучения требуют большого количества данных для обучения. Поэтому данный метод обычно применяется только когда известно, что $\hat{\mu}_{(w)}(X_i)$ линейные.

4 Стратификация

Один из методов решения проблемы зависимости $\mathbb{E}[Y_i(w)|X_i]$ состоит в стратификации выборки. Для этого вся выборка разбивается по ковариатам на непересекающиеся подгруппы, а значения W_i назначаются так, чтобы в каждой подгруппе было равное число контрольных и тестовых объектов. Это помогает избавиться от дисбаланса при равномерном распределении. При маленьких выборках это помогает значительно повысить чувствительность методов. Однако в зависимости от выборки, в каждой подгруппе может оказаться всего по одному наблюдению (если слишком много признаков), что не даёт сделать полноценную стратификацию. Также если ковариаты не дискретные, а непрерывные, это может усложнить разбиение на подгруппы.

5 Мэтчинг

Более общим и предпочтительным способом разбиения выборки на тестовую и контрольную группы является так называемый мэтчинг. В ходе мэтчинга выборка разбивается на пары похожих объектов, чтобы один объект из пары был помещён в тестовую выборку, а второй — в контрольную. Это позволяет добиться того, чтобы и в тестовой, и в контрольной выборках были примерно одинаковые объекты.

Метод мэтчинга состоит из следующих основных шагов:

1. Составление списка возможных мер расстояния ρ.
2. Оценка получаемых разбиений при каждой из возможных мер.
3. Выбор наиболее подходящей меры схожести.

Далее будут рассмотрены некоторые конкретные методы сравнения.

5.1 Полное сходство

Предположим, что имеется два объекта X_1 и X_2:

$$\rho_e(X_1, X_2) = \begin{cases} 0, \texttt{если} X_1 = X_2, \\ \infty, \texttt{иначе}. \end{cases} \tag{1}$$

Данная мера расстояния позволяет находить совпадающие объекты, но обладает существенным недостатком при наличии непрерывных признаков: если они отличаются хоть немного, то объекты рассматриваются как абсолютно различные. Это приводит к использованию так называемого coarsened exact matching, когда непрерывные признаки могут быть не в точности равны, а отличаться на некоторое значение δ (причём δ может подбираться для каждого признака отдельно).

5.2 Расстояние Махаланобиса

Расстояние Махаланобиса является обобщением расстояния Евклида, учитывающим зависимости между признаками:

$$\rho_d(X_1, X_2) = \sqrt{(X_1 - X_2)^T \Sigma^{-1} (X_1 - X_2)},$$

где Σ — ковариационная матрица признаков из x. Если положить, что Σ — единичная матрица, то расстояние Махаланобиса становится простым расстоянием Евклида. Далее при помощи ρ_d вновь находятся пары похожих объектов (расстояние между которыми не превышает заданную константу c), и объекты распределяются по контрольной и тестовой выборкам.

Рассмотренные выше методы мэтчинга, а именно полное сходство и расстояние Махаланобиса, обладают существенным недостатком — вообще говоря, они оценивают $\mathbb{E}[Y_i(1) - Y_i(0)]$ только для маленького подмножества объектов, которые были собраны в пары. Это подмножество может получиться нерепрезентативным и не отражать всех особенностей той выборки, которую мы сформировали изначально. Поэтому в качестве метода балансировки в современных исследованиях используется другой подход.

6 Propensity score

В статье Rosenbaum and Rubin (1983), которая уже упоминалась выше, был показан интересный теоретический результат. Для того чтобы оценка τ была корректной, достаточно для каждого объекта X_i оценить его propensity score, а именно вероятность того, что к объекту X_i будет применён эффект:

$$e(X_i) = \mathbb{P}[W_i = 1 | X_i].$$

Далее объекты собираются в пары не по схожести векторов признаков X_i, а по схожести $e(X_i)$. Однако, как было показано в соответствующей работе, этого достаточно для того, чтобы получить сбалансированную выборку.

Классический алгоритм, примеряемый для оценки τ по propensity score, выглядит следующим образом:

1. Назначить каждый объект в контрольную ($W_i = 0$) или тестовую ($W_i = 1$) выборку.
2. По получившейся выборке оценить $\hat{e}(x)$, к примеру воспользовавшись логистической регрессией.

3. Каждому тестовому объекту из тестовой группы найти наиболее близкий по $\hat{e}(x)$ объект из контрольной выборки.

4. Оценить τ по полученным контрольным и тестовым группам.

В ходе работы данного алгоритма может оказаться, что есть какие-то тестовые объекты, для которых совсем нет ничего похожего среди контрольных объектов. Такие объекты можно просто отбросить.

Преимущества данного подхода состоят в том, что он позволяет оценивать схожесть методов по одному числу, а не по всему набору признаков. Однако для успешной оценки сходства требуется использовать весь набор признаков — игнорирование каких-то важных признаков может привести к смещению.

Сравнение объектов через одно число хоть и является вычислительно эффективным, но приводит к очевидным проблемам. К примеру, в статье King and Nielsen (2019) с говорящим названием "Why propensity scores should not be used for matching" приводится следующая иллюстрация, показывающая превосходство расстояния Махаланобиса перед propensity score.

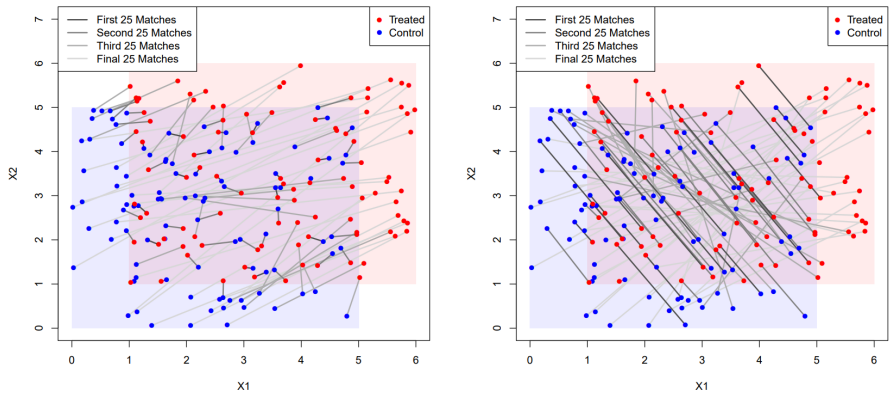

Рис. 1: Propensity score vs расстояние Махаланобиса

По данной картинке можно сделать следующий вывод: ближайшие 25 пар, выделенных расстоянием Махаланобиса, являются действительно близкими объектами — на рисунке они расположены недалеко. На картинке справа же видно, что propensity score может называть схожими объекты, расположенные на значительном расстоянии. На реальных данных это иногда приводит к серьёзным искажениям, в связи с чем данная оценка может оказаться смещённой.

7 Inverse propensity weighting

Алгоритм, описанный выше, страдает от необходимости прямого поиска близких объектов. Подход, названный inverse propensity weighting, является менее интуитивным, но он позволяет получить оценку τ без составления пар объектов.

Для оценки τ вновь используется propensity score, так что предварительно требуется разбить имеющуюся выборку на контрольную и тестовую части и оценить $\hat{e}(x)$. Далее оценка τ строится по следующей формуле:

$$\tau_{IPW} = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[\mathbb{E}[Y_i(1)|X_i] - \mathbb{E}[Y_i(0)|X_i]] =$$

$$= \mathbb{E}\left[\frac{\mathbb{E}[W_i|X_i]\mathbb{E}[Y_i(1)|X_i]}{e(X_i)} - \frac{\mathbb{E}[1 - W_i|X_i]\mathbb{E}[Y_i(0)|X_i]}{1 - e(X_i)}\right] =$$

$$\mathbb{E}\left[\frac{\mathbb{E}[W_i Y_i(1)|X_i]}{e(X_i)} - \frac{\mathbb{E}[(1 - W_i)Y_i(0)|X_i]}{1 - e(X_i)}\right] =$$

$$= \mathbb{E}\left[\frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i)Y_i}{1 - e(X_i)}\right].$$

Последнее равенство выполняется в силу того, что мы полагаем, что объекты назначаются в контроль и тест однозначно.

Интуитивная идея данного метода состоит во взвешивании наблюдений. Наблюдения, которые обладают большим значением $e(X_i)$, являются типичными для тестового множества и поэтому могут быть учтены с меньшим весом. Объекты, обладающие низким $e(X_i)$, являются чем-то уникальным или даже выбросами, поэтому их вес в итоговом значении увеличивается. Аналогично для контрольной группы и $1 - e(X_i)$.

8 Регрессионные подходы

Есть также ряд методов, которые оценивают влияние эффекта через построение регрессий.

8.1 Линейная регрессия

Предположим, что имеется множество объектов X с целевым значением Y и эффектом W. Для оценки влияния W можно построить следующую регрессию:

$$Y_i = \beta + \beta_1 W_i + \beta_2 X_i + \varepsilon_i,$$

где ε_i — нормальная ошибка. После построения данной регрессии и оценки значений коэффициентов β можно будет выяснить, что влияние W_i на значение Y_i равно $\beta_1 W_i$.

В статье Guo et al. (2021) было предложено улучшение данного подхода. Во-первых, вместо обычной линейной регрессии предлагается обучать обобщённую регрессию $g(x)$. Во-вторых, делается предположение, что влияние W_i также нелинейно и зависит от отклонения $g(x)$. Таким образом, для оценки влияния W_i нужно построить следующую модель:

$$Y_i = \beta + \beta_1 W_i + \beta_2 g(X_i) + \beta_3 W_i(g(X_i) - \bar{g}) + \varepsilon_i,$$

где $\bar{g} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$. Таким образом, W_i оказывает влияние, пропорциональное отклонению $g(X_i)$ от среднего предсказания.

9 Результаты

Из этого текста ты узнал:

- *как проводить A/B-тестирование, когда нельзя проводить A/B-тестирование;*
- *как оценить влияние эффекта на целевую переменную.*

Список литературы

Guo, Y., Coey, D., Konutgan, M., Li, W., Schoener, C., and Goldman, M. (2021). Machine learning for variance reduction in online experiments. *Advances in Neural Information Processing Systems*, 34:8637–8648.

King, G. and Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political analysis*, 27(4):435–454.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.