1 Интро

Заметки по ходу чтения книги Judea Pearl "Causality models, reasoning and inference".

Introduction to Probabilities, Graphs, and Causal Models

Какова вообще связь причинности и теории вероятностей? Есть две причины.

Первая состоит в том, что утверждения о причинах и следствиях обычно сопровождаются той или иной степенью уверенности. Часто причины не делают следствие абсолютно обязательным, а лишь повышают его вероятность.

Вторая (она на самом деле довольно сильно связана с первой) состоит в том, что даже весьма очевидные причинно-следственные связи выполняются не всегда, а *почти всегда*: существует множество мелких деталей, которые сложно учесть.

Рассмотрим факторизацию распределения $P(x_1, ...x_N) = \prod_n P(x_n|x_1...x_{n-1})$.

def Марковские родители случайной переменной X_n - минимальное подмножество переменных $PA_n \subset \{X_1...X_{n-1}\}$ такое, что $P(x_n|pa_n) = P(x_n|x_1..x_{n-1})$.

def Байесовская сеть - DAG, построенный с вершинами-переменными и ребрами, соединяющими вершину с её марковскими родителями (рёбра направлены от родителей к детям).

Можно показать, что при заданном упорядочивании переменных марковские родители для каждой переменной определяется однозначно, если распределение $P(X_1, ...X_N)$ строго положительно, то есть любая комбинация переменных имеет вероятность > 0 (понятное дело, если она не содержит значений переменных, маргинальная вероятность которых $= 0$). Понятно, что это будет достаточным условием, чтобы были определены условные вероятности $P(x_n|x_1..x_{n-1}) = \frac{P(x_1,...,x_n)}{P(x_1,..,x_{n-1})}$, так как в этом случае знаменатель не будет нигде обращаться в 0 на области определения $P(x_1, ..., x_N)$.

def Марковская согласованность (Markov Compatibility) - говорят что распределение P марковски согласованно с DAG G, если оно факторизуемо согласно графу, т.е. $P(x) = \prod_n P(x_n|pa_n)$.

Удобным способом характеризации распределений P, согласованных с G, является список независимостей, которые в этих распределениях должны быть. Эти независимости можно графически определить, используя критерий $d-$разделения (можно ознакомиться в Бишопе), но для полноты:

def d-разделение - говорят, что путь p в DAG G d-разделен/заблокирован множеством вершин Z если выполняется хотя бы одно из трёх условий:

1. Он содержит цепочку $a \to b \to c : \ b \in Z$
2. Он содержит вилку $b \to a, b \to c : \ b \in Z$
3. Он содержит $v-$структуру с вершиной, которая не в Z и все наследники которой тоже не в Z: $a \to b, c \to b, b \notin Z, de(b) \cap Z = \emptyset$

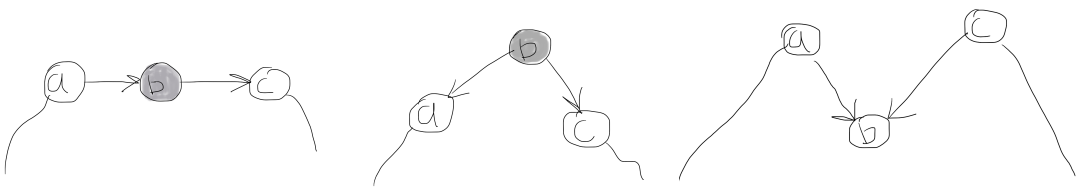


Рис. 1.1: Различные причины $d-$сепарации, заштрихованные вершины $\in Z$

Множество Z $d-$разделяет множества X и Y, если оно блокирует любой путь между X и Y.

Приложения d-разделения

А зачем собственно мы вводили $d-$разделение? А вот зачем:

Теорема 1.1 *Вероятностные следствия d−сепарации*
Если множества X и Y d-разделены множеством Z, то $X \perp\!\!\!\perp Y \mid Z$ в любом распределении, совместимом с G. Обратно, если X и Y не d-разделены Z в G, то существует как минимум одно распределение, согласованное с G: $X \not\perp\!\!\!\perp Y \mid Z$ в нем.

Пруф: Начнем с введения понятия отношения полуграфоида.

def Модель зависимостей - это тернарное отношение над множеством подмножеств 2^V некоторого множества V, тройки которого интерпретируются как утверждения о независимости первого и третьего элемента при условии, что известен второй.

def Полуграфоид (semi-graphoid) - это замыкание модели зависимостей относительно первых четырёх свойств (X, Y, Z, W - непересекающиеся подмножества множества-носителя V):

1. Симметрия: $I(X,Z,Y) \iff I(Y,Z,X)$
2. Декомпозиция: $I(X,Z,Y \cup W) \implies I(X,Z,Y)$ & $I(X,Z,W)$
3. Слабое объединение: $I(X,Z,Y \cup W) \implies I(X,Z \cup W,Y)$
4. Сокращение: $I(X,Z \cup Y,W)$ & $I(X,Z,Y) \implies I(X,Z,Y \cup W)$
5. Пересечение: $I(X,Z \cup Y,W)$ & $I(X,Z \cup W,Y) \implies I(X,Z,Y \cup W)$

Если кроме того полуграфоид замкнут относительно ещё пятого свойства, то он называется **графоидом**.

Примером полуграфоида (собственно, почему они нам в данном контексте интересны), заданным на множестве подмножеств случайных переменных V, будет отношение условной независимости: $I(X,Y,Z) \iff X \perp\!\!\!\perp Y \mid Z$. Если распределение к тому же является строго положительным, то есть для любого набора значений переменных $(x_1 ... x_N)$: $\forall i \in [1..N] \sum\limits_{X_j, j \neq i} P(x_1, ..., x_N) > 0 \implies P(x_1 .. X_N) > 0$, то отношение условной независимости будет графоидом. Почему важно это условие? Рассмотрим, когда будем пруфать свойство 5.

Давайте это докажем, чтобы просто поразминаться.

1. Весьма очевидно: действительно, если $P(X,Y|Z) = P(X|Z)P(Y|Z)$, то и симметричное верно, так как $P(Y,X|Z) = P(X,Y|Z) = P(X|Z)P(Y|Z) = P(Y|Z)P(X|Z)$.

2. Пусть $P(X, YW|Z) = P(X|Z)P(YW|Z)$. Тогда просто просуммируем правую и левую часть по множеству значений W:
Для левой части имеем $\sum\limits_w P(X, YW|Z) = P(X, Y|Z)$.
Для правой части аналогично

$$\sum_w P(X|Z)P(YW|Z) = P(X|Z) \sum_w P(YW|Z) = P(X|Z)P(Y|Z) \tag{1.1}$$

предпоследний переход в силу $Z \cap W = \emptyset$.

По условию левая и правая часть равны, значит $P(X,Y|Z) = P(X|Z)P(Y|Z)$.

3. Пусть $P(X, YW|Z) = P(X|Z)P(YW|Z)$. Тогда по свойству декомпозиции

$$P(X,W|Z) = P(X|Z)P(W|Z) \tag{1.2}$$

Запишем факторизацию

$$P(X,Y,Z,W) = P(X,YW|Z)P(Z) = P(X|Z)P(YW|Z)P(Z) = P(X|Z)P(Y|ZW)P(W|Z)P(Z) \tag{1.3}$$

$$P(X,Y|ZW) = \frac{P(X,Y,Z,W)}{P(Z,W)} = \frac{P(X|Z)P(Y|ZW)P(W|Z)P(Z)}{P(ZW)} = \frac{P(X,W|Z)P(Y|ZW)P(Z)}{P(ZW)}$$
$$= \frac{P(X,Z,W)P(Y|ZW)}{P(ZW)} = P(X|ZW)P(Y|ZW) \tag{1.4}$$

4. Пусть

$$P(X,W|ZY) = P(X|ZY)P(W|ZY) \tag{1.5}$$
$$P(X,Y|Z) = P(X|Z)P(Y|Z) \tag{1.6}$$

Рассмотрим $P(X,YW|Z)$:

$$P(X,YW|Z) = \frac{P(X,Y,Z,W)}{P(Z)} = \frac{P(X,W|ZY)P(ZY)}{P(Z)} = \frac{P(X|ZY)P(W|ZY)P(ZY)}{P(Z)}$$
$$= P(X|ZY)P(W|ZY)P(Y|Z) = P(X|ZY)P(Y,W|Z) = \frac{P(X,Y|Z)}{P(Y|Z)}P(Y,W|Z) \tag{1.7}$$
$$= \frac{P(X|Z)P(Y|Z)}{P(Y|Z)}P(Y,W|Z) = P(X|Z)P(YW|Z)$$

5. Пусть

$$P(X,W|Z,Y) = P(X|Z,Y)P(W|Z,Y) \tag{1.8}$$
$$P(X,Y|Z,W) = P(X|Z,W)P(Y|Z,W) \tag{1.9}$$

Умножим первое тождество на $P(Y)$, второе на $P(W)$:

$$P(X,W,Y|Z) = P(X|Z,Y)P(W|Z,Y)P(Y) = P(X|Z,Y)P(W,Y|Z) \tag{1.10}$$
$$P(X,Y,W|Z) = P(X|Z,W)P(Y|Z,W)P(Y) = P(X|Z,W)P(Y,W|Z) \tag{1.11}$$

Приравняв правые части, и используя свойство положительности (вот тут оно нужно), сократим на $P(Y,W|Z)$, получаем

$$P(X|Z,Y) = P(X|Z,W) \tag{1.12}$$

Видим, что правая часть не зависит от Y, значит и левая не должна зависеть: $P(X|Z,Y) = P(X|Z)$. Аналогично $P(X|Z,W) = P(X|Z)$. Нам же надо показать, что $P(X,Y,W|Z) = P(X|Z)P(Y,W|Z)$. можно заметить, что это следует из 1.10, если использовать независимость $X \perp\!\!\!\perp Y \mid Z$, выведенную ранее.

Ну в общем, вроде всё верно :) Более интересно, что эти свойства достаточны, чтобы определить все свойства вероятностной независимости.

Теперь перейдём к тому, как задать модель зависимостей на данном множестве. Понятно, что можно поступить наивным образом и задать её явно, перечислив список троек (X,Z,Y), для которых отношение независимости выполняется. Однако, этот список будет в общем случае расти экспоненциально с ростом размера множества-носителя, так как экспоненицально растёт число различных его подмножеств. Представление модели зависимостей в виде графа, в свою очередь, может быть интуитивно понятным, компактным, а также с графами можно эффективно работать.

Есть как водится два основных варианта: использовать **неориентированные** и **ориентированные** графы.

В случае **неориентированных** графов, интерпретация довольно простая: элементам множества ставятся в соответствие вершины, и множества вершин X и Y независимы при условии Z, если оно разделяет X и Y в обычном смысле теории графов, то есть если любой путь X в Y обязательно содержит хотя бы одну вершину из Z. Ну, тут стоит отметить, что вообще говоря далеко не любой полуграфоид в таком виде представим точно: в большинстве случаев в графе будут отсутствовать некоторые независимости. Например, если модель зависимостей над множеством из трёх элементов $V = \{x, y, z\}$ содержит единственную независимость $I(\{x\}, \{y\}, \emptyset)$, то никак соответствующий ей полуграфоид (заметим: в полуграфоиде будут две независимости в силу симметрии) не представить, не добавив лишних зависимостей, либо не убрав имеющиеся независимости.

На самом деле, множество полуграфоидов, которые точно задаются неориентированными графами - это замыкание намного более сильного класса свойств:

1. Симметрия: $I(X, Z, Y) \iff I(Y, Z, X)$
2. Декомпозиция: $I(X, Z, Y \cup W) \implies I(X, Z, Y)$ & $I(X, Z, W)$
3. **Сильное** объединение: $I(X, Z, Y) \implies I(X, Z \cup W, Y)$
4. Пересечение: $I(X, Z \cup Y, W)$ & $I(X, Z \cup W, Y) \implies I(X, Z, Y \cup W)$
5. Транзитивность: $I(X, Z, Y) \implies I(X, Z, W) \vee I(Y, Z, W)\ \forall\ W :\ W \cap (X \cup Y \cup Z) = \emptyset$

Ну то, что эти свойства верны для представлений в виде неориентированных графов, весьма понятно. Давайте докажем что отношение, замкнутое относительно этих свойств, является графоидом. По сути, три свойства графоидов совпадают в данном определении, так что вывести остаётся только два: слабое объединение и сокращение.

Начнём со слабого объединения: $I(X, Z, Y \cup W) \implies I(X, Z, Y) \implies I(X, Z \cup W, Y)$, где первый переход в силу свойства декомпозиции, второй - в силу свойства сильного объединения.

Докажем свойство сокращения: $I(X, Z, Y) \implies I(X, Z \cup W, Y)$ в силу сильного объединения, а значит $I(X, Z \cup Y, W)$ & $I(X, Z, Y) \implies I(X, Z \cup Y, W)$ & $I(X, Z \cup W, Y) \implies I(X, Z, Y \cup W)$, где последний переход сделан в силу свойства пересечения.

В общем понятно, неориентированные графы прикольные, но могут представить довольно ограниченное подмножество возможных моделей независимостей (тут и далее будем использовать этот термин как синоним полуграфоида, полагая, что модель зависимостей замкнута относительно свойств 1-4 полуграфоидов).

Вообще говоря, довольно часто нам не требуется идеальное представление модели зависимостей, а вполне достаточно разумного приближения, которое не будет содержать **все** независимости, определенные моделью, но по крайней мере не будет содержать лишние. Такое представление назовём *I-map* (от *independence*).

Перейдём к представлению модели зависимостей в виде ориентированных графов, или, точнее, DAG. интерпретация таких графов проста: ребро означает непосредственную причинную зависимость двух переменных. Увы, простые разрезы графа в данном случае уже не будут отражать независимость, так как обуславливание на какое-то общее следствие двух несвязанных событий может сделать их зависимыми. Поэтому, вместо обычного разделения графа вводится понятие d-разделения (мы о нём уже говорили).

def Хвостовая граница (*tail boundary*) переменной x - это подмножество B множества переменных L меньших x в смысле некоторого полного порядка на множестве переменных такое, что $I(x, B, L \backslash B)$.

def Протокол стратификации $L_\theta = (\theta, B(x))$ это пара из полного упорядочивания переменных θ, и функции $B(x)$ отображающей переменную на её хвостовую границу.

По протоколу стратификации однозначно строится DAG очевидным образом. Ясно, что для заданной модели зависимостей на n переменных существует $n!$ полных упорядочиваний. Для каждого полного упорядочивания в худшем случае существует $2^{n(n-1)/2}$ различных способов задать хвостовые границы (для каждой переменной все предыдущие в худшем случае могут как присутствовать в гра-

нице, так и нет $\implies$ для переменной номер i может оказаться 2^{i-1} различных функций, задающих хвостовую границу). Итого, может существовать до $n!2^{\frac{n(n-1)}{2}}$ разных протоколов стратификации для заданной модели зависимостей.

Утверждается, что если модель зависимостей обладает идеальным представлением в виде DAG (то есть существует такой DAG, в котором есть все независимости из модели, и только они, или что то же самое, он является и I-map и D-map одновременно), то один из протоколов стратификации его задаёт. Докажем это.

Рассмотрим граф D, идеально представляющий модель. Он задаёт частичный порядок на множестве переменных ϕ. Пусть θ - любо полный порядок, согласованный с ϕ. Тогда $L = (\theta, Par(x))$ будет определять протокол стратификации, генерирующий D (нетрудно увидеть, что непосредственные родители x являются хвостовой границей). ■

Если существует идеальное представление модели в виде DAG, то его можно найти, однако проверка на существование - это в общем случае сложная задача. Практически часто достаточно найти минимальный I-map, и следующая теорема покажет, что для любого полуграфоида (не обязательно имеющего идеальное представление в DAG) можно использовать стратификационные протоколы для построения I-map.

Теорема 1.2 *Если M - полуграфоид, и L_θ - любой его протокол стратификации, то DAG, сгенерированный по этому протоколу, будет I-map полуграфоида.*

Пруф по индукции по числу переменных в модели. Понятно, что для модели из одной переменной существует единственный DAG, и он конечно является I-map. Пусть теперь утверждение верно для моделей с числом переменных меньше k. Пусть M имеет k переменных, и имеется её протокол стратификации L_θ, последняя по порядку θ переменная n, $M - n$ - полуграфоид, полученный удалением всех отношений независимости, содержащих переменную n, $G - n$ - DAG с удалённой вершиной n и всеми инцидентными ей рёбрами. n- последняя переменная в упорядочивании, поэтому она не содержится ни в какой хвостовой границе из протокола L_θ, так что $L_\theta - n$ (это L_θ с удаленным правилом для переменной n) будет протоколом стратификации для $M - n$. Графом, который генерирует $L_\theta - n$, будет $G - n$, и по индукции он является I-map $M - n$.

Обозначим M_G модель зависимостей, построенную по G (то есть с использованием всех возможных d-разделений в графе), M_{G-n} - соответственно модель,, сгенерированная по $G - n$. Сайд-ноут: по идее M_G может содержать больше независимостей, чем есть $d-$разделений в G, так как оно строится как замыкание всех независимостей, полученных из G, но кажется нам бы доказать, что там нет лишних независимостей (об этому будет лемма ниже). По индукции, как сказано выше, $M_{G-n} \subset M - n$. Соответственно, нам надо показать, что $M_G \subset M$. Любая тройка $T \in M_G$ может быть отнесена к одной из четырёх непересекающихся категорий: либо n не представлено ни в одном из трёх множеств, составляющих T, либо n в каком-то из этих трёх множеств.

Лемма Пусть G - DAG, и M_G - модель зависимостей, индуцированная им. Тогда G - идеальное представление M_G в виде DAG. Заметим, G является I-map для M_G, так как все независимости из G по построению имеются в M_G. Значит, остаётся показать, что G - D-map. Для этого нужно доказать, что в G выполняются свойства 1-4 полуграфоидов (ведь тогда d-разделенные тройки замкнуты в G, а значит в M_G).

1. Свойство симметрии выполняется очевидно (если $X \perp\!\!\!\perp_G Y \mid Z \implies Y \perp\!\!\!\perp_G X \mid Z$)

2. Свойство декомпозиции в общем тоже очевидно верно: если Z блокирует пути между X и $Y \cup W$, то конечно Z блокирует пути между X и Y.

3. Свойство слабого объединения: пусть $X \perp\!\!\!\perp_G Y \cup W|Z$. Надо показать, что $X \perp\!\!\!\perp_G Y|Z \cup W$. Будем рассуждать от противного, пусть не так. Заметим, что по свойству декомпозиции, $X \perp\!\!\!\perp_G Y|Z$ и $X \perp\!\!\!\perp_G W|Z$. Значит, добавление к Z множества W разблокировало какой-то путь между X и Y. Но это возможно, только если разблокированный путь $X \rightsquigarrow Y$ имеет v-структуру с концом в W, то есть $X \rightsquigarrow ... \rightarrow w \leftarrow ... \rightsquigarrow Y$, где $w \in W$. Ясно, что при этом путь $X \rightsquigarrow w$ разблокирован. Рассмотрим аналогично этот путь (он будет короче предыдущего). Он либо был разблокирован до обуславливания на W, либо стал таким после. Во втором случае мы повторяем логику и откусываем опять префикс

пути, повторя подход пока не окажемся в первом случае. В первом же случае, у нас префикс пути до w не заблокирован при обуславливании на Z, но тогда $X \not\perp\!\!\!\perp_G w|Z$, что противоречит исходному предположению.

4. На десерт, свойство сокращения. Пусть $X \perp\!\!\!\perp_G W|Z \cup Y$ и $X \perp\!\!\!\perp_G |Z$. Нам надо показать, что $X \perp\!\!\!\perp_G Y \cup W|Z$.

Ну, начнём с того, что по условию, Z блокирует все пути $X \rightsquigarrow Y$. Значит, нам остаётся показать, что Z блокирует все пути $X \rightsquigarrow W$. Предположим, это не так. Тогда существует незаблокированный путь $p = X \rightsquigarrow W$. заметим, что по условию $Z \cup Y$ отделяет W от X, значит Y должно блокировать путь p, а значит, $p = X \rightsquigarrow ... \to y \to ...W$ или $p = X \rightsquigarrow ... \leftarrow y \to ... \rightsquigarrow W$, где $y \in Y$, причем префикс пути p вплоть до y не блокируется Z (иначе путь был бы заблокирован и без обуславливания на y). Но это в свою очередь означает, что существует незаблокированный путь от X до y, что противоречит тому, что Z d-разделяет X и Y. ■

В общем, теперь показано, что $(X, Z, Y) \in M_G \iff X \perp\!\!\!\perp_G Y|Z$, то есть отношение d-сепарации задаёт графоид на DAG.

Кейс 1: n не представлено в $T = (X, Z, Y)$. $T \in M_G \implies T \in M_{G-n}$, так как иначе в $G - n$ существует незаблокированный множеством Z путь, но тогда этот же незаблокированный путь есть и в G, так как добавление вершин и рёбер не может заблокировать путь. $G - n$ - I-map для $M - n$, значит $T \in M - n$, и так как $M - n \subset M$, то $T \in M$.

Кейс 2: $T = (Xn, Z, Y)$. Пусть $(n, B, R) \in L_\theta$ - последний триплет протокола стратификации (и конечно единственный, содержащий n), $B = B_X \cup B_Y \cup B_Z \cup B_0$, $R = R_X \cup R_Y \cup R_Z \cup R_0$, причём $X = B_X \cup R_X$, $Y = B_Y \cup R_Y$, $Z = B_Z \cup R_Z$ 1.2.

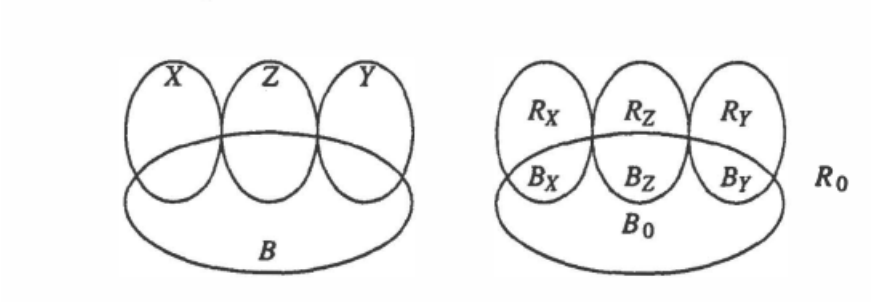


Рис. 1.2: Кейс 2

По построению, из всех вершин B есть ребро в n. Раз $T \in M_G$, то любой путь между n и Y должен быть заблокирован Z, поэтому $B_Y = \emptyset$, иначе был бы путь, состоящий просто из одного ребра, из $b \in B_Y$ в n. Таким образом, $Y = R_Y$, и последний триплет протокола представим в виде $(n, B_X B_0 B_Z, R_X R_Z R_0 Y)$.

Так как M полуграфоид,то по свойству слабого объединения мы можем перенести $R_X R_Z$ из третьего элемента триплета во второй и получить корректное отношение независимости: $(n, BRB_0, Y R_0) \in M$. Также, в силу декомпозиции, можем забить в последнем элементе тройки на R_0 и снова получить элемент M:

$$(n, XZB_0, Y) \in M \tag{1.13}$$

Все элементы B_0 соединены ребром с n, и n d-отделено от Y вершинами Z, значит B_0 тоже d-отделено от Y тем же Z, так как иначе существовал бы путь $B_0 \rightsquigarrow Y$, но тогда в силу того что из $B_0 \to n$, был бы незаблокированный Z путь $Y \rightsquigarrow n$. Теперь, раз X и B_0 d-разделены с Y через Z, то и их объединение тоже отделено от X через Z, так что $(XB_0, Z, Y) \in M_G$. В этой тройке не фигурирует n, значит по кейсу 1 имеем $(XB_0, Z, Y) \in M$. Объединяя это с 1.13 и используя свойство сокращения, получаем $(Y, Z, nXB_0) \in M$, а тогда по свойству декомпозиции $(nX, Z, Y) \in M$.

Кейс 3: $T = (X, nZ, Y)$. Опять представим последний элемент протокола стратификации в виде $(n, B_X B_Y B_Z B_0, R_X R_Y R_Z R_0 \in M$. Заметим, что $B_X = \emptyset \vee B_Y = \emptyset$, так как иначе есть путь $b_x \to n \leftarrow b_y$, разблокированный обуславливанием на n, то есть был бы незаблокированный путь $X \rightsquigarrow Y$, а это бы противоречило тому, что $T \in M_G$. Не умаляя общности, пусть $B_Y = \emptyset$. По соображениям из предыдущего пункта, $(B_0, Z, Y) \in M_G$.

Далее, $(X, nZ, Y) \in M_G \implies (X, Z, Y) \in M_G$, так как n имеет только входящие рёбра, а значит, если бы был незаблокированный Z путь $X \rightsquigarrow Y$, то обуславливание на n не помогло бы его заблокировать. Значит, $(XB_0, Z, Y) \in M_G$, и по кейсу 1 также $(XB_0, Z, Y) \in M$. По рассуждениям из кейса 2, последний триплет протокола представим в виде $(n, B_X B_0 B_Z, R_X R_Z R_0 Y) \in M$ и в итоге $(n, XZB_0, Y) \in M \implies (nXB_0, Z, Y) \in M \implies (XB_0, nZ, Y) \implies (X, nZ, Y)$ (слабое объединение, затем декомпозиция).

Кейс 4: $T = (X, Z, nY)$ - засчёт симметрии сводится к кейсу 2. ■

> По смыслу, как юзать эту теорему, то есть какие следствия? А вот какие: допустим есть модель зависимостей (любой полуграфоид), и мы построили для неё какой-то протокол стратификации L_θ, а по протоколу стратификации построили DAG. Так вот, тогда любое d-разделение множеств в DAG означает принадлежность соответствующей тройки (условную независимость) в модели зависимостей.

Ещё одно простое следствие: если протокол стратификации модели зависимостей M L_θ таков, что все хвостовые границы в нём минимальны (нельзя удалить элемент ни из одной с тем чтобы не нарушить принадлежность триплета M), то построенный по этому протоколу DAG является минимальным I-map модели - очевидно, он I-map по теореме, а минимальный, потому что каждое ребро определяется какой-то хвостовой границей протокола, значит никакое ребро нельзя удалить без нарушения I-map (иначе мы бы восстановили по усечённому DAG обратно урезанный протокол стратификации и он был бы корректен).

Возвращаясь к изначальной теореме, о следствиях d-разделения в контексте вероятностных распределений: мы показали, что если $P(V)$ марковское относительно DAG G то $X \perp\!\!\!\perp_G Y|Z \implies X \perp\!\!\!\perp Y|Z$. Ну, потому что в данном случае $P(V)$ выступает как модель зависимостей, а марковская согласованность позволяет построить протокол стратификации этой модели, соответствующий G. Но вот со вторым утверждением теоремы пока ничего не было сделано: что если в графе нет d-разделения множеств X, Y посредством множества Z, то существует распределение, согласованное с G, в котором $X \not\perp\!\!\!\perp Y|Z$. Пока оставим этот факт недоказанным.

Теорема 1.3 ***Условие упорядоченной марковости***
Необходимым и достаточным условием того, чтобы распределение P было марковски согласованным с G является то, что каждая переменная независима от всех своих предшественников в некотором упорядочивании, согласованном с G при обуславливании на её родителей в G.

Необходимость: пусть P совместимо с G. Это значит, что $P(x_1..x_n) = \prod P(x_i|pa_i)$. Упорядочим переменные, используя топологическую сортировку согласно графу G. Нам надо показать, что $P(x_i|pa_i, x_j) = P(x_i|pa_i)$.

Заметим, что

$$P(x_1...x_n) = \prod_{k \le i} P(x_k|pa_k) \prod_{k > i} P(x_k|pa_k) \tag{1.14}$$

Отсюда легко выводится частное распределение первых i переменных в выбранном нами порядке путём суммирования равенства

$$
\begin{aligned}
P(x_1...x_i) &= \prod_{k\le i} P(x_k|pa_k) \sum_{x_j:j>i} \prod_{k>i} P(x_k|pa_k) \\
&= \prod_{k\le i} P(x_k|pa_k) \sum_{x_j:i<j<n} \prod_{i<k<n} P(x_k|pa_k) \sum_{x_n} P(x_n|pa_n) \\
&= \ldots = \prod_{k\le i} P(x_k|pa_k)
\end{aligned}
\tag{1.15}
$$

Используя 1.15, легко выводим соотношение

$$
P(x_i|x_1,...,x_{i-1}) = \frac{P(x_1,...,x_i)}{P(x_1,...,x_{i-1})} = \frac{\prod\limits_{k\le i} P(x_k|pa_k)}{\prod\limits_{k\le i-1} P(x_k|pa_k)} = P(x_i|pa_i) \tag{1.16}
$$

Таким образом, $X_i \perp\!\!\!\perp \{x_1...x_{i-1}\}\backslash PA_i \mid PA_i$. Мы уже ранее доказывали, что отношение условной независимости для множества случайных переменных образует графоид, а потому по свойству декомпозиции $\forall j < i : X_j \notin PA_i \implies X_i \perp\!\!\!\perp X_j|PA_i$, что собственно и требовалось.

Достаточность: пусть каждая переменная X_i независима от всех переменных с меньшим номером при обуславливании на непосредственных предков переменной в графе G. Покажем, что в таком случае распределение марковски совместимо с G, то есть что оно факторизуется согласно G.

Любое распределение представим в виде следующей факторизации:

$$
P(x_1,...,x_n) = \prod_i P(x_i|x_1,...,x_{i-1}) \tag{1.17}
$$

Заметим, что теперь нам остаётся воспользоваться условием, что каждая переменная условно независима от всех переменных с меньшим номером при обуславливании на её родителей в графе G, чтобы получить необходимое равенство:

$$
P(x_1,...,x_n) = \prod_i P(x_i|pa_i) \tag{1.18}
$$

■

Ещё одна похожая теорема:

Теорема 1.4 ***Условие родительской марковости***
Необходимым и достаточным условием того, чтобы распределение P было марковски согласованным с G является то, что каждая переменная независима от всех своих не наследников в некотором упорядочивании, согласованном с G при обуславливании на её родителей в G.

Необходимость: пусть $P(v)$ совместимо с графом G, и $x_j \notin de(x_i)$. Покажем, что $x_i \perp\!\!\!\perp x_j \mid pa_i$. Заметим, что если $j < i$, то утверждение вытекает из предыдущей теоремы об условии упорядоченной марковости. Пусть $j > i$. Ясно, что в графе G не существует ни направленного пути $x_i \rightsquigarrow x_j$ (так как $x_j \notin de(x_i)$), ни направленного пути $x_j \rightsquigarrow x_i$, так как в таком случае упорядочение переменных было бы не согласовано с топологией графа. А значит, в любом пути, соединяющем x_i и x_j, существует как минимум одно соединение вида $\rightarrow x_k \leftarrow$, и конечно $x_k \notin pa_i$, ведь тогда в неё не могло бы быть двух стрелок в пути. Но любой такой путь заблокирован множеством pa_i, значит x_i и x_j d-разделены в G множеством pa_i, а значит, условно независимы при обуславливании на pa_i согласно вероятностным следствиям d-разделения.

Достаточность: тут всё просто, если каждая переменная независима от своих ненаследников, то она уж точно независима от всех переменных с меньшими, чем её, номерами, а значит по предыдущей теореме распределение согласовано с G. ■

def Наблюдаемая эквивалентность Два графа называют наблюдаемо эквивалентными, если любое распределение, согласованное с первым, согласовано со вторым, и наоборот.

Теорема 1.5 ***О наблюдаемой эквивалентности***
Два графа наблюдаемо эквивалентны тогда и только тогда, когда они имеют один и тот же скелет и набор $v-$структур.

Необходимость: Пусть два графа наблюдаемо эквивалентны. Докажем, что у них один и тот же скелет и набор v-структур.

Начнём со скелета. Пусть в графе G_1 имеется ребро $x - y$, а в G_2 нет. Рассмотрим такое распределение:

$$P(x_1, ..x_n) = uniform_{[0,1]}(x)bernoulli_x(y) \tag{1.19}$$

где все переменные, кроме двух рассматриваемых, константны. Заметим, что такое распределение будет согласованным с G_1, так как факторизуется согласно ему. В то же вермя, оно не согласованно с G_2. Действительно, пусть оно согласованно. Согласно теореме об упорядоченной марковости 1.3, в этом случае существует согласованное с G_2 упорядочивание, при котором каждая переменная независима от переменных с меньшими номерами при обуславливании на её родителей в G_2. Не умаляя общности будем считать, что x упорядочилась до y. Заметим, что $x \notin PA_y$, так как между ними нет ребра в G_2, а значит $P(y|x) = \sum\limits_{pa_y} P(y, pa_y|x) = \sum\limits_{pa_y} P(y|x, pa_y)P(pa_y|x) = \sum\limits_{pa_y} P(y|pa_y) = P(y)$, то есть y должно быть независимо от x в P, но это не так. Мы пришли к противоречию, значит P несовместимо с G_2, но тогда G_1 и G_2 наблюдаемо неэквивалентны, что противоречит исходному предположению, значит, в G_2 тоже нет ребра $x - y$. Значит, скелеты графов совпадают.

Докажем теперь, что в графах совпадают v-структуры. Ну, это довольно просто доказать: рассмотрим любые такие два графа, и предположим, что утверждение ложно. Мы уже знаем, что скелеты графов совпадают, значит совпадают и неориентированные пути. Раз набор v-структур не совпадает, существует как минимум один путь $x \rightsquigarrow y$, заблокированный Z в первом графе, и не заблокированный во втором (симметричный кейс аналогичен). Рассмотрим этот путь в обоих графах. Разница в блокировке может быть связана только с различной ориентацией стрелок в какой-то вершине v. Существует всего четыре способа ориентации стрелок в пути:

1. $... \rightarrow v \leftarrow ...$, но в таком случае v имеет одинаковые стрелки в обоих графах, так как является вершиной v-структуры, значит эта вершина не может различать блокировку пути в графах.
2. $\rightarrow v \rightarrow$
3. $\leftarrow v \leftarrow$
4. $\leftarrow v \rightarrow$

Последние три вариант могут наблюдаться в обоих графах независимо. Однако, при любой комбинации, если $v \in Z$, то в обоих графах v блокирует путь, а если $v \notin Z$, то не блокирует. То есть, в любом случае, любое допустимое направление стрелок в v не может различать заблокированность пути в графах, значит предположение неверно.

Достаточность: пусть у двух графов одинаковый скелет и набор v-структур. Нам надо показать, что они наблюдаемо эквивалентны.

Тут придётся немного попотеть. Доказывать будем по индукции по числу вершин в графах. С одной вершиной всё понятно верно, так что считаем, что база есть.

Пусть мы доказали утверждение для всех графов с число вершин, меньшим n. Рассмотрим два графа на n вершинах с одинаковым скелетом и набором v-структур. Предположим, что $P(v)$ согласованно с G_1, и покажем, что тогда оно согласовано и с G_2.

Раз $P(v)$ согласовано с G_1, то оно факторизуется согласно ему:

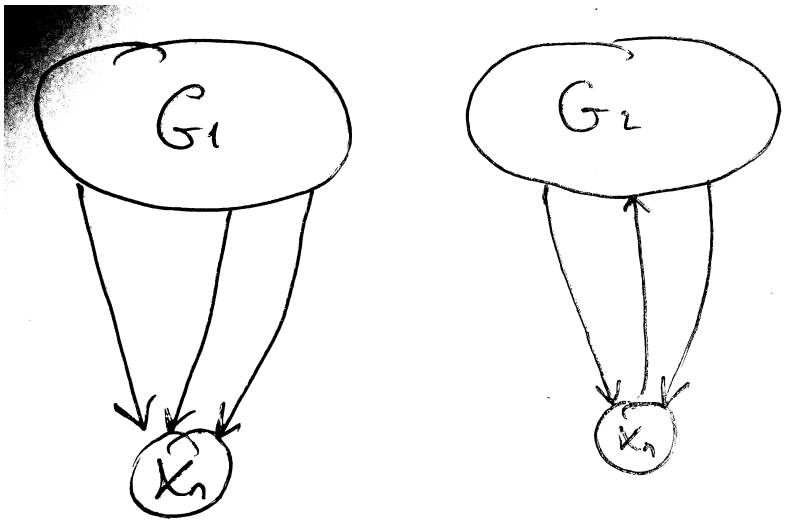


Рис. 1.3: Два рассматриваемых графа G_1 и G_2

$$P(x_1, ...x_n) = \prod_i P(x_i|pa_i^1) \tag{1.20}$$

Будем считать, что переменные упорядочены согласно топологии G_1, pa_i^1 - родители вершины X_i в G_1. Аналогично, далее будем обозначать pa_i^2 - родители вершины X_i в G_2. Понятно, что в общем случае $pa_i^1 \neq pa_i^2$.

Рассмотрим частное распределение $P(x_1, .., x_{n-1})$. Оно очевидно согласованно с $G_1 - x_n$. Далее, заметим, что оно согласовано и с $G_2 - x_n$, так как у этих двух графов на $n-1$ вершине одинаковые скелеты и v-структуры (скелеты понятно, а v-структуры - потому что мы не могли разрушить или добавить никакую v-структуру удалением x_n из графа G_2, так как в G_1 эта переменная не участвует в формировании каких-либо v-структур, а значит аналогично и в G_2). Поэтому мы можем записать:

$$P(x_1, ...x_{n-1}) = \prod_i P(x_i|pa_i^1) = \prod_i P(x_i|pa_i^2 \backslash \{x_n\}) \tag{1.21}$$

В последнем равенстве мы учитываем, что при удалении x_n из G_2 некоторые вершины (а именно все, в которые из x_n в G_2 ведёт ребро) из списка родителей вершины в графе $G_2 - x_n$ приходится убирать x_n.

Перейдём обратно к полному распределению:

$$P(x_1, ...x_n) = P(x_1, ...x_{n-1})P(x_n|pa_n^1) = \prod_i P(x_i|pa_i^2 \backslash \{x_n\})P(x_n|pa_n^1) \tag{1.22}$$

Разобьём мысленно множество pa_n^1 на три $pa_n^1 = Q \cup R \cup S$

Здесь Q - множество вершин, образующих v-структуру с x_n, то есть $x_i, x_j \in Q \iff x_i \to x_n \leftarrow x_j$ и нет ребра, соединяющего x_i с x_j. R - множество вершин, из которых в G_2 есть ребро в x_n. S - самое интересное множество, так как это множество вершин, ориентация рёбер между которыми и x_n в G_2 инвертирована относительно G_1, то есть это все вершины, для которых в G_2 вершина x_n является родителем. Для наглядности, можно посмотреть на картинку 1.4

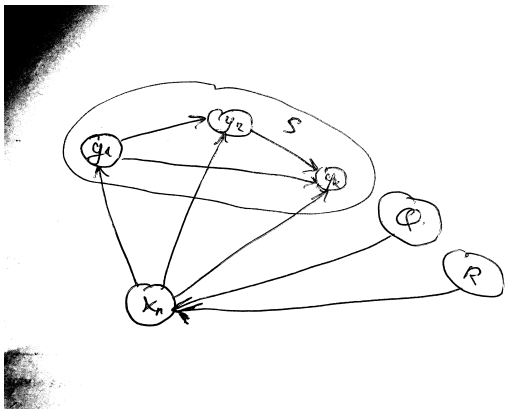


Рис. 1.4: Разбиение связанных с x_n вершин на Q, R, S

Таким образом, $pa_n^1 = Q \cup R \cup S$, $pa_n^2 = Q \cup R$. Остановимся на множестве S: надо понять, что эти вершины образуют клику, то есть все они связаны рёбрами между собой. Действительно, если бы это было не так, то существовали бы две вершины $x_i, x_j \in S$: из x_n есть в них ребро, но они не связаны между собой. Но тогда они в G_1 образовывали бы v-структуру с вершиной в x_n, но тогда они согласно нашему разбиению бы оказались не в S, а в Q. Тогда, мы можем представить факторизацию $P(v)$ следующим образом

$$P(x_1, ...x_n) = \prod_{x_i \notin S} P(x_i|pa_i^2) \prod_{x_i \in S} P(x_i|pa_i^2 \backslash \{x_n\}) P(x_n|Q, R) \tag{1.23}$$

Пронумеруем вершины в S в топологическом порядке в графе G_2. Ясно, что раз они все между собой связаны, это можно сделать единственным образом. Итак, пусть $S = \{y_1, y_2, ..., y_k\}$. Заметим, что $pa_{y_j}^2 \subset R \cup Q \cup \{y_1, ...y_{j-1}\} \cup \{x_n\}$. Ну действительно, $y_{j+1}, ...y_k$ в родителях нет в силу того, что мы топологически пронумеровали вершины. Никаких прочих вершин z там нет, так как если бы такая вершина была, то обязательно должно было бы присутствовать ребро $z \to x_n$, так как иначе $x_n \to y_j \leftarrow z$ образуют в G_2 v-структуру, которой нет в G_1, а значит $z \in Q \cup R$.

Нам надо показать, что $P(v) = \prod_i P(x_i|pa_i^2)$. Сравнивая с 1.23, приходим к тому, что надо показать

$$\prod_{x_i \in S} P(x_i|pa_i^2 \backslash \{x_n\}) P(x_n|Q, R, S) = \prod_{x_i \in S} P(x_i|pa_i^2) P(x_n|pa_n^2) \tag{1.24}$$

Начнём сворачивать формулу в левой части, начиная со множителя для y_k и x_n. Заметим, что Q, R не содержат наследников y_k в $G_2 - x_n$, иначе был бы цикл через x_n. Кроме того, $P(v \backslash \{x_n\})$ совместимо с $G_2 - x_n$ по предположению индукции. Значит, согласно теореме о родительском марковском условии, $P(y_k|pa_{y_k}^2 \backslash \{x_n\}) = P(y_k|y_1, ...y_{k-1}, Q, R)$.

Тогда

$$\begin{aligned} &p(y_k|pa_{y_k}^2 \backslash \{x_n\}) P(x_n|Q, R, S) = P(y_k|y_1, ...y_{k-1}, Q, R) P(x_n|y_1, ...y_k, Q, R) = P(x_n, y_k|y_1, ...y_{k-1}, Q, R) \\ &= P(y_k|x_n, y_1, ...y_{k-1}, Q, R) P(x_n|y_1, ...y_{k-1}, Q, R) = P(y_k|pa_{y_k}^2) P(x_n|y_1, ...y_{k-1}, Q, R) \end{aligned} \tag{1.25}$$

Далее мы можем действовать аналогично, сворачивая формулу для $y_{k-1} \dots y_1$. В результате получим ровно то, что нужно было показать в 1.24. ■

Таким образом, наблюдаемая эквивалентность определяет границы, в рамках которых возможно определение ориентаций в байесовской сети. Чтобы предпочесть одну эквивалентную байесовскую сеть

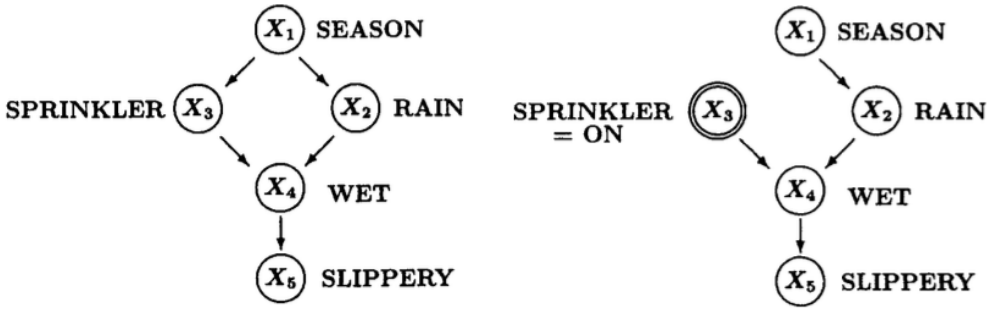
другой, нужна дополнительная информация об очерёдности событий, ну либо проводить эксперименты со вмешательством.

Причинные байесовские сети

Вообще говоря, до сего момента мы рассматривали байесовские сети как способ представления модели зависимостей, порождённой вероятностным распределением, соответственно рёбра в DAG означали всего лишь непосредственную зависимость между переменными. Однако, оказывается удобным наделять ориентацию рёбер дополнительно причинным смыслом, и, соответственно. строить байесовские сети по возможности так, чтобы направление рёбер отражало причинную связь переменных.

Плюсом такого подхода является модулярность - можно вносить интервенции с минимальными изменениями - по сути, затрагивая только рёбра, инцидентные одной конкретной вершине, если что-то меняется в поведении именно этой вершины. Профит в том, что каждое направленное ребро в графе отражает некий фундаментальный физический закон, который не влияет на другие физические законы в модели, и потому может быть потвикан независимо.

Рассмотрим пример небольшой байесовской сети:



(a) Небольшая причинная байесовская сеть (b) После интервенции Sprinkler := On

В ней все переменные бинарные (Да/Нет), кроме сезона, который понятно принимает одно из четырёх значений лето/осень/зима/весна. таким образом, для этой сети

$$P(x) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2,x_3)P(x_5|x_4) \tag{1.26}$$

Допустим теперь, что мы совершаем интервенцию путём принудительного включения поливалки. Ясно, что в этом случае нам надо удалить ребро $x_1 \to x_3$, так как теперь сезон никак не может влиять на поливалку - по сути, своим **действием** мы поменяли физический закон, определяющий x_3.

Тут скрывается как раз разница между наблюдением $X_3 = On$ и действием $do(X_3 = On)$: в первом случае мы просто обуславливаемся в исходной байесовской сети, во втором - в сети, которая меняется из-за интервенции. Когда мы наблюдаем, что поливалка включена, мы можем предполагать, что сейчас жаркий сезон, что наверно не было дождя и т.д. Когда мы сами включаем поливалку, понятное дело, все такие предположения становятся беспочвенными.

def Байесовская причинная сеть Пусть $P(v)$ - вероятностное распределение над переменными V, и $P_x(v)$ - распределение, которое получается в результате интервенции $do(X = x)$. Обозначим через P_* множество всех распределений с интервенциями $P_* = \{P_x(v)|X \subset V, x = instance(X)\}$. DAG G называется причинной байесовской сетью, согласованной с P_* тогда и только тогда, когда $\forall P_x \in P_*$ выполняется:
1. $P_x(v)$ марковское относительно G
2. $P_x(v_i) = 1$ $\forall V_i \in X$ если v_i консистентно с $X = x$
3. $P_x(v_i|pa_i) = P(v_i|pa_i)$ $\forall V_i \notin X$, если pa_i консистентно с $X = x$

Такие ограничения на множество интервенций P_* позволяют компактно представить его в виде одного графа G, и просто вычислять интервенционные распределения по правилу усечённой факторизации:

$$P_x(v) = \prod_{\{i|V_i \notin X\}} P(v_i|pa_i),\ \forall v, \text{консистентного с } x \tag{1.27}$$

Покажем, что это правило выводится из свойств байесовской причинной сети. Действительно, по свойству 1, $P_x(v) = \prod_i P(v_i|pa_i)$. $P_x(pa_i) = \sum_{v_i'} P_x(pa_i|v_i')P(v_i)$. Заметим далее, что если $V_i \in X$, то $P(v_i') = \delta(v_i' = v_i)$, и потому $P_x(pa_i) = P_x(pa_i|v_i)$ в этом случае, а значит PA_i независимо от V_i в распределении P_x. Это в свою очередь значит, что $P_x(v_i|pa_i) = P_x(v_i) = 1$, что и требовалось нам показать.

Нетрудно показать верность следующих двух свойств для G - байесовской причинной сети, согласованной с $P_*(v)$

Свойство 1: $\forall i\ P(v_i|pa_i) = P_{pa_i}(v_i)$

Ну действительно, $P(v_i|pa_i) = P_{pa_i}(v_i|pa_i) = \frac{P_{pa_i}(v_i,pa_i)}{P_{pa_i}(pa_i)} = P_{pa_i}(v_i)$, где первый переход сделан по условию 3 из определения байсовской причинной сети с $X = PA_i$, второй - так как $P_{pa_i}(v_i) = \sum_{PA_i} P_{pa_i}(v_i, PA_i) = P_{pa_i}(v_i, pa_i)$, так как P_{pa_i} принимает ненулевое значение только при $PA_i = pa_i$, ну и соответственно $P_{pa_i}(pa_i) = 1$.

По смыслу, это свойство означает что наблюдаемое распределение переменной v_i при условии, что мы пронаблюдали pa_i, совпадает с распределением v_i, когда мы внешним вмешательством сделали $PA_i = pa_i$.

Свойство 2: $\forall i,\ \forall S \subset V : S \cap (\{V_i\} \cup PA_i) = \emptyset \implies P_{pa_i,s}(v_i) = P_{pa_i}(v_i)$

По условию 3 мы имеем $P_{pa_i,s}(v_i|pa_i) = P(v_i|pa_i)$. Аналогично свойству 1, легко показать, что $P_{pa_i,s}(v_i) = P_{pa_i,s}(v_i, pa_i)$. Далее, $P_{pa_i,s}(v_i, pa_i) = P_{pa_i,s}(v_i|pa_i)P_{pa_i,s}(pa_i) = P_{pa_i,s}(v_i|pa_i)$, так как по условию 2 $P_{pa_i,s}(pa_i) = 1$. А значит, $P_{pa_i,s}(v_i) = P_{pa_i,s}(v_i|pa_i) = P(v_i|pa_i) = P_{pa_i}(v_i)$ (предпоследний переход по условию 3, последний - аналогично свойству 1).

По смыслу, это свойство описывает понятие инвариантности: как только мы контролируем PA_i - все непосредственные причины V_i - все прочие действия не влияют на поведение V_i.

Есть важное отличие между причинными связями, и вероятностным: причинные связи более "стабильны". Что это значит? По сути, причинные связи описывают механику внешнего мира, в то время как вероятностные - то, что мы знаем о мире/ во что верим. Таким образом, причинные связи не меняются при открытии нами новой информации, в этом плане они стабильны, по крайней мере до тех пор, пока физические законы (в широком смысле - законы, по которым работает окружающая среда) неизменны.

Вот пример по картинке с нашей простой байесовской причинной сетью: рассмотрим причинное отношение S_1 "Включение опрыскивателя не влияет на наличие дождя и вероятностное S_2 "Состояние распрыскивателя независимо от наличия дождя". Согласно картинке 1.5а, S_2 поменяется с False на True, если мы узнаем текущее время года, и обратно с True на False, если мы узнаем, мокрый асфальт, или нет. В свою очередь, узнавание любого из этих фактов не влияет на истинность S_1.

Функциональные причинные модели

Функциональная причинная модель состоит из набора уравнений вида

$$x_i = f_i(pa_i, u_i),\ i = 1..n \tag{1.28}$$

Здесь pa_i - множество переменных, непосредственно определяющих значение X_i, а U_i - помехи, связанные с ненаблюдаемыми переменными.

Сравним фичи, которыми обладают функциональные модели, с фичами причинных байесовских сетей. Рассмотрим три типа запросов, от менее требовательных к детализации знаний к более требовательным:

1. Предсказания - будет ли покрытие мокрым, если опрыскиватель выключен?

2. Интервенции - будет ли покрытие мокрым, если мы выключим опрыскиватель?

3. Контрфакты - было бы покрытие мокрым, если бы мы выключили опрыскиватель, если мы знаем, что покрытие скользкое и опрыскиватель включён?

Предсказания

По функциональной причинной модели построим диаграмму, проведя рёбра из PA_i в X_i, полученный граф G называют **причинной диаграммой**. Если G - DAG, то соответствующая модель называется *полумароквской*, и значения X однозначно определяются значениями U, $P(x)$ определяется через $P(u)$ однозначно. Если кроме того U взаимонезависимы, то модель называется *марковской*.

Теорема 1.6 ***Причинное марковское условие***
Любая марковская причинная модель M задаёт распределение $P(x_1...x_n)$ которое удовлетворяет родительскому марковскому условию относительно диаграммы G, построенной по M. То есть, $X_i \perp\!\!\!\perp Y|PA_i$ для любого Y не содержащего наследников X_i.

Действительно, PA_i, U_i однозначно определяют значение X_i, значит конечно $P(x_1...x_n, u_1..u_n)$ марковски совместимо с расширенным графом $G(X, U)$ (напомним, это просто значит, что распределение факторизуется согласно графу). Но тогда утверждение теоремы легко доказать, используя критерий d-разделения в расширенном графе и теорему о вероятностных следствиях d-разделения. ■

Свойство марковости причинной модели можно гарантировать, если договориться о двух правилах построения модели:

1. Если x_i и x_j зависимы, то либо одна из них является причиной другой (не обязательно непосредственной), либо существует третья переменная, являющаяся их общей причиной (нет корреляции без причины и нет взаимной причинности).

2. Если более, чем одна переменная является следствием данной переменной y, то данная переменная включена в модель, т.е. $y \notin U$ - иначе бы мы наблюдали необъяснимые моделью корреляции между x_i, x_j, для которых y - общая причина.

Второе предположение гарантирует, что $U_i \perp\!\!\!\perp U_j$. Первое - то, что нет циклических зависимостей, таким образом, мы получаем марковскую причинную модель.

В чем же плюсы использования функциональных моделей по сравнению с причинными байесовскими сетями?

Во-первых, функциональная спецификация обычно более краткая и содержит меньше параметров.

Во-вторых, в функциональных моделях проще рассуждать об условной независимости, потому что для людей естественно выводить такие предположения из предположении об отсутствии каких-то ненаблюдаемых общих причин. Человеку намного проще проверить, что все непосредственные причины события учтены, нежели проверять что переменная независима от своих ненаследников при условии известности непосредственных родителей.

Наконец, если что-то меняется в законах окружающей среды, проще учесть это в функциональном описании, нежели чисто вероятностном, так как в первом случае обычно нужно проапдейтить всего несколько уравнений.

Тем не менее, предсказание всё-таки является наиболее простой из трёх задач и решается в терминах условных вероятностей, без обязательной необходимости привлечения структурных уравнений или даже причинных байесовских сетей (обычные подойдут).

Интервенции

Интервенции легко описывать на языке функциональных моделей. Достаточно для всех переменных, поведение которых определяется действиями, убрать соответствующие уравнения, заменив их на $x_i =$

c_i.

Кроме такого удобства, есть и другие плюсы у функциональных моделей. Во-первых, большая гибкость: если отойти от марковских моделей, мы можем описывать и циклические зависимости, таким образом, отвечая на вопросы, связанные с политиками.

Во-вторых, интервенции, меняющие параметры в уравнениях более понятны в отличие от тех, что меняют условные вероятности, так как уравнения обычно рассматриваются как описание более-менее стабильных физических процессов, в отличие от условных вероятностей. Условные же вероятности мы обычно рассматриваем как то, что выводится из общего распределения, а не как то, что его генерирует.

Для изучения влияния интервенций обычных байесовских сетей уже недостаточно - нужны причинные байесовские сети, ну либо структурные уравнения.

Контрфакты

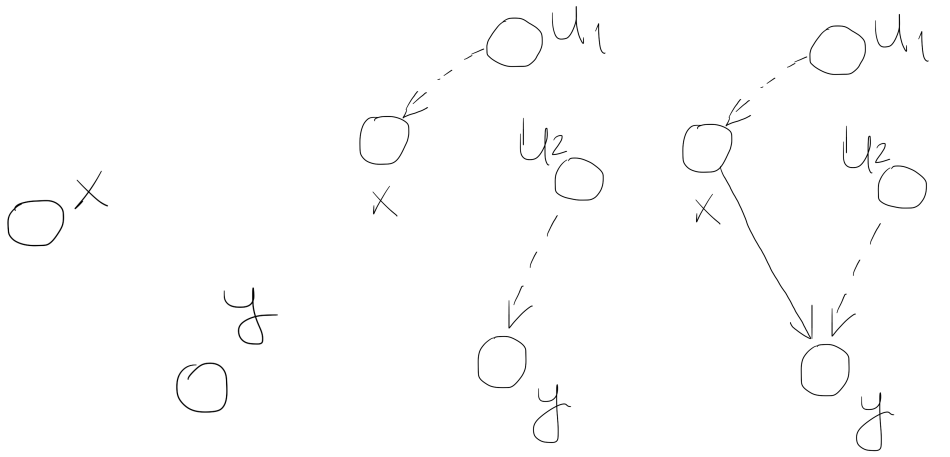
Контрфактические утверждения в принципе не могут быть определены в терминах стохастических причинных моделей (стохастические это то же что байесовские сети). Рассмотрим простой пример: пусть есть модель на двух переменных X - было ли дано человеку лечение или нет, и Y - выжил он в итоге или нет. Допустим, что некий индивидум Джо умер после получения лекарства, и мы хотим ответить на контрфактический вопрос "Какова вероятность того, что Джо выжил **бы**, если бы ему не дали лекарство?".

Понятно, что ответить на такой вопрос по имеющимся данным невозможно: действительно, Джо умер, и никакой информации о том, что с ним было, когда он не принимал лекарство, у нас нет. Перефразирование вопроса в терминах частоты аналогичного контрфактического события в популяции не помогает в данном случае, поэтому вообще многие стат аналитики воспринимали долгое время контрфактические вопросы как метафизические, на которые невозможно ответить непосредственным тестированием.

Тем не менее, в повседневной жизни мы постоянно оперируем контрфактическими утверждениями, что является поводом считать, что такие утверждения не лишены смысла и способны нести полезную информацию / быть выводимыми из какой-то информации о мире.

Пусть в нашем примере $P(y|x) = 0.5\ \forall x, y$. Тогда понятно, что $P(x, y) = 0.25$. Рассмотрим две функциональные модели, каждая из которых генерирует такое совместное распределение, но ведущие к различному значению искомой вероятности контрфакта Q - что субъект, умерший после лечения $(x = 1, y = 1)$, не умер бы $(y = 0)$, если бы его не лечили $(x = 0)$.

Байесовская сеть, описывающая данное распределение, имеет вид, как на 1.6а



(a) Байесовская сеть для описываемого распределения (b) Причинная диаграмма для модели 1 (c) Причинная диаграмма для модели 2

Модель 1

$$\begin{aligned} x &= u_1 \\ y &= u_2 \end{aligned} \tag{1.29}$$

Модель 2

$$\begin{aligned} x &= u_1 \\ y &= u_2 x + (1-x)(1-u_2) \end{aligned} \tag{1.30}$$

В первой модели, исход не зависит от лечения; во второй - исход для любого пациента зависит от лечения. Дело в том, что модель 2 описывает смесь двух подпопуляций: одной ($u_2 = 0$), в которой пациент выздоравливает только если принимает лекарство, во второй ($u_2 = 1$) пациент выздоравливает, только если лекарства ему не дают.

Эффект от лечения в этих двух моделях различен. В модели 1, $y = u_2$ и лечение не влияет на исход, так что замена X с 1 на 0 для тех, кто умер и лечился, не поменяет $y = 1 = u_2$ с 1 на 0, поэтому $Q = 0$. Во второй модели, в свою очередь, замена X с 1 на 0 для тех, кто умер (а это субпопуляция $u_2 = 1$), трагический исход поменяется на позитивный ($y = 1 \cdot x = 0$), то есть эффект отсутствия лечения будет $Q = 1$.

Как видим, стохастическая модель не позволяет в данном случае вычислять контрфактические вероятности: нужно знание процесса, который стоит за $P(y|x)$. С другой стороны, причинные структурные модели позволяют такие величины вычислять. Делали в данном случае мы это вычисление в три шага: сперва, мы применили имеющиеся под рукой данные $e : \{x = 1, y = 1\}$. Зная их, мы вывели, каковы совместимые значения U_1, U_2 - в данном случае только $u_1 = 1, u_2 = 1$. Вторым шагом, мы подставили $x = 1$ в структурные уравнения, игнорируя уравнение для X $x = u_1$. Наконец, мы решили уравнения относительно y и получили что $y \equiv 0$, то есть вероятность излечения в таком раскладе единична.

Этот подход обобщается для определения вероятности любого контрфакта $Y = y$ на данных e при условии $X = x$ по структурной модели в три шага.

1. **Отведение** - заменяем $P(u) \rightsquigarrow P(u|e)$
2. **Действие** - заменяем уравнения, определяющие X, на $X = x$
3. **Предсказание** - используем модифицированную модель для вычисления вероятности $Y = y$.

Если перекладывать этот метод на временные метафоры, то первый шаг - это объяснение прошлого с учётом имеющихся доказательств/данных; второй шаг - минимальное изменение курса истории чтобы согласовать мир с $X = x$; третий шаг - это предсказание будущего Y при условии нашего нового понимания прошлого и нашего альтернативного настоящего.

В целом, это самая сложная из трёх задач, и она уже в свою очередь не решается без привлечения механизма структурных уравнений.

Причинная и статистическая терминология

def Вероятностный параметр - любая величина, определяемая через совместное распределение переменных.

def Статистический параметр - любая величина. определяемая через совместное распределение наблюдаемых переменных, без каких либо предположений о существовании/несуществовании ненаблюдаемых переменных. Примерами являются условное матожидание $E(Y|x)$, коэффициент регрессии r_{XY}, значение плотности вероятности в $x = 0, y = 1$.

def Причинный параметр - любая величина, определяемая через причинную модель (которая задана в виде структурных уравнений) как в 1.28 и не являющаяся в то же время статистическим параметром. Примерами является параметр функции $f_i(pa_i, u_i)$, имеет ли X_9 влияние на X_3 при некотором u, ожидаемое значение Y при совершении интервенции $do(X = 0)$.

def Статистическое предположение - предположение о совместном распределении наблюдаемых переменных: например, что оно марково относительно некоторого DAG, или является многомерным нормальным.

def Причинное предположение - любое ограничение на причинную модель, которое не представимо статистическими предположениями. Например, что f_i линейно, или что $x_3 \notin pa_4$. Причинные предположения могут как иметь статистические следствия, так и нет. В первом случае говорят, что причинное предположение "проверяемо"или "фальсифицируемо". Часто, хоть и не всегда, причинные предположения фальсифицируемы через эксперимент, в этом случае говорят об "экспериментальной фальсифицируемости". Например, предположение, что X не имеет эффекта на $E[Y]$ в рассмотренной ранее модели два фальсифицируемо экспериментом, в то время как предположение "X может вылечить определенного субъекта популяции нет.

Два ментальных барьера к причинному анализу

Есть две сложности, с которыми мы сталкиваемся: во-первых, за каждым причинным высказыванием стоят причинные предположения, которые не выводятся из совместного распределения, и соответственно, они не проверяемы с помощью одних лишь наблюдаемых данных. Эти предположения предоставляются людьми, то есть приходится полагаться на какое-то экспертное мнение.

Во-вторых, для математического описания причинной теории нужна новая нотация: как мы выяснили, чисто вероятностного языка недостаточно. Например, на языке теории вероятностей (то есть через функции распределения) невозможно выразить причинную связь симптома и болезни: мы можем только описать их зависимость через $P(disease|symptom)$, но мы не можем отличить, причинная эта зависимость или статистическая.

2 Теория вывода причинности

Интуиция

Начнем с интуиции, которая стоит за причинно-следственными связями. Обычно необходимым условием является временная зависимость - причина происходит до следствия. Однако, очевидно, это далеко не всегда является достаточным условием для наличия причинной связи, поэтому остается вопрос, же ее установить?

Возможно ли в целом какое-то выявление причинно-следственных связей? На самом деле, да. Рассмотрим пример, где есть три события A, B, C и мы знаем, что A зависимо с B, B зависимо с C, но A и C независимы. В таком случае, если немного подумать, выходит, что наиболее простой граф, описывающий такую конфигурацию, выглядит как на 2.1

Фреймворк

Будем рассматривать задачу определения причинно-следственных связей в виде индукционной игры (индукционность в смысле что про некоторым примерам выводится какое-то общее правило), в которую ученый играет с природой (красивая формулировка конечно XD). предполагается, что у природы есть стабильные причинно-следственные механизмы, которые можно определить функциональными зависимостями между переменными, некоторые из которых впрочем ненаблюдаемы.

def Причинная структура (causal structure) множества переменных V - это DAG, в котором вершинам соответствуют переменные, а рёбрам - прямая функциональная зависимость между соответствующими переменными.

Причинная структура - это грубо говоря макет для **причинной модели** - точного определения того, как одни переменные влияют на другие.

def Причинная модель (causal model) - пара (D, Θ_D) из причинной структуры D и множества параметров Θ_D, ей соответствующих, то есть описывающих конкретные функциональные зависимости

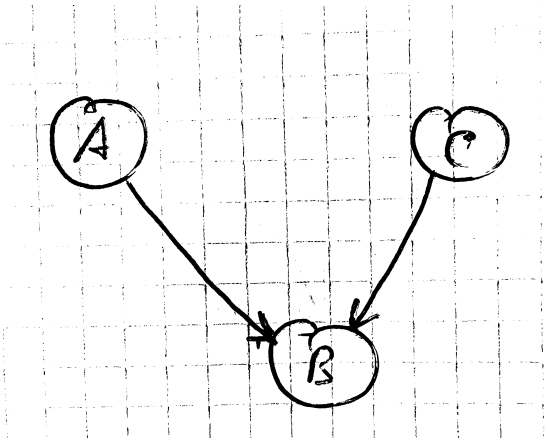


Рис. 2.1: A, C безусловно независимы, но зависимы при наблюдаемом следствии B

между переменными V в виде $x_i = f_i(pa_i, u_i)\ \forall x_i \in V$, где PA_i - родители x_i согласно D, U_i - случайный шум, вероятностное распределение над которым также определяется Θ_D, $U_i \perp\!\!\!\perp U_j$.

Шум, влияющий на значение переменных, можно рассматривать например как следствие ненаблюдаемости некоторых переменных. Такая модель, согласно данному ранее определению, является марковской, а потому по теореме 1.6 задаёт распределение, марковски согласованное с D.

Теперь задачу, поставленную перед гипотетическим учёным, можно сформулировать в виде восстановления причинной структуры, а затем и модели, при условии что он наблюдает лишь значения некоторого подмножества переменных $O \subset V$.

Выбор модели (бритва Оккама)

Вообще говоря, так как V неизвестно, можно придумать сколь угодно много разных моделей, которые смогу зафитить данное (эмпирически определённое) распределение $P(O)$, путём различного введение скрытых переменных. Например, можно ввести одну скрытую переменную U, которая будет причиной всех наблюдаемых переменных O, при этом никаких причинно-следственных связей между наблюдамемыми переменными в такой модели не будет, причинная структура для такой модели представлена на 2.2.

С другой стороны, считая $V = O$, но не имея никаких временных подсказок, учёный не может отбросить возможность того, что подлежащая структура - это полный DAG, где все вершины связаны со всеми в произвольном порядке - такая структура может имитировать поведение любой структуры.

Идея выбора модели состоит в том, чтобы в некотором смысле она была наиболее простой/минимальной относительно тех данных, которые наблюдаемы.

Дальше введем не совсем формальное пока-что определение выведенной причинности (пока полагаем, что все переменные наблюдаемые)

def Выведенная причинность (предв.) Переменная X имеет причинное влияние на переменную Y, если существует направленный путь из X в Y в любой минимальной **причинной** структуре, согласованной с данными.

def Скрытая структура (latent structure) это пара $L = (D, O)$, где D - причинная структура над V, $O \subset V$ - множество наблюдаемых переменных.

def Предпочтение структуры (structure preference) структура $L = (D, O)$ предпочтительнее структуры $L' = (D', O')$ (пишут $L \preceq L'$) если D' эквивалентно D на множестве наблюдаемых переменных O, т.е. тогда и только тогда, когда $\forall \Theta_D\ \exists \Theta_{D'} : P_{[O]}((D', \Theta_{D'})) = P_{[O]}((D, \Theta_D))$.

Латентные структуры называются эквивалентными, если $L \preceq L'$ и $L' \preceq L$.

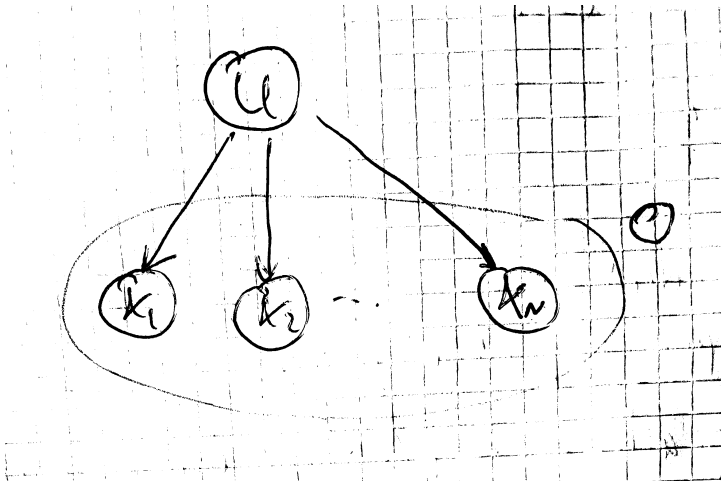


Рис. 2.2: Довольно бесполезная причинная структура

def Минимальность (minimality) структуры L относительно класса структур C означает её предпочтительность относительно всех других структур этого класса: $\forall L' \in C \; L \preceq L'$.

def Согласованность латентной структуры $L = (D, O)$ с распределением $\hat{P}$ над O означает возможность разместить $\hat{P}$ в данной латентной структуре, то есть что $\exists \Theta_D : P((O, \Theta_D)) = \hat{P}$.

def Выведенная причинность С данной $\hat{P}$ над O, переменная X имеет причинное влияние на переменную Y, если существует направленный путь из X в Y в любой минимальной **латентной** структуре.

Надо отметить, что экспрессивная мощность латентной структуры тем выше, чем меньше в ней закодировано независимостей между переменными: таким образом, структуры с меньшим числом независимостей, согласованные с данными, будут менее предпочтительны, чем структуры с большим числом независимостей в причинной структуре.

Стабильные распределения

Концепция минимальности латентной структуры позволяет корректно и непротиворечиво получать выводы о причинных связях переменных. Однако, это не всегда вычислительно просто - различных конфигураций структур может быть очень много, и проверять каждую из них на минимальность может быть очень дорого. К тому же, вообще говоря, может же оказаться, что настоящий процесс, генерировавший данные, все таки был порожден моделью, отличной от минимальной? Чтобы упростить себе жизнь, предлагается ввести в рассмотрение ещё один принцип, помимо минимальности - принцип *стабильности*.

Начнем с небольшого примера. Рассмотрим процесс, в котором есть две честные монетки. Множеством событий будет выпадение монетки A, выпадение монетки B, и событие C - "монетки выпали одинаковой стороной". нетрудно заметить, что любая пара переменных безусловно независима, но зависима при условии третьей переменной (например, $P(A = 1) = P(A = 1|B) = 0.5 \; \forall B$, но $P(A = 1|B = 1) = 0.5 \neq P(A = 1|C = 1, B = 1) = 1$). Таким образом, любая из структур на 2.3 допустима с точки зрения данных и является минимальной. В то же время, если чуть пошатать параметры распределения, например сделать $P(A = 1) = 0.6, P(A = 0) = 0.4$, то уже однозначно не подойдет структура, где C и B независимы безусловно, так как будет $P(C = 1) = 0.5$, но $P(C = 1|B = 1) = 0.6$. Аналогично можно пошатать вероятности для второй монетки, сделав её не совсем честной, и отбросить модель, где A и C независимы.

Для того, чтобы разрулить такие неоднозначности, вводится понятие стабильности:

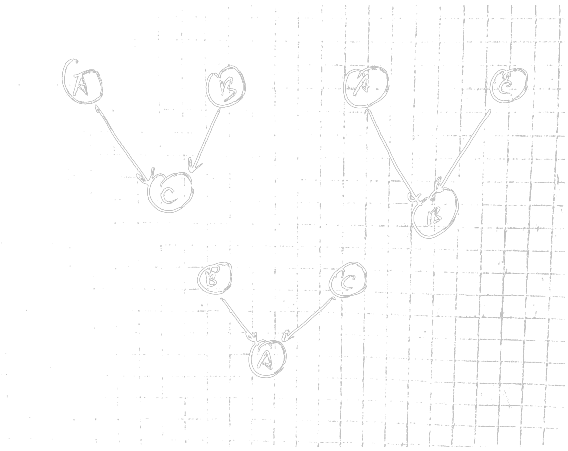


Рис. 2.3: Какую из трех причинных структур выбрать?

def Стабильность (распределения) Пусть $I(P)$ - множество всех независимых отношений переменных, заданных через P. Причинная модель $M = (D, \Theta_D)$ генерирует стабильное распределение тогда и только тогда когда в $P((D, \Theta_D))$ нет никаких лишних независимостей, т.е. $I(P((D, \Theta_D))) \subset I(P(D, \Theta_{D'}))\ \forall \Theta_{D'}$.

По смыслу, при варьировании параметров от Θ к Θ' никакие независимости не должны рушиться, если распределение стабильно. Что пока непонятно - а как выбирать, какие из вероятностей шатать: видимо те которые не ноль? Тогда и правда из стабильности остается только один вариант из трех в приведенном выше примере.

Стабильность означает, что распределение P имеет **идеальное** представление в виде DAG, то есть граф не просто I-map, но и D-map, и соответственно изоморфен распределению.

Реконструкция причинной структуры (DAG)

Когда все переменные наблюдаемы, если использовать принципы минимальности и стабильности, мы всегда будем получать единственную (с точностью до эквивалентности) причинную структуру (эквивалентные структуры - которые шарят одни и те же независимости, то есть один и тот же скелет и v-структуры).

Так как у подлежащей структуры мб эквивалентные, полученный DAG не будет однозначно определяться, поэтому лучшее, что можно сделать - определить его класс эквивалентности. Такой класс эквивалентности называют **шаблоном** (**pattern**), и он представляет из себя частично ориентированный граф (ориентируются только те рёбра, которые одинаково направлены во всех графах данного класса эквивалентности).

IC (Inductive Causation) Algorithm

Вход: $\hat{P}$ - стабильное распределение над переменными V.

Выход: $H(\hat{P})$ - шаблон, согласованный с $\hat{P}$.

Шаг 1: Строится неоринетированный граф D на вершинах V. Ребром соединяются любые две вершины a, b такие, что $\nexists S_{ab} \subset V \backslash \{a, b\} :\ a \perp\!\!\!\perp b | S_{ab}$

Шаг 2: $\forall a, b \in V : (a, c) \notin E$ перебираются их общие соседи $c : (a, c) \in E, (b, c) \in E$ и проверяется, $c \in S_{ab}$ или нет: если нет, рёбра a, c и b, c ориентируются в сторону c (то есть создаётся новая $v-$структура).

Шаг 3: Ориентируем оставшиеся рёбра, если это можно сделать однозначно при условии, что надо соблюсти условия

- ацикличности

- не добавления новых $v-$структур

Первый шаг отвечает за то, чтобы построить скелет графа - действительно, следует соединять только те вершины, которые не разделимы никаким множеством других вершин. Второй - за то, чтобы ориентировать все обязательные v-структуры - потому что раз $c \notin S_{ab}$, то $a - c - b$ должен быть заблокирован без обуславливания на c, а значит, $a \to c \leftarrow b$. Ну а третий шаг просто итеративно добавляет ориентации тем рёбрам, про которые можно однозначно вывести их направление после всех предыдущих манипуляций. При этом, на третьем шаге мы не добавляем новые v-структуры, так как все необходимые добавлены на шаге 2.

Реконструкция латентной структуры

Когда природа решает "спрятать"некоторое подмножество переменных, наблюдаемое распределение $\hat{P}$ может уже оказать нестабильным относительно наблюдаемого множества переменных $O \subset V$.

def Проекция латентной структуры $L_{[O]} = (D_{[O]}, O)$ для структуры L это латентная структура со следующими свойствами:

1. Любая ненаблюдаемая переменная в $D_{[O]}$ является общей причиной *ровно двух* наблюдаемых переменных

2. Для любого стабильного распределения P, генерируемого L, существует распределение P', генерируемое $L_{[O]}$, имеющее те же независимости, что и P: $I(P_{[O]}) = I(P'_{[O]})$.

Теорема 2.1 *У любой латентной структуры существует как минимум одна проекция.*

Оставим пока это утверждение без доказательства. ■

Проекции удобно представлять в виде двунаправленного графа, в котором вершины соответствуют только наблюдаемым переменным. Любое двунаправленное ребро в графе означает существование общей скрытой причины двух наблюдаемых переменных.

Можно показать, что имея проекцию произвольной минимальной латентной структуры, генерирующей $\hat{P}$, можно разметить рёбра проекции так, что по этой разметке можно проверять наличие причинного пути в любой минимальной модели для $\hat{P}$. Для построения проекции с последующей разметкой рёбер существует алгоритм IC^*, разбивающий рёбра на 4 типа:

1. Помеченная стрелка $a \xrightarrow{*} b$ - означает наличие направленного ребра из a в b в подлежащей модели.

2. Непомеченная стрелка $a \to b$ - означает либо наличие направленного ребра в подлежащей модели, либо существование общей скрытой причины $a \leftarrow L \to b$.

3. Непомеченная двунаправленная стрелка $a \leftrightarrow b$ означает наличие общей скрытой причины в подлежащей модели $a \leftarrow L \to b$

4. Неориентированное ребро $a - b$ означает что в подлежащей модели есть либо ребро $a \to b$, либо $a \leftarrow b$, либо общая причина $a \leftarrow L \to b$

IC* (Inductive Causation with Latent Variables) Algorithm

Вход: $\hat{P}$ - распределение над переменными O, стабильное относительно некоторой латентной структуры.

Выход: $core(\hat{P})$ - размеченный шаблон, согласованный с $\hat{P}$.

Шаг 1: Строится неориентированный граф D на вершинах O. Ребром соединяются любые две вершины a, b такие, что $\nexists S_{ab} \subset V \backslash \{a, b\} : \; a \perp\!\!\!\perp b | S_{ab}$

Шаг 2: Для любых двух не соединенных ребром вершин a, b с общим соседом c, проверяется, $c \in S_{ab}$. Если нет, то стрелки направляются в сторону c, образуя v-структуру: $a \to c \leftarrow b$, иначе ничего не делается.

Шаг 3: В полученном частично ориентированном графе, ориентируем все возможные концы рёбер, итеративно добавляя стрелки рёбрам по двум правилам:

1. Для любой пары a, b не соединенных ребром вершин, если $a \to c$ и нет стрелки из $b \to c$, то добавить стрелку $c \xrightarrow{*} b$

2. Если a, b смежны и существует направленный путь $a \xrightarrow{*} ... \xrightarrow{*} b$ состоящий из только из помеченных стрелок, то добавить $a \to b$.

Локальный критерий для вывода причинных связей

def Потенциальная причина переменная X является потенциальной причиной Y, если выполняются следующие условия:

1. $\nexists S : X \perp\!\!\!\perp Y|S$, то есть X, Y зависимы в любом контексте. Под контекстом понимается набор переменных с конкретными значениями.

2. Существует контекст S и переменная Z:

- $X \perp\!\!\!\perp Z|S$
- $Y \not\perp\!\!\!\perp Z|S$

Для лучшего понимания можно обратиться к рисунку 2.4: ясно, что Y *не может* быть причиной X, иначе будет существовать незаблокированный S путь $Z \rightsquigarrow Y \to X$, а потому либо X - истинная причина Y, либо они случайно ассоциированны из-за того, что существует общая ненаблюдаемая причина для них.

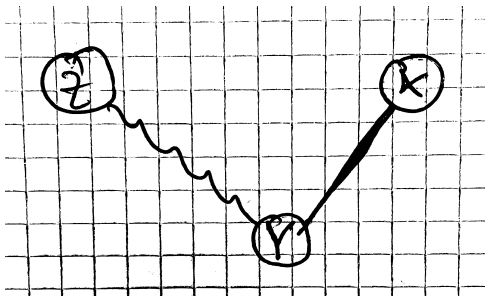


Рис. 2.4: X - потенциальная причина Y, Y - точно не может быть причиной X

def Подлинная причина Переменная X имеет подлинное причинное влияние на переменную Y если существует переменная Z, такая, что выполняется одно из двух условий:

1. X, Y зависимы в любом контексте, и существует контекст S:

* Z - потенциальная причина X в смысле предыдущего определения
* $Z \not\perp\!\!\!\perp Y|S$
* $Z \perp\!\!\!\perp Y|S \cup X$

2. X, Y находятся в транзитивном замыкании первого пункта.

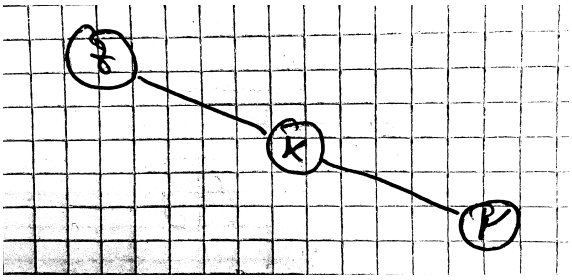


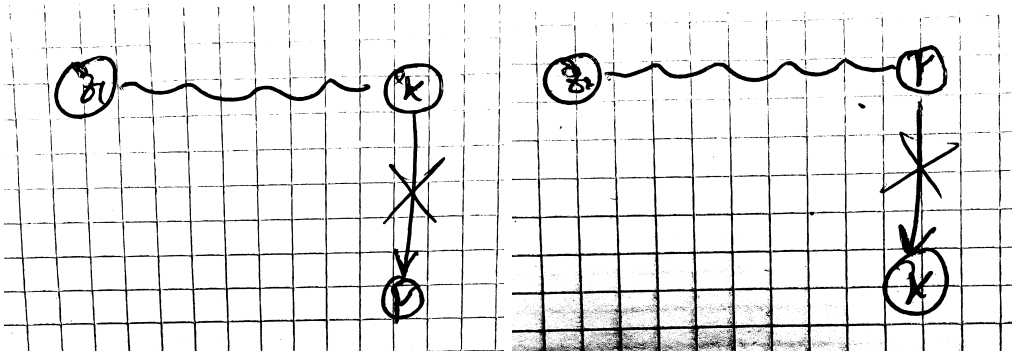
Рис. 2.5: X - истинная причина Y, если Z - потенциальная для X, и X блокирует путь из Z в Y

def Случайная ассоциация Две переменных X, Y случайно ассоциированы, если они зависимы в каком-то контексте и существуют две другие переменные Z_1, Z_2 и два контекста S_1, S_2:

1. $X \not\perp\!\!\!\perp Z_1|S_1$
2. $Y \perp\!\!\!\perp Z_1|S_1$
3. $X \perp\!\!\!\perp Z_2|S_2$

4. $Y \not\perp\!\!\!\perp Z_2 | S_2$

Условия 1 и 2 не позволяют назначить X причиной Y: и правда, раз существует незаблокированный S путь $Z_1 \rightsquigarrow X$, то не может оказаться, что X - причина Y, ведь иначе был бы незаблокированный путь из Z_1 до Y (2.6a). Аналогично условия 3,4 не позволяют назначить Y потенциальной причиной X (2.6b), что значит, что единственное объяснение их ассоциации - общая скрытая причина.



(a) X не может быть причиной Y (b) Y не может быть причиной X

Когда у нас появляется темпоральная информация, определение причинности упрощается: так, любое событие, которое предшествует другому, является его потенциальной причиной (ведь второе не может быть причиной первого, так как причина не может быть позднее следствия). Это приводит к более лаконичным определениям.

def Подлинная причина с темпоральной информацией Переменная X имеет подлинное причинное влияние на переменную Y, если существует контекст S и переменная Z, произошедшие до X:

1. $Y \not\perp\!\!\!\perp Z | S$
2. $Y \perp\!\!\!\perp Z | S \cup X$

По сути, это всё то же определение подлинной причины без темпоральной информации, просто теперь для определения потенциальной причинности достаточно того, что X случилось после Z.

def Случайная ассоциация с темпоральной информацией - две переменные случайно ассоциированы, если X произошла до Y, они зависимы в некотором контексте S, и существует переменная Z:

1. $Z \perp\!\!\!\perp Y | S$
2. $Z \not\perp\!\!\!\perp X | S$

3 Причинные диаграммы и установление причинных эффектов

В прошлой главе мы занимались тем, что изучали способы вывода причинных связей по сырым данным, без привлечения каких-либо дополнительных предположений. В этом же разделе мы будем изучать то, какие выводы можно сделать по комбинации данных и качественных причинных предположений, которые считаются приемлемыми в данном домене. Основной задачей будет выяснение, достаточно ли наблюдаемых данных (без необходимости проведения эксперимента) для выявления тех или иных причинных эффектов.

Причинные эффекты позволяют понять, как система реагирует на интервенции. В этой главе для описания интервенций будем использовать причинные диаграммы, и определять постинтервенционные распределения через преинтервенционные. Мы покажем, что эффект любой интервенции может быть расчитан по исключительно наблюдаемым данным при условии, что в причинной диаграмме нет ненаблюдаемых переменных и эта диаграмма - DAG.

Когда не все переменные наблюдаемы возникает вопрос устанавливаемости причинных связей и эффектов. В этом разделе будет рассмотрен фреймворк для непараметрического выявления таких связей.

Интервенции в марковских моделях

Графы как модели интервенций

Мы ранее установили, что причинные модели, в отличие от вероятностных, позволяют вычислять эффекты от интервенций. Для этого надо, чтобы совместное распределение P было снабжено также причинной диаграммой - DAG, определяющим причинные связи между переменными.

Самый простой тип интервенций - когда мы устанавливаем фиксированное значение x_i какой-то переменной X_i. Такая интервенция, называемая атомарной, состоит в том, что X_i перестаёт вести себя по закону, описанному в соответствующем функциональном уравнении $x_i = f_i(pa_i, u_i)$, а вместо этого ведёт себя по закону $x_i = x_i$ (тут слева x_i это не константа, а просто обозначение переменной, знак равенства не симметричный в функциональных уравнениях). Для обозначения такой интервенции используют запись $do(X_i = x_i)$ или просто $do(x_i)$. В полученной новой модели (с замененным механизмом поведения X_i), если вычислить функцию распределения для X_j, можно получить причинный эффект X_i на X_j, который обозначим $P(x_j|\hat{x}_i)$.

В общем случае, когда интервенции подвергается множество переменных путем присвоения им фиксированных значений, мы удаляем соответствующие этим переменным уравнения, и получаем новое распределение, описывающее постинтервенционный мир.

def Причинный эффект (causal effect) Пусть X, Y - два непересекающихся множества переменных. Причинным эффектом X на Y, обозначающимся $P(y|\hat{x})$ или $P(y|do(x))$, называется функция, отображающая X на пространство вероятностных распределений над Y. Для любой конкретной реализации $X = x$, $p(y|\hat{x})$ определяет вероятность $Y = y$ порожденную постинтервенционной причинной моделью.

Как мы уже показывали ранее в разделе про причинные байесовские сети, граф, соответствующий усеченному множеству уравнений - это подграф исходного графа, в котором удалены все входящие рёбра в переменные, над которыми проведена интервенция.

Интервенции как переменные

В некоторых случаях удобно рассматривать возмущение, описывающее интервенцию, тоже как переменную в графе: то есть смотреть на f_i как инстанс переменной F_i, и записывать структурное уравнение в виде

$$x_i = I(pa_i, f_i, u_i) \tag{3.1}$$

Здесь I - некоторая трёхместная функция, удовлетворяющая условию $I(a, b, c) \equiv f_i(a, c)$ если $b = f_i$.

Соответствующая такой интерпретации интервенции диаграмма содержит новое ребро по сравнению с оригинальным графом:

Вычисление эффекта интервенции

Рассмотрим атомарную интервенцию $do(X_i = x'_i)$. Она задаёт распределение $P(x_1, .., x_n|\hat{x}'_i) = \prod_{j \neq i} P(x_j|pa_j)$, когда $x_i = x'_i$, и равное 0 во всех остальных случаях. Разделим и умножим это равенство на $P(x_i|pa_i)$, и получим

$$P(x_1, ..., x_n|\hat{x}'_i) = \begin{cases} \frac{P(x_1,...,x_n)}{P(x_i|pa_i)} & x_i = x'_i \\ 0 & x_i \neq x'_i \end{cases} \tag{3.2}$$

На это равенство интересно посмотреть с точки зрения перераспределения масс: под действием интервенции, масса (вероятностная мера) точки $(x_1, ..., x_n)$ возрастает в $\frac{1}{P(x'_i|pa_i)}$. Таким образом, для

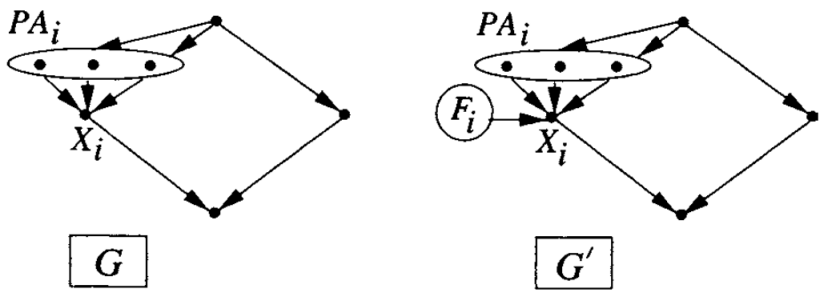


Рис. 3.1: Представление интервенции через расширенную диаграмму

тех точек, для которых условная вероятность $P(x_i'|pa_i)$ мала, существенно увеличивают свою вероятность при интервенции, в то время как точки, для которых значение $X_i = x_i'$ довольно естественно ($P(x_i'|pa_i) \approx 1$) не сильно поменяют свою массу.

В обычном байесовском подходе исключенные точки $x_i \neq x_i'$ переносят массу во все остальные точки посредством нормализационной константы $\frac{1}{P(x_i')}$. В нашем же случае происходит другая трансформация: масса точки распределяется между точками, шарящими с ней то же самое множество pa_i. Действительно: $P(pa_i|do(x_i')) = P(pa_i)$ - то есть полная масса, распределенная по pa_i, сохраняется, а значит масса каждой точки $X = (s_i, x_i, pa_i)$, где $s_i = instance(S)$, $S = V \backslash (PA_i \cup \{X_i\})$ распределилась как-то между точками с тем же pa_i и $x_i = x_i'$. Коэффициент, на который увеличивается масса каждой точки, для данного pa_i при этом один и тот же - $\frac{1}{P(x_i'|pa_i)}$.

Можно посмотреть на равенство 3.2 по-другому:

$$P(x_1, ..., x_n|\hat{x}_i') = \begin{cases} P(x_1, ..., x_n|x_i, pa_i)P(pa_i) & x_i = x_i' \\ 0 & x_i \neq x_i' \end{cases} \tag{3.3}$$

Теорема 3.1 ***Учёт непосредственных причин*** *Пусть PA_i - множество непосредственных причин переменной X_i и $Y \cap \{PA_i \cup X_i\} = \emptyset$. Тогда эффект интервенции $do(X_i = x_i')$ на Y вычисляется по формуле*

$$P(y|\hat{x}_i') = \sum_{pa_i} P(y|x_i', pa_i)P(pa_i) \tag{3.4}$$

Теорема следует непосредственно из суммирования формулы 3.3 по всем переменным кроме Y ■.

Операция, состоящая в таком обуславливании с последующим усреднением результата, взвешенным априорными вероятностями pa_i, называется "поправкой на PA_i".

В более общем случае, когда интервенция состоит в установке какого-то подмножества переменных S в константы $do(S = s)$, мы получаем аналогичную формулу для постинтервенционного распределения

$$P(x_1, ..., x_n|\hat{s}) = \begin{cases} \prod_{i|X_i \notin S} P(x_i|pa_i) & x_1, ..., x_n \text{ согласованное с } s \\ 0 & \text{иначе} \end{cases} \tag{3.5}$$

Можно пойти ещё дальше, и рассмотреть произвольную интервенцию, заменяющую некоторые причинные механизмы на новые. Например, если заменить механизм, определяющий X_i: $P(x_i|pa_i) \rightsquigarrow$

$P^*(x_i|pa_i^*)$ (может в том числе поменяться множество причин переменной X_i), то результирующее распределение после интервенции будет

$$P^*(x_1,..,x_n) = P(x_1,...,x_n)\frac{P^*(x_i|pa_i^*)}{P(x_i|pa_i)} \tag{3.6}$$

Выводы: имея причинную диаграмму, в которой все непосредственные причины переменных, для которых проводится интервенция, были наблюдаемы, мы можем определить постинтервенционное распределение через преинтервенционное, используя формулу усеченной факторизации 3.5.

Вообще говоря, интереснее ситуация, когда не все прямые причины наблюдаемы, что мешает непосредственному определению $P(x_i'|pa_i)$. Далее будут разработаны графические способы проверки того, можно ли оценить $P(x_j|\hat{x}_i')$, но для начала определим формально, что значит возможность оценить причинную величину Q по пассивным наблюдениям - возможность *идентификации.*

Идентификация причинных величин

Причинные величины, в отличие от статистических параметров, определяются относительно причинной модели M, а не относительно одного лишь совместного распределения $P_M(v)$ над множеством наблюдаемых переменных. Так как неэкспериментальные данные дают информацию только о $P_M(v)$, и так как вообще говоря зачастую несколько моделей могут генерировать одно и то же распределение, существует возможность, что интересующая величина не будет однозначно определяться по имеющимся данным, причем независимо от того, как много семплов есть. Идентифицируемость гарантирует, что добавленные предположения о причинной модели (например, причинная структура или нулевые коэффициенты в структурных уравнениях) предоставят необходимые дополнительные сведения без необходимости задания модели M целиком.

def Идентифицируемость Пусть $Q(M)$ - любая вычисляемая величина в модели M. Мы говорим, что $Q(M)$ идентифицируема в классе моделей $\mathbf{M}$, если для любых двух моделей из этого класса M_1 и M_2 оказывается $P_{M_1}(v) = P_{M_2}(v) \implies Q(M_1) = Q(M_2)$. Если наши наблюдения ограничены и мы знаем только некоторое множество фич F_M распределения $P_M(v)$, то мы говорим, что $Q(M)$ идентифицируемо по F_M если $F_{M_1} = F_{M_2} \implies Q(M_1) = Q(M_2)$.

Идентифицируемость необходима для объединения статистических данных с неполными причинными знаниями $\{f_i\}$, так как позволяет консистентно оценивать величины Q по большим семплам $P(v)$ без детального определения M: достаточно некоторых общих характеристик класса моделей $\mathbf{M}$. На данный момент, нас будет в качестве причинной величины Q интересовать причинный эффект $P_M(y|\hat{x})$, который конечно вычислим по заданной модели M (используя его определение), но его обычно нам как раз надо вычислить, не имея полной спецификации модели M.

Мы будем рассматривать следующий класс $\mathbf{M}$ причинных моделей, внутри которого будем определять идентифицируемость:

1. Модели шарят общую причинную структуру

2. Определяют положительные распределения на множестве наблюдаемых переменных v: $P(v) > 0$.

def Идентифицируемость причинного эффекта Причинный эффект X на Y идентифицируем из графа G, если величина $P(y|\hat{x})$ может быть однозначно вычислена из любого положительного распределения наблюдаемых переменных, то есть если $P_{M_1}(y|\hat{x}) = P_{M_2}(y|\hat{x})$ для любой пары моделей M_1, M_2, для которых $P_{M_1}(v) = P_{M_2}(v)$ и графы которых совпадают.

Идентифицируемость $P(y|\hat{x})$ гарантирует, что причинный эффект $do(X = x)$ на Y можно вычислить по двум источникам данных:

1. Пассивным наблюдениям, позволяющим восстановить $P(v)$

2. Причинному графу G, который определяет (качественно), какие переменные задают стабильные причинные механизмы (какие переменные определяют другие переменные).

Положительность распределения гарантирует то, что в формуле 3.2 мы не будем делить на 0: логично, если бы вероятность наблюдать x_i' в контексте pa_i была нулевой, мы бы не смогли вывести эффект от действия $do(X_i = x_i')$.

Отметим, чтобы установить неидентифицируемость, достаточно предоставить в качестве примера два множества структурных уравнений, задающих одинаковые распределения над наблюдаемыми переменными, но имеющих различные причинные эффекты.

Теорема 3.2 *При заданной причинной диаграмме G любой марковской модели, в которой некоторое подмножество V переменных наблюдаемо, причинный эффект $P(y|\hat{x})$ идентифицируем если $\{X \cup Y \cup PA_x\} \subset V$. Выражение для $P(y|\hat{x})$ в этом случае определяется путем учёта непосредственных причин по формуле 3.4.*

Контроль путающего смещения (confounding)

Когда мы хотим оценить эффект какого-то фактора X на другой Y, возникает вопрос нужно ли нам стандартизировать измерения на различные значения прочих факторов Z, также известных как ковариации/конфаундеры. Стандартизация состоит в разбиении популяции на подгруппы, которые гомогенны относительно значения Z, затем оценивании эффекта в каждой подгруппе и усреднение результата.

Иллюзорная природа стандартизации была обнаружена еще пирсоном в 1899, когда он обнаружил парадокс Симпсона: любая статистическая связь между двумя переменными может быть инвертирована путем включения дополнительных факторов в анализ. Соответственно встаёт вопрос - какие переменные нам следует учитывать во избежание конфаундинга?

Критерий задней двери

def Backdoor criterion Множество переменных Z удовлетворяет критерию задней двери относительно упорядоченной пары переменных (X_i, X_j) в DAG G если

1. Никакая вершина в Z не является наследником X_i
2. Z блокирует любой путь между X_i и X_j, в котором есть ребро, направленное в X_i

В общем случае, когда есть два множества переменных X, Y, говорят что Z удовлетворяет критерию задней двери, если оно удовлетворяет ему для любой упорядоченной пары $(X_i, X_j) : X_i \in X, X_j \in Y$

Теорема 3.3 ***Поправка задней двери*** *Если множество переменных Z удовлетворяет критерию задней двери относительно (X, Y), то причинный эффект идентифицируем и задаётся формулой*

$$P(y|\hat{x}) = \sum_z P(y|x,z)P(z) \tag{3.7}$$

Докажем теорему. Для доказательства удобнее всего рассматривать расширенный граф, в котором представлены вершины, задающие причинные механизмы. Раз все заднедверные пути заблокированы Z, то любой путь, $F_x \rightsquigarrow Y$ идет через детей X, но эти пути заблокированы при обуславливании на X. В итоге, мы приходим к тому, что Y независимо от F_x при условии X, Z:

$$P(y|x,z,F_x = do(x)) = P(y|x,z,F_x = idle) = P(y|x,z) \tag{3.8}$$

что означает, что наблюдение $X = x$ неотличимо в контексте Z от интервенции $do(X = x)$.

Действительно, если P' - распределение, порожденное расширенным графом G', то

$$P(y|\hat{x}) = P'(y|F_x) = \sum_z P'(y|z,F_x)P'(z|F_x) = \sum_z P'(y|z,x,F_x)P'(z|F_x) \tag{3.9}$$

Последнее равенство верно, так как $F_x \implies X = x$. Теперь нам остается избавиться от двух F_x в правой части равенства. Заметим, что F_x - это корневые вершины в G', с детьми X. Поэтому,

$F_x \perp\!\!\!\perp Z|X$, так как они d-отделены, ведь Z - не наследники X по первому пункту бэкдор критерия. Поэтому

$$P'(z|F_x) = P'(z) = P(z) \tag{3.10}$$

Что касается $P'(y|z, x, F_x)$, то согласно 3.8 оно эквивалентно $P(y|x, z)$. В итоге, мы получаем ровно то, что и требовалось. ■

Фронтальный критерий

Рассмотрим следующую диаграмму:

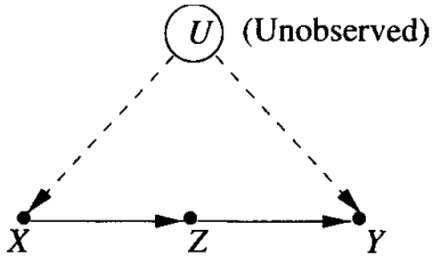


Рис. 3.2: Диаграмма для представления фронтального критерия

Соответствующее диаграмме распределение факторизуется согласно

$$P(x, z, y, u) = P(u)P(x|u)P(z|x)P(y|z, u) \tag{3.11}$$

Интервенция $X = do(x)$ приводит нас к факторизации без множителя для X:

$$P(z, y, u|\hat{x}) = P(u)P(z|x)P(y|z, u) \tag{3.12}$$

Просуммируем по u, z для вычисления эффекта на y:

$$P(y|\hat{x}) = \sum_z P(z|x) \sum_u P(u)P(y|z, u) \tag{3.13}$$

Этим просто так не воспользоваться, так как у нас u ненаблюдаемая - от неё надо как-то избавиться. Заметим, что по d-разделению в графе имеем следующие независимости:

$$P(u|z, x) = P(u|x) \tag{3.14}$$
$$P(y|z, u, x) = P(y|u, z) \tag{3.15}$$

Используя эти равенства, рассмотрим компонент 3.13

$$\begin{aligned}\sum_u P(u)P(y|z,u) &= \sum_x \sum_u P(u|x)P(x)P(y|z,u) \\ &= \sum_x \sum_u P(u|x,z)P(x)P(y|z,u,x) \\ &= \sum_x \sum_u P(y,u|z,x)P(x) \\ &= \sum_x P(y|z,x)P(x)\end{aligned} \tag{3.16}$$

что позволяет нам преобразовать 3.13 в

$$P(y|\hat{x}) = \sum_z P(z|x) \sum_x P(y|z,x)P(x) \tag{3.17}$$

В данной формуле все множители в правой части могут быть оценены по наблюдаемым данным, из чего следует, что $P(y|\hat{x})$ так же можно оценить. Таким образом, при условии что Z удовлетворяет 3.14, 3.15 мы можем несмещенно оценить эффект X на Y.

Формулу 3.17 можно интерпретировать как двухшаговое применение бэкдор-выравнивания: первым шагом мы вычисляем эффект X на Z: так как нет ни одного разблокированного бэкдор-пути из Z в X, то

$$P(z|\hat{x}) = P(z|x) \tag{3.18}$$

Затем, мы вычисляем эффект Z на Y. Тут у нас уже есть незаблокированный бэкдор-путь $Y \leftarrow U \rightarrow X \rightarrow Z$, который однако можно заблокировать, обусловившись на X, так что X удовлетворяет в данном случае бэкдор-критерию и потому

$$P(y|\hat{z}) = \sum_{x'} P(y|z,x')P(x') \tag{3.19}$$

В итоге, мы комбинируем два причинных эффекта чтобы получить

$$P(y|\hat{x}) = \sum_z P(z|\hat{x})P(y|\hat{z}) \tag{3.20}$$

def Фронтальный критерий Множество переменных Z удовлетворяет фронтальному критерию относительно упорядоченной пары переменных (X,Y), если

1. Z блокирует все направленные пути $X \rightsquigarrow Y$
2. Нету незаблокированного бэкдор-пути $Z \rightsquigarrow X$
3. Все бэкдор-пути $Y \rightsquigarrow Z$ заблокированы X.

Теорема 3.4 ***Фронтальная корректировка*** *Если Z удовлетворяет фронтальному критерию относительно (X,Y) и $P(x,z) > 0$, то причинный эффект X на Y идентифицируем и может быть вычислен как*

$$P(y|\hat{x}) = \sum_z P(z|x) \sum_{x'} P(y|z,x')P(x') \tag{3.21}$$

Исчисление интервенций

Нотация

Пусть X, Y, Z - какие-то непересекающиеся множества переменных. Обозначим через $G_{\overline{X}}$ граф, полученный удалением из G всех рёбер, указывающих на X. Аналогично, через$G_{\underline{X}}$ обозначим граф, полученный из G удалением всех рёбер, исходящих из X. Выражение $P(y|\hat{x}, z) \equiv \frac{P(y,z|\hat{x})}{P(z|\hat{x})}$ описывает вероятность $Y = y$ при условии, что мы установили $X = x$ и пронаблюдали при этом $Z = z$.

Правила вывода

Теорема 3.5 ***Правила do-исчисления*** *Пусть G - DAG с заданной причинной моделью и P - распределение, индуцированное ею. Для любых непересекающихся множеств X, Y, Z, W переменных справедливы правила:*

1. *Вставка/удаление наблюдений*

$$P(y|\hat{x}, z, w) = P(y|\hat{x}, w) \text{ если } Y \perp\!\!\!\perp_{G_{\overline{X}}} Z|X, W \tag{3.22}$$

2. *Заменяемость наблюдений/действий*

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, z, w) \text{ если } Y \perp\!\!\!\perp_{G_{\overline{X}\underline{Z}}} Z|X, W \tag{3.23}$$

3. *Вставка/удаление действий*

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, w) \text{ если } Y \perp\!\!\!\perp_{G_{\overline{X}\overline{Z(W)}}} Z|X, W \tag{3.24}$$

Где $Z(W) \subset Z : de(Z(W)) \cap W = \emptyset$ в $G_{\overline{X}}$

Первое правило утверждает, что $d-$сепарация является валидным тестом на независимость в постинтервенционном распределении: действительно, удаление структурных уравнений в результате интервенции не приводит к появлению новых зависимостей.

Второе правило описывает условие, при котором внешнее наблюдение z совпадает с интервенцией: для того, чтобы это выполнялось, нужно соблюдение бэкдор критерия: любой бэкдор-путь из Y в Z должен быть заблокирован, и это ровно то, что является условием применения этого правила: в $G_{\overline{X}\underline{Z}}$ существуют только бэкдор-пути из Y в Z (так как все рёбра исходящие из Z в этом графе удалены по определению), и раз в таком графе $Y \perp\!\!\!\perp Z|W, X$, то W, X блокируют все бэкдор-пути.

Третье правило - самое коварное из трех. Для его доказательства рассмотрим расширенный граф G', в котором добавлены вершины F_z и рёбра $F_z \to Z$.

Рассмотрим два случая - когда $Z_i \in Z(W)$ и когда $Z_i \notin Z(W)$.

В первом случае, имеем условно картинку 3.3. В $G_{\overline{X},\overline{Z(W}}$ фронтальные пути $Z_i \rightsquigarrow Y$ совпадают с путями в графе $G_{\overline{X}}$, и раз в первом графе они заблокированы W, X, то и во втором (исходном) они тоже заблокированы. Это значит, что любой путь $F_{Z_i} \rightsquigarrow Y$ проходящий по ребру из Z_i заблокирован. Пути же из F_{Z_i}, которые являются бэкдор-путями относительно Z_i (содержат ребро, ведущее в Z_i), заблокированы самим Z_i, так как образуют v-структуру.

Во втором случае, имеем картинку 3.4. В графе $G_{\overline{X},\overline{Z(W)}}$ никакие инцидентные Z_i рёбра не удалены, потому из того, что в манипулированном графе Z и Y d-разделены через X, W следует, что и в исходном они d-разделены, а значит и F_z тоже d-отделено от Y, что и требовалось показать ■.

Почему мы в данном случае рассматриваем неманипулированный граф? На самом деле, потому, что из утверждения про манипулированный граф в данном случае не вывести утверждение для исходного графа: путь $F_z - \to Z_i \leftarrow U \to Y$ окажется разблокированным, несмотря на то, что мы не обуславливаемся на Z_i, потому что мы тем не менее обуславливаемся на наследников Z_i.

Следствие: если путем применения конечного числа правил 1-3 к причинному эффекту $P(y|\hat{x}_1, ..., \hat{x}_m)$ можно получить формулу без do-операторов, содержащую только наблюдаемые величины, то причинный эффект идентифицируем.

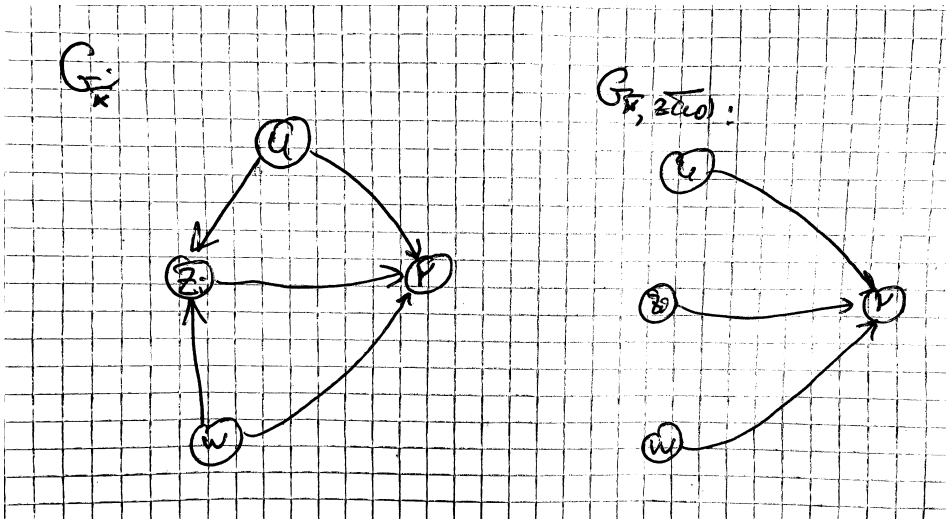


Рис. 3.3: $Z_i \in Z(W)$

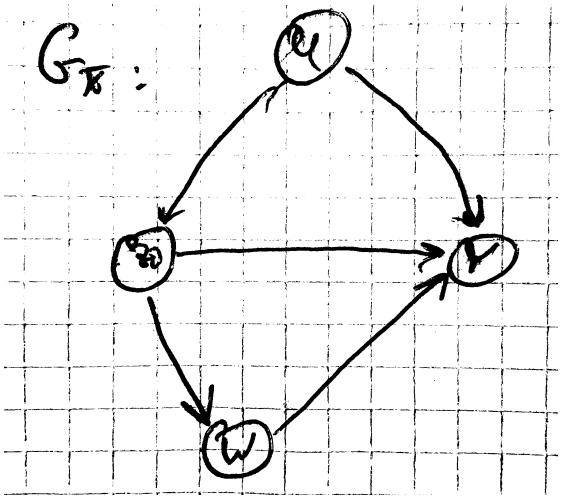


Рис. 3.4: $Z_i \notin Z(W)$

Было показано, что эти правила полны - то есть, если эффект идентифицируем, это можно показать, применив набор таких правил. Однако, задача определения, существует ли такой набор правил остается сложной и не систематизированной.

Примеры применения правил вывода

Рассмотрим структуру на рисунке 3.5 и повыводим для неё эффекты

Задачка 1. $P(z|\hat{x})$

Заметим, что $X \perp\!\!\!\perp_{G_{\underline{X}}} Z$ (единственный путь заблокирован ненаблюдаемым Y), а потому по второму правилу $P(z|\hat{x}) = P(z|x)$

Задачка 2. $P(y|\hat{z})$

$P(y|\hat{z}) = \sum\limits_{x} P(y|\hat{z}, x)P(x|\hat{z})$

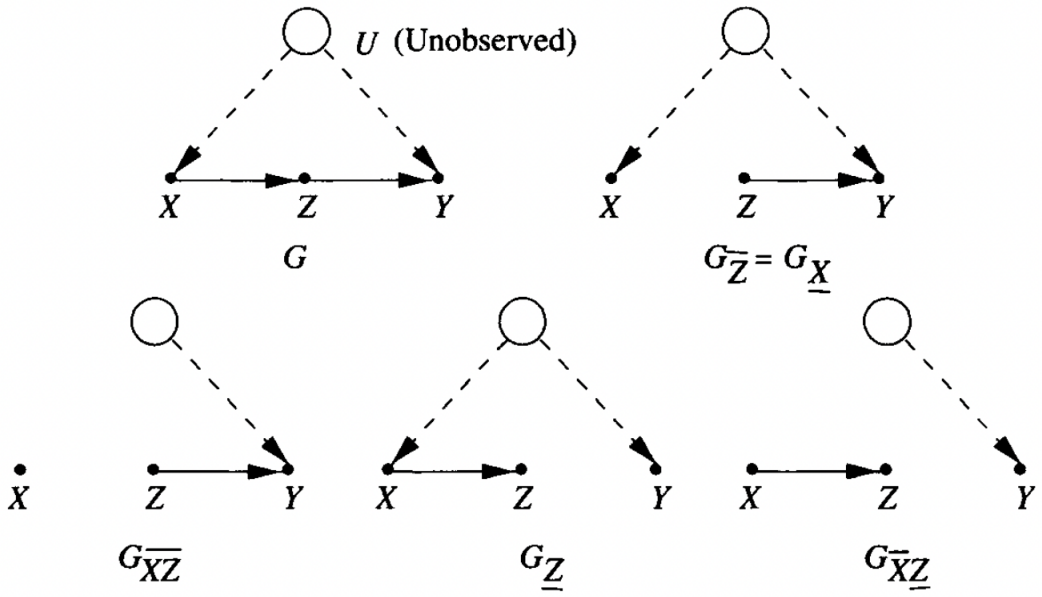


Рис. 3.5: $Z_i \notin Z(W)$

По правилу три, так как $X \perp\!\!\!\perp_{G_{\overline{Z}}}$, то $P(x|\hat{z}) = P(x)$.

По правилу два, так как $Y \perp\!\!\!\perp_{G_{\underline{Z}}} Z|X$, то $P(y|\hat{z}, x) = P(y|z, x)$.

В итоге, получаем знакомую формулу для бэкдор-поправки $P(y|\hat{z}) = \sum\limits_{x} P(y|z, x)P(x)$

Задачка 3. $P(y|\hat{x})$

$P(y|\hat{x}) = \sum\limits_{z} P(y|\hat{x}, z)P(z|\hat{x}) = \sum\limits_{z} P(y|\hat{x}, z)P(z|x)$

Последнее равенство получено как результат решения первой задачки, таким образом, остается вывести $P(y|\hat{x}, z)$.

По правилу 2, $P(y|\hat{x}, z) = P(y|\hat{x}, \hat{z})$; по правилу 3 $P(y|\hat{x}, \hat{z}) = P(y|\hat{z})$. По задачке 2 $P(y|\hat{z}) = \sum\limits_{x} P(y|z, x)P(x)$.

Таким образом, получаем $P(y|\hat{x}) = \sum\limits_{z} P(z|x) \sum\limits_{x'} P(y|z, x')P(x')$ - что совпадает с фронтальной поправкой

Задачка 4. $P(y, z|\hat{x})$

$P(y, z|\hat{x}) = P(y|z, \hat{x})P(z|\hat{x}) = P(z|x) \sum\limits_{x'} P(y|z, x')P(x')$.

Задачка 5. $P(y, x|\hat{z})$

$P(y, x|\hat{z}) = P(y|x, \hat{z})P(x|\hat{z}) = P(y|x, z)P(x)$

Последнее равенство верно, так как X блокирует все бэкдор-пути из Y в Z (по сути, второе правило применили), а для получения второго члена мы применили правило 3.

Графические тесты на идентифицируемость

На картинке 3.6 можно увидеть некоторые примеры ситуаций, когда причинный эффект X на Y неидентифицируем. Во всех этих случаях причиной проблемы является то, что у нас имеется путающая связь (*confounding arc*). Такие связи характеризуются тем, что они образуют бэкдор-путь, который

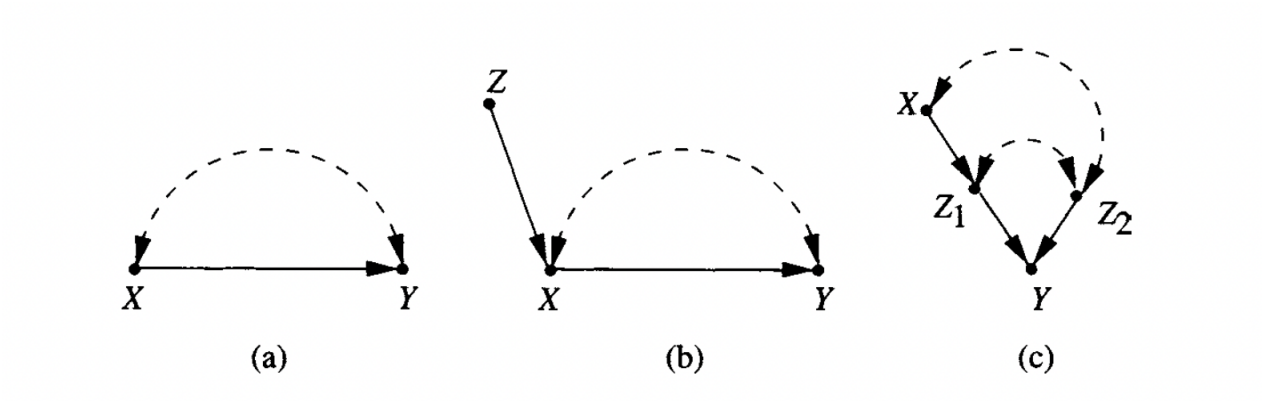


Рис. 3.6: Путающие связи

содержит только ненаблюдаемые переменные и не имеет v-структур: важно отметить, что это бэкдор путь: действительно, если бы было не так, то у нас либо был бы либо направленный путь $X \rightsquigarrow Y$, либо направленный путь $Y \rightsquigarrow X$. В общем же, можно вспомнить, как мы характеризовали случайные ассоциации, в их определении важно то, что ни одну из двух ассоциированных переменных нельзя назвать причиной другой.

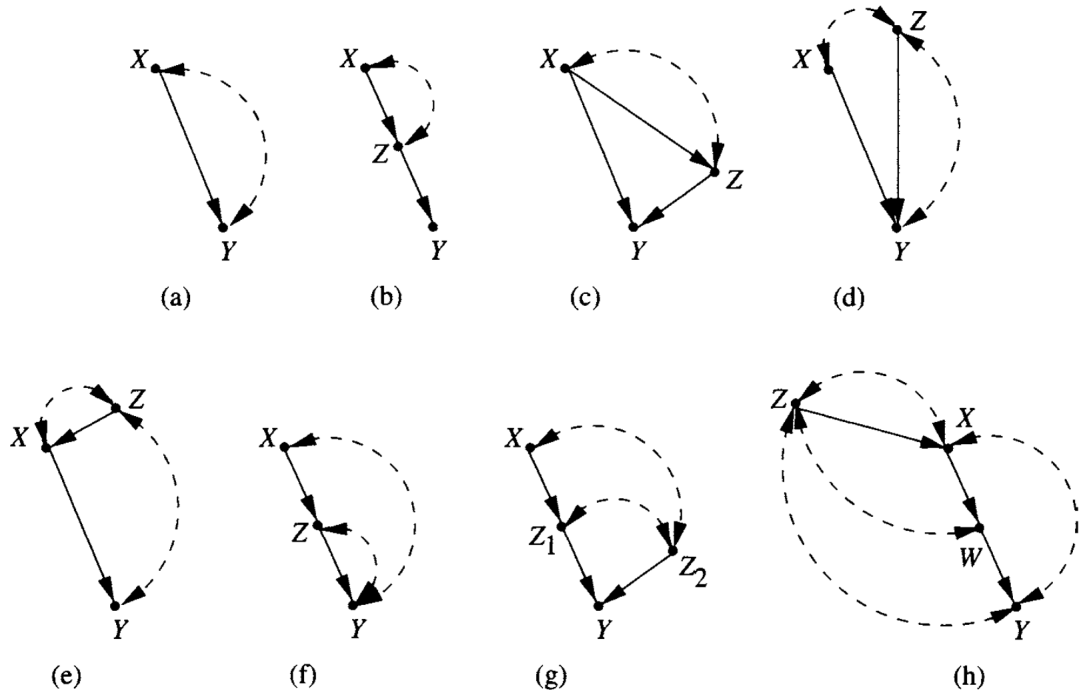


Рис. 3.7: Примеры, когда из-за путающих связей (дуг) нет идентифицируемости

Характерно то, что добавление переменных и ребер не способно привести к идентифицируемости, если такая дуга существует, как на рисунке 3.6 (b).

Другое наблюдение состоит в том, что добавление двунаправленных дуг может ухудшать идентифицируемость, но не улучшать её: таким образом, если причинные выводы недоступны в исходной диаграмме, то в диаграмме с дополнительными дугами тем более.

Любопытно, что возможность оценить эффект на отдельные части не всегда означает возможность оценить совокупный эффект: например, на 3.6 (c) мы можем оценить $P(z_1|\hat{x}) = P(z_1|x)$ (так как $\emptyset$ удовлетворяет бэкдор-критерию относительно (X, Z_1)), и в книге также утверждается возможность вычисления эффекта $P(z_2|\hat{x})$, но невозможность вычисления эффекта $P(z_1, z_2|\hat{x})$. Впрочем, мне пока что не удалось показать вычисляемость эффекта $P(z_2|\hat{x})$:(UPD кажется понял: надо заметить, что $Z_2 \perp\!\!\!\perp_{G_{\overline{X}}} X$, а потому по третьему правилу $P(z_2|\hat{x}) = P(z_2)$

Ещё одно интересное наблюдение: оказывается эффект от совместной интервенции может быть расчитываем, в то время как эффекты от частичных интервенций - нет: так, обращаясь снова к 3.6 (c) можно увидеть, что $P(y|\hat{x}, \hat{z}_2)$ идентифицируем, но $P(y|\hat{x})$ - нет.

Когда у нас нет идентифицируемости? Оказывается, достаточным условием невозможности вычислить $P(y|\hat{x})$ будет наличие дуги между $X \leftrightarrow Z$, где $Z \in de(X), Y \in de(Z)$. Необходимым условием является наличие неблокируемого бэкдор-пути из Y в X (неблокируемого ненаследниками X).

Например, рассмотрим 3.7 (b): запишем $P(y|\hat{x}) = \sum_z P(y|\hat{x}, z)P(z|\hat{x})$. $P(y|\hat{x}, z) = P(y|z)$ по правилу 3, однако, $P(z|\hat{x})$ неидентифицируемо, так как оно соответствует ситуации на картинке (a).

4 Действия, планы и прямые эффекты

Введение: действия, поступки и вероятности

Действия можно интерпретировать с двух сторон: реактивной и осознанной (пассивной и активной). Реактивная интерпретация рассматривает действие как следствие агентских предположений, текущей ситуации и входов от внешней среды, например "Адам съел яблоко, так как Ева его ему дала". Осознанная интерпретация рассматривает действие как выбор в ходе принятия решения, обычно включающий в себя сравнение последствий, например "Адаму было любопытно, что сделает Бог, если он съест яблоко". Мы будем разделять эти две точки зрения, называя первое *поступком*, а второе *действием*.

Поступок наблюдается снаружи, а *действие* - изнутри. Таким образом, *поступок* можно предсказывать и он может служить в качестве доказательства некоторых мотивов и стимулов актора (при условии что актор - часть нашей модели). *Действия*, в свою очередь, не могут быть ни предсказаны, ни предоставлять какое бы то ни было доказательство, так как по определению они ожидают принятия обдуманного решения и превращаются в *поступки*, когда оказываются исполненными.

Основанная на здравом смысле теория принятия решений предписывает рациональным агентам выбирать опцию x, которая максимизирует ожидаемую выгоду:

$$U(x) = \sum_y P(y|do(x))u(y) \tag{4.1}$$

Где $u(y)$ - выгода от исхода y. В свою очередь, "доказательная"теория принятия решений предписывает максимизировать условное матожидание

$$U_{ev}(x) = \sum_y P(y|x)u(y) \tag{4.2}$$

где x (некорректно!) интерпретируется как наблюдаемый поступок.

Парадоксы, которые возникают из этого заблуждения забавны и очевидны: например, можно прийти к выводу, что пациентам не стоит ходить к врачам, чтобы уменьшить вероятность серьезной болезни. Действительно, рассмотрим такие данные

болен	пошел к доктору?	вероятность	выгода
да	да	0.35	-80
нет	да	0.15	90
да	нет	0.15	-100
нет	нет	0.35	100

(4.3)

Тогда $U_{ev}(x = \text{пойти к доктору}) = -29$, в то время как $U_{ev}(x = \text{не пойти к доктору}) = +40$.

Аналогично, рабочим не стоит спешить на работу, чтобы не опоздать, а студентам не стоит готовиться к экзаменам, если они хотят их сдать. Обобщая, приходим к парадоксу, гласящему, что предпринимания мер по улучшению ситуации следует избегать, иначе эти меры увеличат вероятность того, что их действительно стоило предпринимать.

Странность такой логики проистекает из того, что мы пытаемся рассматривать действия как поступки, которые определяются предыдущими ассоциациями, вместо того чтобы воспринимать их как объекты свободного выбора, как это диктует семантика *do*-оператора. Доказательная теория принятия решений утверждает, что нельзя игнорировать статистические доказательства (в нашем случае, это доказательства в пользу того, нужны ли были те или иные поступки или нет), но корректная теория напоминает нам, что действия по определению делают такие доказательства неважными по отношению к принятию решения в данный момент, так как действия *меняют* вероятности, которым поступки обычно подчиняются.

Стишок классный, процитирую as-is

Whatever evidence an act might provide
On what could have caused the act,
Should never be used to help one decide
On whether to choose that same act.

Условные действия и стохастические политики

Ранее мы рассматривали только простейшие интервенции, состоящие в установке некоторого множества переменных в какое-то фиксированное значение. В общем случае, интервенция может состоять в том, что переменная начинает вести себя по какому-то специфичному закону, зависящему от значения других переменных: $x = g(z)$, либо через стохастическое отношение, когда X устанавливается в x с вероятностью $P^*(x|z)$. Далее мы покажем, что идентификация таких интервенций эквивалентна идентификации $P(y|\hat{x}, z)$.

$$\begin{aligned} P(y|do(X = g(z))) &= \sum_z P(y|do(X = g(z)), z)P(z|do(X = g(z))) \\ &= \sum_z P(y|\hat{x}, z)_{x=g(z)}P(z) \\ &= \mathbb{E}_z P(y|\hat{x}, z)_{x=g(z)} \end{aligned} \tag{4.4}$$

$P(z|do(X = g(z))) = P(z)$, так как Z не является наследником X (если бы являлось, у нас бы после интервенции получился бы граф с циклом, чего нам наверно не хотелось бы и вряд ли на практике такая интервенция имела бы физический смысл), а потому никакая интервенция X не может влиять на Z (в постинтервенционном мире X не зависит от всех своих не-наследников, это можно вывести например из правила 3 для do-исчисления). Таким образом, эффект от интервенции с политикой $do(X = g(z)$ может быть вычислен напрямую из выражения $P(y|\hat{x}, z)$ путем подстановки $x = g(z)$ и последующим взятием матожидания от этой величины.

Критерий идентифицируемости для условной политики строже, чем для безусловной: ну действительно, если $P(y|do(X = g(z)))$ идентифицируем, то подставив $g(z) = x$ (константа) мы получим, что

идентифицируема и атомарная интервенция. Обратное, однако) не всегда верно: обуславливание на Z может привнести дополнительные зависимости, которые могут помешать преобразованию $P(y|\hat{x}, z)$ в выражение без do-операторов.

Стохастическая политика, когда X ведет себя по $P^*(x|z)$, можно рассмотреть подобным образом: такая политика означает, что мы делаем интервенцию $do(X = x)$ с вероятностью $P^*(x|z)$. Таким образом, можем записать

$$P(y|do(X \sim P^*(x|z))) = \sum_x \sum_z P(y|\hat{x}, z)P^*(x|z)P(z) \tag{4.5}$$

мы приходим к тому, что необходимым и достаточным условием идентифицируемости стохастической политики будет также идентифицируемость $P(y|\hat{x}, z)$.

Когда эффект от действия идентифицируем?

Далее в двух теоремах будут описаны необходимое и достаточное условие идентифицируемости эффекта одной переменной на другую.

Теорема 4.1 *Пусть X и Y - две переменные в полумарковской модели, характеризуемой графом G (DAG, но неизмеренные переменные могут иметь более одного наследника). Достаточным условием идентифицируемости $P(y|\hat{x})$ будет то, что G удовлетворяет хотя бы одному из четырёх условий*

1. Не существует бэкдор-пути из X в Y, то есть $X \perp\!\!\!\perp_{G_{\underline{X}}} Y$.

2. Нет направленного пути из X в Y.

3. Существует множество вершин B, которое блокирует все бэкдор-пути из X в Y такое, что $P(b|\hat{x})$ идентифицируемо (специальный случай - когда все вершины в B являются ненаследниками X, тогда $P(b|\hat{x}) = P(b)$).

4. Существуют множества $Z_1 \subset de(X), Z_2 : Z_2 \cap de(X) = \emptyset$:

(i) Z_1 блокирует любой направленный путь из X в Y в $G_{\overline{Z_1}}$: то есть, $Y \perp\!\!\!\perp_{G_{\overline{Z_1 X}}} X|Z_1$. В данном случае, мы рассматриваем так модифицированный граф, потому что нам во-первых, надо убрать все бэкдор-пути (ведь у нас утверждение о направленных путях, значит нас не интересуют пути с ребрами, входящими в X), а во-вторых, мы хотим, чтобы в этом графе Z_1 блокировало только направленные пути: значит нам не страшно, что оно возможно разблокирует v-структуру, именно поэтому мы рассматриваем граф с удаленными входящими в Z_1 ребрами - ясно что все направленные пути что в таком графе, что в графе с неудаленными входящими в Z_1 ребрами, будут заблокированы Z_1.

(ii) Z_2 блокирует все бэкдор-пути между Z_1 и Y в $G_{\overline{X}}$: $Y \perp\!\!\!\perp_{G_{\underline{Z_1}\overline{X}}} Z_1|Z_2$. Почему тут мы рассматриваем граф с удаленными ребрами из Z_1 - понятно: ведь нас интересует блокирование только бэкдор-путей, а они по таким ребрам не идут.

(iii) Z_2 блокирует все бэкдор-пути между X и Z_1: $X \perp\!\!\!\perp_{G_{\underline{X}}} Z_1|Z_2$ - тут мотивация рассматривать такой граф тоже очевидна.

(iv) Z_2 не активирует никакой бэкдор-путь из X в Y: $X \perp\!\!\!\perp_{G_{\overline{X(Z_2)}\overline{Z_1}}} Y|Z_1, Z_2$. Тут мы рассматриваем граф с удаленными входящими ребрами в $X(Z_2)$, так как такие вершины не могут разблокировать при обуславливании на них никакой бэкдор путь из X в Y. Вершины же, которые $X \backslash X(Z_2)$ - это такие вершины из X, у которых есть наследники в Z_2, и входящие ребра в них в графе не удаляются чтобы гарантировать, что при обуславливании на Z_2 мы не разблокировали как раз никакой бэкдор-путь в них.

Пруф:

1. Это условие означает, что $Y \perp\!\!\!\perp_{G_{\underline{X}}} X$, значит можно непосредственно применить второе правило do-исчисления, то есть $P(y|\hat{x}) = P(y|x)$.

2. Это условие означает, что Y - не наследник X в графе, а значит, $P(y|\hat{x}) = P(y)$ - это следует из правила 3 do-вычислений: $Y \perp\!\!\!\perp_{G_{\overline{X}}} X$

3. В этом случае, запишем $P(y|\hat{x}) = \sum_b P(y|\hat{x}, b)P(b|\hat{x})$

Далее, используя бэкдор критерий (ну или правило 2 do-вычислений) $P(y|\hat{x}, b) = P(y|x, b)$, и так как $P(b|\hat{x})$ тоже идентифицируемо по условию, то $P(y|\hat{x}) = \sum_b P(y|x, b)P(b|\hat{x})$ идентифицируемо.

4. Тут придется попотеть:
$P(y|\hat{x}) = \sum_{z_1,z_2} P(y|\hat{x}, z_1, z_2)P(z_1, z_2|\hat{x})$.
$P(y|\hat{x}, z_1, z_2) = P(y|\hat{x}, \hat{z}_1, z_2)$ по правилу 2, так как Z_2 блокирует все бэкдор-пути между Z_1 и Y.
Далее, $P(y|\hat{x}, \hat{z}_1, z_2) = P(y|\hat{z}_1, z_2)$ по правилу 3, так как $Y \perp\!\!\!\perp_{G_{\overline{Z_1}\overline{X(Z_2)}}} X|Z_2, Z_1$.
Чтобы избавиться от шапочки тут, мы можем записать $P(y|\hat{z}_1, z_2) = \sum_{x'} P(y|\hat{z}_1, z_2, x')P(x'|\hat{z}_1, z_2)$.
Заметим, что $P(y|\hat{z}_1, z_2, x') = P(y|z_1, z_2, x')$, так как по (ii) и (iii) единственный возможный бэкдор-путь между Z_1 и Y, не заблокированный Z_2 должен иметь ребро, входящее в X и ребро выходящее из X: и правда, если бы не было входящего в X ребра, то из (ii) следует, что такой путь заблокирован в $G_{\underline{Z_1}}$ множеством Z_2, то есть и в исходном графе тоже (раз он бэкдор относительно Z_1). Если же нет исходящего из X ребра, то X образует в этом пути v-структуру, а значит Z_2 должно блокировать часть этого пути между Z_1 и X по (iii), ведь эта часть является бэкдор-путем из X в Z_1. Но такой путь (вида $Z_1 \rightsquigarrow ... \to X \to ... \rightsquigarrow Y$) будет заблокирован, если мы обусловимся на X, что мы собственно и делаем.

Используя рассуждения выше, приходим к выводу, что по правилу 2 $P(y|\hat{z}_1, z_2, x) = P(y|z_1, z_2, x)$.

Заметим далее, что $P(x'|\hat{z}_1, z_2) = P(x'|z_2)$, так как Z_1 - наследники X, а значит действие над ними на X не влияет.

$$P(y|\hat{x}, z_1, z_2) = \sum_{x'} P(y|z_1, z_2, x')P(x'|z_2) \tag{4.6}$$

Теперь рассмотрим $P(z_1, z_2|\hat{x}) = P(z_1|\hat{x}, z_2)P(z_2|\hat{x})$. Заметим, что раз Z_2 не содержит наследников X, то $P(z_2|\hat{x}) = P(z_2)$ (это следует из правила 3 do-исчисления). $P(z_1|\hat{x}, z_2) = P(z_1|x, z_2)$, так как Z_2 блокирует все бэкдор-пути из X в Z_1 по (ii).

Таким образом, приходим к полному выражению:

$$P(y|\hat{x}) = \sum_{z_1,z_2} P(z_1|x, z_2)P(z_2) \sum_{x'} P(y|z_1, z_2, x')P(x'|z_2) \tag{4.7}$$

■

Теорема 4.2 ***О необходимом условии идентифицируемости*** *Как минимум одно из условий теоремы 4.1 является необходимым для идентифицируемости $P(y|\hat{x})$.*

Доказывать будем от противного: предположим, что существует граф G и $P(y|\hat{x})$ такие, что никакое из четырех условий не выполняется, но $P(y|\hat{x})$ идентифицируемо, и покажем, что это приводит к противоречию.

Итак, раз выражение идентифицируемо, то существует какая-то конечная последовательность применений правил 1-3 do-вычислений, приводящее выражение к бесшапочному виду. Чтобы убрать шапочку над x, у нас есть всего два правила из трех:
Кейс 1: применив правило 2: $P(y|\hat{x}, \hat{z}, w) \to P(y|x, \hat{z}, w)$ если $Y \perp\!\!\!\perp_{G_{\overline{Z}\underline{X}}} X|Z, W$
Кейс 2: применив правило 3: $P(y|\hat{x}, \hat{z}, w) \to P(y|\hat{z}, w)$ если $Y \perp\!\!\!\perp_{G_{\overline{Z}\overline{X(W)}}} X|Z, W$ (Кейс 2)

Рассмотрим каждый из этих двух кейсов по очереди.

Кейс 1 Существуют некоторые подмножества X, W: $Y \perp\!\!\!\perp_{G_{\overline{Z}\underline{X}}} X|Z, W$. Не умаляя общности будем считать, что Z, W - минимальные, иначе уберем из них лишние вершины и будем доказывать для таких множеств.

Если бы было $Y \perp\!\!\!\perp_{G_{\underline{X}}} X|Z,W$, то множество $Z \cup W$ блокировало бы все бэкдор-пути из X в Y, но тогда выполнялось бы условие 3 4.1. Правда, непонятно пока а с чего мы решили что $P(z,w|\hat{x})$ идентифицируемо? Я правда пока не понимаю, выглядит как неточность в доказательстве.

Пруф этой теоремы можно найти в Galles, Pearl (1995), мне пока лень искать. ■

Заметки об эффективности

Условия 3 и 4 теоремы 4.1 требуют исчерпывающего перебора возможных множеств B и Z_1, Z_2. Однако, следующие теоремы позволяют существенно уменьшить пространство поиска.

Теорема 4.3 *Если $P(b_i|\hat{x})$ идентифицируемо для какого-то минимальноого множества B_i, то $P(b_j|\hat{x})$ идентифицируемо для любого другого минимального множества B_j.*

Эта теорема позволяет осуществить проверку идентифицируемости для любого минимального множества, блокирующего все бэкдор-пути, и вынести по результату вердикт - выполняется ли условие 3 теоремы 4.1 или нет.

Для доказательства этой теоремы воспользуемся леммой

Лемма 4.1 *Если выражение $P(y|\hat{x})$ идентифицируемо и множество вершин Z лежат на направленном пути из X в Y, то $P(z|\hat{x})$ тоже идентифицируемо.*

Теорема 4.4 *Пусть Y_1, Y_2 - два подмножества вершин таких, что выполняется одно из двух:*
(i) $Y_1 \cap de(X) = \emptyset$
(ii) $Y_1 \cup Y_2 \subset de(X)$ и $Y_1 \cap de(Y_2) = \emptyset$
Тогда $P(y_1, y_2|\hat{x})$ идентифицируемо $\iff$ $P(y_1|\hat{x})$ и $P(y_2|\hat{x}, y_1)$ идентифицируемо.

Ну, стрелка влево очевидна: $P(y_1, y_2|\hat{x}) = P(y_1|\hat{x})P(y_2|\hat{x}, y_1)$, так что из выводимости правой части и правда следует выводимость левой.

Стрелка вправо - не столь очевидна.

Теорема 4.5 *Если существует Z_1, удовлетворяющее условию 4 теоремы 4.1, то множество, состоящее из детей X, пересеченное с предшественниками Y также будет удовлетворять условиям, предъявляемым к Z_1.*

Эту теорема делает поиск множества Z_1 тривиальным.

Алгоритм вывода $P(y|\hat{x})$

Алгоритм ClosedForm

Вход: Запрос $P(y|\hat{x})$

Выход: Либо выражение для $P(y|\hat{x})$ в замкнутой форме, содержащее только наблюдаемые переменные, либо FAIL, если выражение не идентифицируемо.

1. Если $X \perp\!\!\!\perp_{G_{\overline{X}}} Y$ то вернуть $P(y)$
2. Иначе если $X \perp\!\!\!\perp_{G_{\underline{X}}}$ то вернуть $P(y|x)$
3. Иначе, обозначим $B = BlockingSet(X, Y)$, $Pb = ClosedForm(P(b|\hat{x}))$. Если $Pb \neq FAIL$, вернуть $\sum_b P(y|x, b)Pb$
4. Иначе, пусть $Z_1 = Children(X) \cap (Y \cup Ancestors(Y))$, $Z_3 = BlockingSet(X, Z_1)$, $X_4 = BlockingSet(Z_1, Y)$, $Z_2 = Z_3 \cup Z_4$. Если $Y \notin Z_1$ и $X \notin Z_2$, вернуть $\sum_{z_1,z_2} P(z_1|x, z_2)P(z_2) \sum_{x'} P(y|z_1, z_2, x')P(x'|z_2)$
5. Иначе, вернуть FAIL.

Вот с этим алгоритмом тоже есть такая хитрость: при рекурсивном вызове вообще говоря может оказаться, что $P(b|\hat{x})$ вызывается с $|B| > 1$, но изначальная теорема про необходимость и достаточность одного из четырех условий доказана только для одиночных переменных X, Y.

Идентификация динамических планов

Мотивация

Рассмотрим пример с причинным графом, как на 4.1. X_1, X_2 обозначают лечения, которые предписывают врачи, Z - наблюдения, по которым второй врач выносит решение о типе X_2, а Y - выжил ли пациент. Ненаблюдаемы переменные: U_1 - часть истории пациента, U_2 - склонность пациента к излечению.

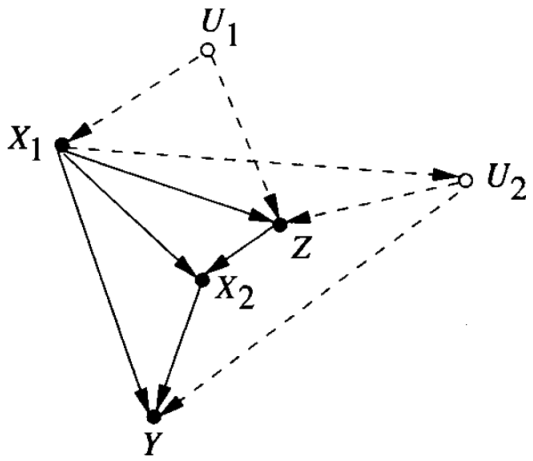


Рис. 4.1: Задача оценки эффекта плана $(do(x_1), do(x_2))$ на Y

Реальным примером такого графа может служить сопутствующее спиду заболевание PCP - pneumocystis pneumonia - соответствует Z в графе. Она эффективно лечится, поэтому не имеет прямого эффекта на выживаемость Y, но является индикатором того, что иммунное состояние человека U_2 не в лучшем состоянии, что уже в свою очередь может приводить к смерти. X_1, X_2 в данном случае обозначают прием бактрима - лекарства, борющегося с PCP и также способное препятствовать смерти по другим причинам. Для предписания бактрима доктора руководствуются в том числе более ранней PCP историей пациента, но предположим что она не была записана для нашего анализа.

Перед нами поставлена следующая задача: есть большой набор собранных данных (X_1, Z, X_2, Y). Далее, к нам приходит пациент, и мы должны оценить эффект от безусловного плана $(do(x_1), do(x_2))$ на его выживание. В целом, что нам надо сделать - разработать план действий на основании наблюдаемых уже исполненных действий (актов) других акторов, чьи стратегии для нас неизвестны.

В данном случае мы сталкиваемся вновь с задачей идентификации, которая как и ранее не является тривиальной из-з наличия путающих связей. В случае идентификации планов, однако, есть дополнительная сложность - некоторые конфаундеры аффектятся контролируемыми переменными (например, Z - аффектится X_1).

В данном случае, мы можем идентифицировать $P(y|\hat{x}_1, \hat{x}_2)$:

$$\begin{aligned} P(y|\hat{x}_1, \hat{x}_2) &= P(y|x_1, \hat{x}_2) \\ &= \sum_z P(y|x_1, z, \hat{x}_2) P(z|x_1, \hat{x}_2) \\ &= \sum_z P(y|x_1, z, x_2) P(z|x_1) \end{aligned} \tag{4.8}$$

Первый переход следует из правила 2 (так как $Y \perp\!\!\!\perp_{G_{\overline{X}_2\underline{X}_1}} X_1|X_2$). Второй переход - это просто маргинализация совместного распределения на $Y, Z|X_1, X_2$. Третий - используем то, что Z не наследник X_2, значит интервенция на X_2 на него не влияет (по правилу 3), а X_1, Z блокирует все бэкдор пути между X_2 и Y, так что первый множитель преобразуется по правилу 2.

Нотация и предположения при идентификации планов

Мы начинаем с того, что предполагаем известность причинной диаграммы, которая предоставляет качественное представление о процессе, генерирующем данные.

Задача контролирования определяется DAG G с вершинами состоящими из четырех непересекающихся множеств $V = \{X, Z, U, Y\}$ где

X - множество контролируемых переменных

Z - множество наблюдаемых переменных-ковариаций

U - множество ненаблюдаемых (латентных) переменных

Y - переменная-исход

Пронумеруем контролируемые переменные $X_1, X_2, ..., X_n$ так, чтобы X_k не имело направленных путей в $X_1, ..., X_{k-1}$, а также Y - наследник X_n. Планом называется упорядоченная последовательность $(\hat{x}_1, ..., \hat{x}_n)$. Условный план - упорядоченная последовательность $\hat{g}_1(z_1), ..., \hat{g}_n(z_n)$ где g_k - функция из Z_k в X_k, $\hat{g}_k(z_k)$ означает установить X_k в $g(z_k)$, и Z_k не содержит наследников X_k относительно G.

Задача стоит в вычислении эффекта безусловного плана на переменную Y $P(y|\hat{x}_1, ...\hat{x}_n)$.

Идентификация плана: последовательный бэкдор-критерий

Теорема 4.6 *$P(y|\hat{x}_1, ...\hat{x}_n)$ идентифицируем если* $\forall 1 \leq k \leq n \; \exists Z_k \subset Z$:

1. $\forall j > k \; Z_k \cap de(X_j) = \emptyset$

2. $Y \perp\!\!\!\perp_{G_{\underline{X}_k \overline{X}_{k+1}, .., \overline{X}_n}} X_k|X_1, ..., X_{k-1}, Z_1, ..., Z_k$

В этом случае эффект от плана выражается по формуле

$$\sum_{z_1, .., z_n} P(y|z_1, ..., z_n, x_1, ..., x_n) \prod_{k=1}^{n} P(z_k|z_1, ..z_{k-1}, x_1, ..., x_{k-1}) \tag{4.9}$$

Пруф: раз Z_k не содержит наследников $X_j \; \forall \; j > k$, то по правилу 3 do-исчисления

$$P(z_k|z_1, .., z_{k-1}, x_1, .., x_{k-1}, \hat{x}_k, ..., \hat{x}_n) = P(z_k|z_1, .., z_{k-1}, x_1, .., x_{k-1}) \tag{4.10}$$

Условие 2 теоремы позволяет нам применить правило 2:

$$P(y|z_1, .., z_k, x_1, .., x_{k-1}, \hat{x}_k, ..., \hat{x}_n) = P(z_k|z_1, .., z_{k-1}, x_1, .., x_k, \hat{x}_{k+1}, ..., \hat{x}_n) \tag{4.11}$$

убрав шапку с x_k.

Таким образом, получаем

$$
\begin{aligned}
P(y|\hat{x}_1,...,\hat{x}_n) &= \sum_{z_1} P(y|z_1,\hat{x}_1,...,\hat{x}_n)P(z_1|\hat{x}_1,...,\hat{x}_n) \\
&= \sum_{z_1} P(y|z_1,x_1,...,\hat{x}_n)P(z_1) \\
&= \sum_{z_1,z_2} P(y|z_1,z_2,x_1,\hat{x}_2,...,\hat{x}_n)P(z_1)P(z_2|x_1,\hat{x}_2,...,\hat{x}_n) \\
&= \sum_{z_1,z_2} P(y|z_1,z_2,x_1,x_2,...,\hat{x}_n)P(z_1)P(z_2|x_1,z_1) \\
&= \dots \\
&= \sum_{z_1,...,z_n} P(y|z_1,z_2,...,z_n,x_1,x_2,...,x_n) \prod_{k=1}^{n} P(z_k|x_1,...,x_{k-1},z_1,...,z_{k-1})
\end{aligned}
\tag{4.12}
$$

■

Задача контролирования называется *G-идентифицируемой*, если $P(y|\hat{x}_1,...,\hat{x}_n)$ идентифицируемо с использованием критерия из предыдущей теоремы. Последовательность соответствующих ковариаций Z называется *приемлемой*.

Стоит отметить, что G-идентифицируемость $\implies$ идентифицируемость, но не наоборот: действительно, на k-м шаге общая идентифицируемость может выполняться в том числе с использованием обуславливания на Z_k, содержащие наследников X_k - на переменные, на которые может повлиять действие $do(X_k = x_k)$, но в G-идентифицируемости это запрещено.

Идентификация плана: процедура

Пока что мы лишь декларативно описали, как идентифицировать план. Однако, пока не понятно, как выбирать множества Z_k, и любой ли выбор подойдет. На рисунке 4.2 можно увидеть пример, когда выбор неудачного Z_1 не позволяет сделать дальнейшие выборы множеств.

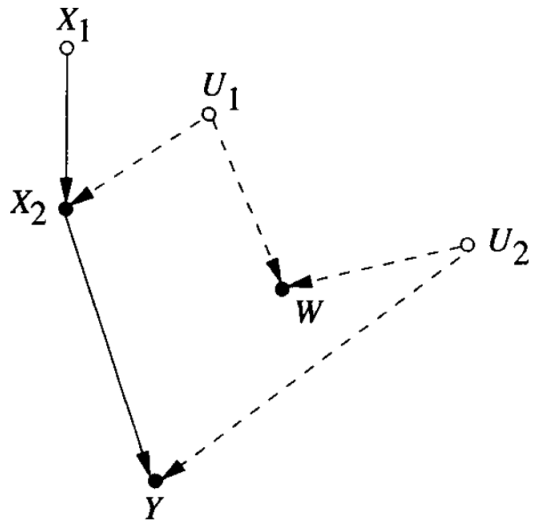


Рис. 4.2: Выбор $Z_1 = \{W\}$ делает невозможным выбор какого-либо допустимого Z_2

Теорема 4.7 *Если существует допустимая последовательность $Z_1^*, ..., Z_n^*$, то для любой* ***минимальной*** *допустимой подпоследовательности $Z_1, ..., Z_{k-1}$ существует допустимое множество Z_k.*

Пруф в Pearl and Robins (1995) ■

Следствие: задача контролирования G-идентифицируема тогда и только тогда, когда следующий алгоритм возвращает успех:

1. $k := 1$
2. Выбрать любое $Z_k \subset Z \backslash (de(X_k) \cup ... \cup de(X_n))$
3. Если Z_k не нашлось, вернуть FAIL, иначе $k := k + 1$
4. Если $k = n + 1$, вернуть OK

Следующая теорема позволяет эффективно выбирать Z_k

Теорема 4.8 *Эффект плана $P(y|\hat{x}_1, ..., x_n)$ G-идентифицируем $\equiv \forall k \in [1, ..., n]$ $Y \perp\!\!\!\perp_{G_{\underline{X}_k, \overline{X}_{k+1}, ..., \overline{X}_n}} X_k | X_1, .., X_{k-1}, W_1, ..., W_k$ где W_k - такие ковариации, которые не содержат наследников $X_k, ..., X_n$ и имеют наследниками либо Y, либо X_k в графе $G_{\underline{X}_k, \overline{X}_{k+1}, ..., \overline{X}_n}$.*

В таком случае эффект плана выражается как

$$\sum_{w_1, ..., w_n} P(y|w_1, w_2, ..., w_n x_1, x_2, ..., x_n) \prod_{k=1}^{n} P(w_k|x_1, ..., x_{k-1}, w_1, ..., w_{k-1}) \tag{4.13}$$

Пруф снова в Pearl and Robins (1995) ■

Стоит отметить, что несмотря на то, что детерминированный эффективный алгоритм в 4.8 позволяет определить G-идентифицируемость, он все-таки зависит от выбранного порядка $X_1, ..., X_n$. Не исключена ситуация, когда при одном допустимом графом G порядке контролируемых переменных допустимая последовательность ковариаций существует, а при другом - нет.

Это означает, что для для проверки того, что план не G-идентифицируем необходимо проверить все допустимые графом порядки контролирующих переменных.

Прямые и непрямые эффекты

Причинный эффект, который мы до этого исследовали $P(y|\hat{x})$ - это полный эффект переменной/множества переменных X на Y. Во многих случаях, однако, нас интересует не он, а прямой эффект - эффект, который не проводится через какие-то сторонние переменные - или, чуть точнее - эффект от изменения X когда все другие переменные фиксированы.

Примером, когда может быть интересен прямой эффект, является ситуация, когда хочется оценить влияние таблеток для контроля рождаемости на образование тромбов: с одной стороны, есть гипотеза, что таблетки повышают тромбообразование. С другой стороны, есть непрямой негативный эффект на появление тромбов, так как беременность повышает их вероятность.

Другой пример состоит в анализе наличия дискриминации по половому признаку: в данном случае интересует именно вопрос непосредственного влияния пола на результат найма, при этом влияние пола или расы на квалификацию, как и влияние квалификации не является предметом разбирательства.

Прямые эффекты: определение и идентификация

def Прямой эффект определяется как $P(y|\hat{x}, \hat{s}_{XY})$, где s_{XY} - множество всех эндогенных (наблюдаемых) переменных системы кроме X, Y.

Заметим, что контролировать все возможные переменные вообще говоря не обязательно - достаточно контролировать всех непосредственных родителей Y: действительно, в этом случае никакой направленный путь в Y, кроме ребра $X \to Y$, не будет активен.

Следствие 4.1 *Непосредственный эффект X на Y задаётся $P(y|\hat{x}, \hat{pa}_{Y \backslash X})$, где $\hat{pa}_{Y \backslash X}$) - любая реализация всех родителей Y, кроме X.*

Ясно, что если X не является родителем Y, то $P(y|\hat{x}, \hat{pa}_{Y\setminus X}) = P(y|\hat{pa}_Y)$ является константным относительно x, что соответствует по смыслу тому, что X не имеет непосредственного влияния на Y.

Полагая, что X - родитель Y, мы приходим к выводу, что непосредственный эффект X на Y идентифицируем, если $P(y|\hat{pa}_Y)$. Этот эффект идентифицируем, в свою очередь, если идентифицируем план для родителей:

Теорема 4.9 *Пусть $PA_Y = \{X_1, ..., X_k, ..., X_m\}$. Непосредственный эффект X_k на Y если соблюдается условие G-идентифицируемости плана $(\hat{x}_1, ..., \hat{x}_m)$ относительно некоторого согласованного упорядочивания родителей Y.*

5 Причинность и структурные модели в социальной науке и экономике

Интро

Модели структурных уравнений было разработано генетиками (Wright, 1921) и экономистами (Haavelmo, 1943). На вопрос, при каких условиях можно причинно интерпретировать коэффициенты структурных уравнений, эти ребята бы ответили "Всегда!". *Структурным* уравнение $y = \beta x + \epsilon$ делает как раз утверждение, что причинная связь между x и y характеризуется β, и никакие статистические связи между x и ϵ не могут это поменять.

На удивление, это понимание причинной семантики в структурных уравнениях со временем утратилось, и большинство перестало воспринимать их как носитель причинной семантики от слова совсем. Проблема состоит в том, что знак равенства в структурных уравнениях, изначально воспринимавшийся как ассиметричный (не как алгебраическое равенство), со временем утратил эту асимметрию.

Графы и тестирование моделей

Как мы уже знаем, по множеству структурных уравнений можно построить причинную диаграмму с ребрами из PA_i в X_i, где PA_i - множество переменных, через которые X_i определяется в структурном уравнении $x_i = f(pa_i, \epsilon_i)$.

NB здесь и далее для краткости иногда буду писать SEM = structural equation model.

Если шумы ε_i распределены по многомерному нормальному распределению (что часто предполагается в SEM), то X_i тоже будут распределены по многомерному нормальному, и будут целиком характеризоваться корреляционными коэффициентами ρ_{ij}.

Удобное свойство многомерных нормальных распределений состоит в том, что условные дисперсия $\sigma^2_{X|z}$, ковариация $\sigma_{XY|z}$ и соответственно коэффициент корреляции $\rho_{XY|z} = \frac{\sigma_{XY|z}}{\sigma_{X|z}\sigma_{Y|z}}$ не зависят от значения z, поэтому их часто обозначают $\sigma^2_{X\cdot Z}$, $\sigma_{XY\cdot Z}$, $\rho_{XY\cdot Z}$. Эти условные коэффициенты также называют частичными.

Частичный коэффициент регрессии определяется как $r_{YX\cdot Z} = \rho_{XY\cdot Z}\frac{\sigma_{Y\cdot Z}}{\sigma_{X\cdot Z}}$. Другой важный факт состоит в том, что $r_{YX\cdot Z} = 0 \Leftrightarrow X \perp\!\!\!\perp Y|Z$.

Для разминки, докажем его:

1. $\Leftarrow$: пусть $X \perp\!\!\!\perp Y|Z$. Это значит, что $p(x|z, y) = p(x|z)$. Запишем

$$\begin{aligned}\sigma_{XY|z} &= \mathbb{E}_{x,y\sim p(x,y|z)}[(x - \mathbb{E}_{x\sim p(x|z)})(y - \mathbb{E}_{y\sim p(y|z)})] \\ &= \mathbb{E}_{x,y\sim p(x,y|z)}[xy] - \mathbb{E}_{x\sim p(x|z)}\,\mathbb{E}_{y\sim p(y|z)}\end{aligned} \tag{5.1}$$

$$\begin{aligned}\mathbb{E}_{x,y\sim p(x,y|z)}[xy] &= \int\limits_{x,y} p(x,y|z)xy\ dx\ dy \\ &= \int\limits_{y} p(y|z)y\ dy \int\limits_{x} p(x|z)x\ dx \\ &= \mathbb{E}_{x\sim p(x|z)}\,\mathbb{E}_{y\sim p(y|z)}\end{aligned} \tag{5.2}$$

.

Таким образом, видим, что действительно в этом случае частичная ковариация, а значит и коэффициент корреляции, равны 0.

2. $\Rightarrow$: пусть $\sigma_{XY|z} = 0$. Надо показать, что переменные условно независимы при обуславливании на Z.

Рассмотрим блочную матрицу

$$\Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix} \tag{5.3}$$

Найдем выражение для блочного представления обратной к этой матрице: пусть

$$\Sigma^{-1} = \begin{bmatrix}A & B \\ C & D\end{bmatrix} \tag{5.4}$$

Запишем уравнения, которые следуют из обратности матрицы $\Sigma\Sigma^{-1} = I$:

$$\Sigma_{11}A + \Sigma_{12}C = \mathbf{I} \tag{5.5}$$
$$\Sigma_{11}B + \Sigma_{12}D = \mathbf{0} \tag{5.6}$$
$$\Sigma_{21}A + \Sigma_{22}C = \mathbf{0} \tag{5.7}$$
$$\Sigma_{21}B + \Sigma_{22}D = \mathbf{I} \tag{5.8}$$
$$\tag{5.9}$$

Из третьего и четвертого уравнения получаем

$$C = -\Sigma_{22}^{-1}\Sigma_{21}A \tag{5.10}$$
$$D = \Sigma_{22}^{-1} - \Sigma_{22}^{-1}\Sigma_{21}B \tag{5.11}$$

Подставляя в 1, получаем выражение для A:

$$A = (\Sigma_{11}^{-1} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \tag{5.12}$$

Обозначим $\Sigma_* = \Sigma_{11}^{-1} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, тогда

$$A = \Sigma_*^{-1} \tag{5.13}$$
$$C = -\Sigma_{22}^{-1}\Sigma_{21}\Sigma_*^{-1} \tag{5.14}$$
$$B = C^T = -\Sigma_*^{-1}\Sigma_{12}\Sigma_{22}^{-1} \tag{5.15}$$
$$D = \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}\Sigma_*^{-1}\Sigma_{12}\Sigma_{22}^{-1} \tag{5.16}$$

Распишем теперь расстояние Махаланобиса:

$$
\begin{aligned}
d(y,\mu) = (y-\mu^T)\Sigma^{-1}(y-\mu) &= \begin{bmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{bmatrix}^T \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{bmatrix} \\
&= \begin{bmatrix} (y_1-\mu_1)^T A + (y_2-\mu_2)^T C & (y_1-\mu_1)^T B + (y_2-\mu_2)^T D \end{bmatrix} \begin{bmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{bmatrix} \\
&= (y_1-\mu_1)^T A(y_1-\mu_1) + (y_2-\mu_2)^T C(y_1-\mu_1) + \\
&+ (y_1-\mu_1)^T B(y_2-\mu_2) + (y_2-\mu_2)^T D(y_2-\mu_2) \\
&= (y_1-\mu_1)^T \Sigma_*^{-1}(y_1-\mu_1) - (y_2-\mu_2)^T \Sigma_{22}^{-1}\Sigma_{21}\Sigma_*^{-1}(y_1-\mu_1) + \\
&- (y_1-\mu_1)^T \Sigma_*^{-1}\Sigma_{12}\Sigma_{22}^{-1}(y_2-\mu_2) + \\
&+ (y_2-\mu_2)^T(\Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}\Sigma_*^{-1}\Sigma_{12}\Sigma_{22}^{-1})(y_2-\mu_2) \\
&= (y_1-\mu_1)^T \Sigma_*^{-1}(y_1-\mu_1) \\
&= (y_1-(\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(y_2-\mu_2)))^T \Sigma_*^{-1}(y_1-(\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(y_2-\mu_2))) \\
&+ (y_2-\mu_2)\Sigma_{22}^{-1}(y_2-\mu_2) \\
&= (y_1-\mu_*)\Sigma_*^{-1}(y_1-\mu_*) + (y_2-\mu_2)\Sigma_{22}^{-1}(y_2-\mu_2)
\end{aligned}
\tag{5.17}
$$

Посмотрев внимательно на формулу, ясно, что при обуславливании на y_2 у нас будет получаться нормальное распределение $N(\mu_*, \Sigma_*^{-1})$. Заметим, что матрица ковариации условного распределения действительно не зависит от значения y_2 - оно влияет только на смещение матожидания для y_1, что прикольно.

Рассмотрим разбиение, при котором $y_2 = z$, $y_1 = [x, y]$, при этом x, y - одиночные переменные, z может быть вектором. При обуславливании на Z, мы получаем просто новое нормальное распределение двух переменных (о нем чуть выше сказано).

Раз у нас $\sigma_{XY|z} = 0$, то $\mathbb{E}_{x\sim p(x|z,y)}[x] = \mathbb{E}_{x\sim p(x,y|z)} = \mathbb{E}_{x\sim p(x|z)}[x]$. Аналогично и дисперсия $\sigma^2_{x|y,z} = \sigma^2_{x|z}$, но это и значит для нормального распределения, что $p(x|z,y) = p(x|z)$, что и требовалось.

■

Мы доказали, что из условной независимости следует нулевая условная корреляция для любого распределения (не обязательно многомерного нормального), а значит справедливо следствие d-сепарации:

Следствие 5.1 *В любой марковской модели, соответствующей DAG G, частичная корреляция $\rho_{XY\cdot Z}$ исчезает как только Z d-разделяет X, Y, вне зависимости от параметров модели. Более того, никакие другие корреляции не будут 0 для любых параметров модели.*

Теорема 5.1 ***d- сепарация в произвольных линейных моделях*** *Для любой линейной модели, согласованной с диаграммой D, в которой могут быть двунаправленные дуги и циклы, частичная корреляция $\rho_{XY\cdot Z}$ пропадает, если Z d-разделяет X, Y.*

Теперь, предположим мы хотим протестировать модель на предмет того, соотносится она с данными или нет. Проверять равенство 0 всех возможных корреляций - это вероятно слишком дофига. К счастью, частичные корреляции не независимы друг от друга: можно выбрать небольшой базис корреляций, который будет достаточен для определения их всех.

def Базис Пусть S - множество частичных корреляций. Базис B для S - это множество нулевых частичных корреляций, из которых выводится 0 всех элементов S, и никакое подмножество B таким свойством не обладает.

Очевидный выбор базиса - множество $B = \{\rho_{ij\cdot pa_i} = 0 | j < i\}$, где i пробегает все вершины в D. Это отражает свойство марковости марковских моделей (кек). Так как это все вероятностные данные, закодированные в DAG, то их проверки достаточно для проверки всех статистических утверждений о линейной марковской модели.

Эквивалентность моделей

В стандартных структурных моделях уравнений предполагается линейная зависимость между переменными, и данные характеризуются матрицами ковариаций. Две таких модели являются наблюдаемо эквивалентными, если они ковариационно эквивалентны, то есть если любая матрица ковариаций, генерируемая одной моделью, мб сгенерирована и другой.

Теорема 5.2 *Две марковские линейно-нормальные модели ковариационно эквивалентны тогда и только тогда, когда они имеют один и тот же список нулевых частных корреляций. Более того, две таких модели ковариационно эквивалентны ⇔ у соответствующих им графов один и тот же набор ребер и v-структур.*

В полумарковских моделях правила определения эквивалентности моделей более сложные. Будем рассматривать любое двунаправленное ребро $X \leftrightarrow Y$ как общую скрытую причину $X \leftarrow L \rightarrow Y$. Достаточным условием эквивалентности моделей все еще будет одинаковость ребер и v-структур - но необходимым это уже не будет - условия послабее в ту сторону.

Конкретно ситуация в том, что теперь у нас есть латентные переменные, и из-за этого создание и разрушение некоторых v-структур может быть допустимым, если оно не меняет множество нулевых частичных корреляций. Для полумарковских моделей надо уточнить понятие v-структуры как структуры с двумя сходящимися стрелками, начала которых *d-разделимы.*

Графы и идентифицируемость

Идентификация параметров в причинных моделях

Рассмотрим направленное ребро $X \rightarrow Y$ в диаграмме SEM, снабженное коэффициентом α. Известно (??), что коэффициент регрессии $r_{YX} = \rho_{XY}\frac{\sigma_Y}{\sigma_X}$ может быть представлен в виде суммы

$$r_{YX} = \alpha + I_{YX} \tag{5.18}$$

где I_{YX} не является функцией α. Почему?

Ну, потому что мы по сути при регрессии пытаемся наилучшим образом приблизить $y \approx rx$, то есть минимизировать $\mathbb{E}[(\alpha x + \varepsilon - rx)^2]$ относительно r, где ε это все прочие слагаемые, определяемые причинной диаграммой. Далее, представив $r = \alpha + I_{YX}$, получаем

$$\begin{aligned} I_{YX} &= \operatorname*{argmin}_{I_{YX}} \mathbb{E}[(I_{YX} - \varepsilon)^2] \\ I_{YX} &= \frac{\sigma_{X\varepsilon} + \mu_x\mu_\varepsilon}{\sigma_x^2 + \mu_x^2} = \frac{\sigma_{x\varepsilon}}{\sigma_x^2} \end{aligned} \tag{5.19}$$

Где последний переход сделан в предположении, что матожидания переменных равны 0.

Таким образом, если окажется, что после удаления ребра $X \rightarrow Y$ у нас нулевая корреляция между X, Y, то значит $I_{YX} = 0$ и соответственно $\alpha = r_{YX}$ идентифицируемо.

В общем случае, если I_{YX} не 0, но его можно сделать 0, если обусловиться на некоторое множество переменных Z, которые лежат на различных d-путях в графе из X в Y, то получится опять то же разбиение $r_{YX\cdot Z} = \alpha + I_{YX\cdot Z}$, где $I_{YX\cdot Z}$ не зависит функционально от α при условии, что в Z нет наследников Y. $I_{YX\cdot Z}$ представляет частичную корреляцию между X, Y при установке $\alpha = 0$, то есть при удалении ребра $X \rightarrow Y$ из диаграммы. Если в итоге Z d-разделяет X, Y, то $I_{YX\cdot Z} = 0$ и соответственно α определяется через соответствующий коэффициент частичной регрессии, то есть α можно вычислить из регрессии

$$y = \alpha x + \beta_1 z_1 + ... + \beta_k z_k + \varepsilon \tag{5.20}$$

Таким образом, мы получаем теорему

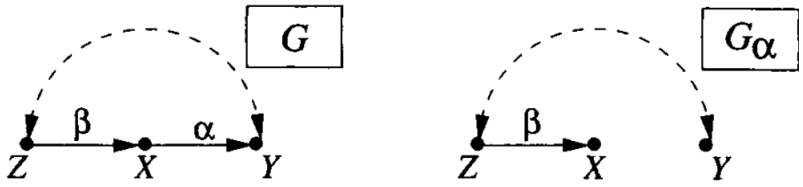


Рис. 5.1: α можно оценить как $r_{YX \cdot Z}$

Теорема 5.3 ***Критерий единственной двери для прямых эффектов*** *Пусть G - любая путевая диаграмма, где α - коэффициент, навешенный на ребро $X \to Y$, а G_α - диаграмма, полученная удалением ребра $X \to Y$ из G. Коэффициент α идентифицируем если существует множество Z:*

(i) $de(Y) \cap Z = \emptyset$

(ii) Z d-разделяет X, Y в G_α

В этом случае $\alpha = r_{YX \cdot Z}$

Обратно, если Z не удовлетворяет этим условиям, то $r_{YX \cdot Z}$ не является консистентной оценкой α.

Может сложиться ситуация, что прямой эффект X на Y оценить невозможно, как например в случае диаграммы 5.2. Однако, может быть все еще возможно оценить полный эффект. Чтобы его идентифицировать, необходимо выбрать такое множество Z, чтобы никакие пути, кроме направленных $X \rightsquigarrow Y$, не давали вклад в $r_{YX \cdot Z}$, то есть были заблокированы Z.

Соответственно, имеем теорему

Теорема 5.4 *Для любых двух переменных X, Y полный причинный эффект X на Y идентифицируем, если*

(i) $Z \cap de(X) = \emptyset$

(ii) Z d-разделяет X, Y в $G_{\underline{X}}$, то есть блокирует все бэкдор-пути их X в Y

Если эти два условия выполнены, то полный эффект оценивается как $r_{YX \cdot Z}$

По аналогии с этими двумя экстремальными случаями - полного эффекта и прямого эффекта, в общем можно ввести понятие частичного эффекта, который определяется как часть эффекта X на Y, действующего через некоторое подмножество направленных путей. Вычислять его по аналогии можно, если заблокировать все прочие пути (направленные и ненаправленные) некоторым множеством Z.

Любопытно, что иногда полный эффект не определим сам по себе, но требует определения каждой компоненты, которая его составляет по отдельности. В целом это логично, ведь предыдущая теорема не утверждает, что это *необходимый* критерий идентифицируемости полного эффекта - что логично, ведь не только бэкдор критерием ранее мы научились проводить идентификацию эффектов.

Например, в случае 5.1, мы не можем используя последнюю теорему оценить эффект Z на Y, так как нам никак не заблокировать бэкдор-путь $Z \rightsquigarrow Y$. Однако, мы можем оценить полный эффект Z на X: $\beta = r_{XZ}$, и мы можем оценить полный эффект X на Y - $\alpha = r_{YX \cdot Z}$, откуда можно выразить полный эффект Z на Y равный $\alpha\beta$.

Другой пример, когда надо ухищряться - на картинке 5.3, где α не оценить напрямую, но можно оценить эффект Z на X: $\beta = r_{XZ}$, и эффект Z на Y: $\alpha\beta = r_{YZ}$, откуда $\alpha = \frac{r_{YZ}}{r_{XZ}}$

Сравнение с непараметрическим подходом

Мощность параметрических методов конечно больше, чем у непараметрических: так, обычно в непараметрическом случае идентификация возможна, а в параметрическом точно определить тип функциональной зависимости не получается.

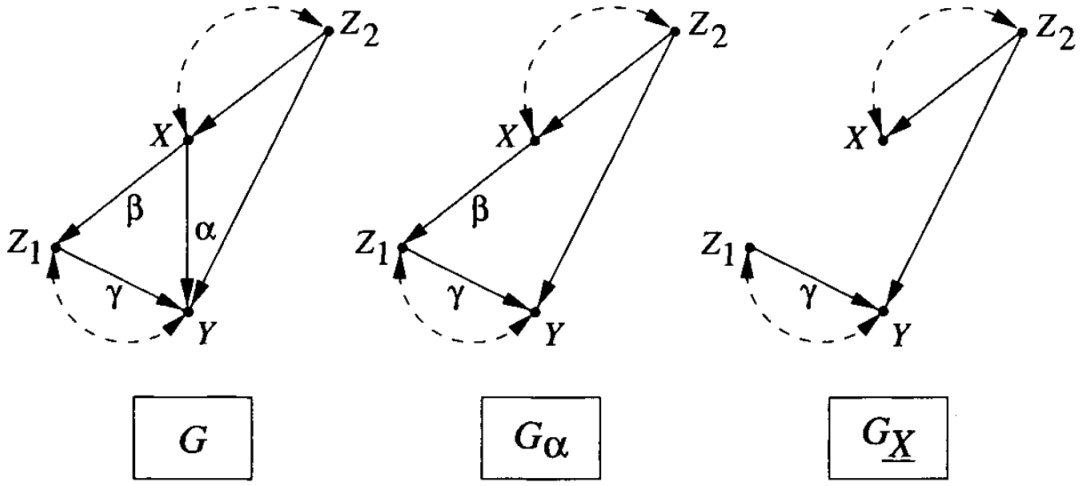


Рис. 5.2: α не идентифицируемо, но полный эффект X на Y, равный $\alpha + \beta\gamma$ оценить можно как $r_{YX|\cdot Z_2}$

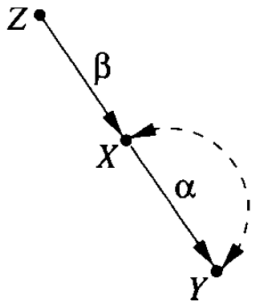


Рис. 5.3: α определяется с использованием инструментируемой переменной Z

Однако, идентификация - это не самоцель, и иногда не необходима. Например, если нужно решать только задачи предсказания, без интервенций, то в целом все необходимое закодировано в матрице ковариаций, и нам в общем неважны конкретные значения функций f_i, определяющих переменные функционально.

Опять повторим, что смысл структурных уравнений в том, что они описывают физические законы, по которым та или иная переменная генерируется, и равенство в этих моделях не симметричное. Для предсказания эффекта интервенций важно, чтобы у нас была корректная модель генерации данных.

Некоторые концептуальные замечания

Что в действительности значат структурные коэффициенты?

Рассмотрим SEM $y = \beta x + \varepsilon$. Если мы проинтерпретируем β как изменение в $E[Y]$ при единичном изменении X, то может появиться соблазн написать $x = \frac{y-\varepsilon}{\beta}$ и проинтерпретировать в свою очередь $\frac{1}{\beta}$ как то, насколько изменится $E[X]$ при единичном изменении Y. Но это конфликтует и с интуицией, и с предсказанием модели: изменение в $E[x]$ должно быть 0 при изменении Y, если Y не является

независимой переменной в оригинальном, структурном уравнении для X.

Есть два подхода к тому, чтобы разобраться с этой ситуаций, но они оба хреновые. Первый состоит в том, чтобы вообще не наделять β причинным смыслом, а только говорить, что β - это мера уменьшения дисперсии Y, объяснимого через X, но это вступает в конфликт с желанием использовать SEM для выбора политик поведения и в целом с интуицией. Второй подход - в том, чтобы наложить на ε ограничение, что он не скоррелирован с X. А типа тогда он будет скоррелирован с Y, и симметрия пропадает. Но в этом случае мы сталкиваемся с проблемой, что как Y можно определить в терминах такого $y = \beta x + \varepsilon_1$, так и симметрично $x = \alpha y + \varepsilon_2$, $\alpha = \beta \frac{\sigma_X^2}{\sigma_Y^2}$, где $cov(x, \varepsilon_1) = cov(y, \varepsilon_2) = 0$, и тогда мы вынуждены наделять как β, так и α причинным смыслом так как они оказываются равноправны - но это тоже не сочетается с интуицией, состоящей в том, что X не должен реагировать на интервенции в Y, если в структурном уравнении для X его не было.

В чем же тогда смысл структурных уравнений, коэффициента β, и слагаемого ошибки? Ответ дает следующее определение:

def Операционное определение SEM уравнение $y = \beta x + \varepsilon$ называется структурным, если его интерпретируют следующим образом: в идеальном эксперименте, где мы устанавливаем $X = x$, а любые другие переменные Z не содержащие X, Y установлены в некоторое фиксированное значение z, значение y определяется согласно структурному уравнению, при этом ε не является функцией x, z.

Заметим, что операционное определение не утверждает что-либо о поведении X при фиксировании Y: если мы наблюдаем X и Y, то симметрия есть: из наблюдения $Y = 0$ можно заключить $X = -\frac{\varepsilon}{\beta}$, но когда у нас появляются интервенции, симметрии уже нет - это очевидно, когда мы смотрим на диаграмму соответствующую уравнениям, там стрелки явно позволяют задать эту асимметрию.

Самое сильное утверждение, которое следует из SEM $y = \beta x + \varepsilon$ состоит в том, что Y причинно зависит *только* от X, таким образом, $P(y|do(x), do(z)) = P(y|do(x))$, то есть любые интервенции, не влияют на Y, если мы контролируем X.

def Операционное определение структурных параметров Смысл коэффициента β состоит в

$$\beta = \frac{\partial}{\partial x} E[Y|do(x)] \tag{5.21}$$

При этом в данной интерпретации неважно, скоррелированы ли X, ε или нет. Уже обсуждено ранее, что β в общем случае не равно коэффициенту регрессии, и не связано напрямую с $E[Y|x]$.

Наконец, операционное определение шума есть

$$\varepsilon = y - \mathbb{E}[y|do(x)] \tag{5.22}$$

Таким образом, шум можно оценить и по контролируемым данным (если мы проведем серию экспериментов с $do(x)$ и запишем данные Y, а дальше вычислим матожидание как выборочное среднее), и по наблюдаемым данным по формуле $\varepsilon = y - \beta x$, если мы знаем β, ведь βне меняется от того, наблюдаем мы или делаем интервенцию.

Аналогично, можно оценить и ковариации шумов для любых двух несмежных переменных X, Y:

$$\mathbb{E}[\varepsilon_Y \varepsilon_X] = E[YX|do(pa_X), do(pa_Y)] - E[X|do(pa_X)]E[y|do(pa_Y)] \tag{5.23}$$

ну тут тоже, если мы уже установили коэффициенты модели, то справа у нас будут известные линейные функции, и можно все посчитать. Если не установили - проводим контролируемые эксперименты с фиксацией PA_X, PA_Y.

6 Парадокс Симпсона, конфаундинг и схлопываемость

Анатомия парадокса Симпсона

В данной секции разберемся, по какой причине вообще эффект, описанный Симпсоном в 1951 году (но впервые на который обратил внимание Пирсон в 1899) вообще воспринимается как парадокс.

Парадокс состоит в следующем: пусть есть некоторое событие C, которое увеличивает вероятность события E в популяции, но в то же время, уменьшает вероятность события в каждой из двух подпопуляций F и $\neg F$:

$$P(E|C) > P(E|\neg C) \tag{6.1}$$
$$P(E|C,F) < P(E|\neg C,F) \tag{6.2}$$
$$P(E|C,\neg F) < P(E|\neg C,\neg F) \tag{6.3}$$

С точки зрения простой теории вероятности, в такой ситуации нет ничего удивительного; парадоксальным она начинает казаться, когда мы наделяем вероятностные соотношения дополнительным *причинным* смыслом.

Например, если мы станем подразумевать под C прием определенного препарата, под E - исход лечения болезни, а под F - пол (женский/мужской) то при придании причинного смысла выражениям, можно прийти к выводу, что и для женщин, и для мужчин лекарство вредно, но в целом оно полезно.

В чем же дело? А в том, что $P(E|C)$ не следует воспринимать в качестве эффекта от лечения: это то, что мы пассивно наблюдаем в выборке; эффект же от лечения (от действия, совершаемого доктором) выражается как $P(E|do(C))$. $P(E|C) > P(E|\neg C)$ в свою очередь - это всего лишь некоторое подтверждение того, что C может положительно влиять на излечение - либо в силу причинного эффекта, либо из-за случайного кофаундинга через общие причины.

	Combined	E	$\neg E$		Recovery Rate
(a)	Drug (C)	20	20	40	50%
	No drug ($\neg C$)	16	24	40	40%
		36	44	80	
	Males	E	$\neg E$		Recovery Rate
(b)	Drug (C)	18	12	30	60%
	No drug ($\neg C$)	7	3	10	70%
		25	15	40	
	Females	E	$\neg E$		Recovery Rate
(c)	Drug (C)	2	8	10	20%
	No drug ($\neg C$)	9	21	30	30%
		11	29	40	

Рис. 6.1: Пример данных для парадокса Симпсона

В данном случае, если считать что пол является общей причиной для принятия лекарства и для выздоровления, эффект лекарства следует оценивать взвешенно для мужчин и женщин. Таким образом, если пол F - единственный конфаундинг-фактор, то 6.2, 6.3 корректно отображают эффективность лекарства в соответствующих подпопуляциях, а 6.1 обозначает его доказательный вес в случае отсутствия информации о поле (грубо говоря,)

В статистической литературе особенно не связывали парадокс с причинностью, потому что её воспринимали чисто воображаемым конструктом.

Что характерно, за все время существования парадокса статистики толком не разбирались, а почему собственно это обращения знаков неравенства вызывает такой когнитивный диссонанс у людей: в конце концов, нет ничего противоестественного, казалось бы, в том что вероятности при обуславливании ведут себя так, а не иначе.

В 1981 году Lindlie и Novick первыми показали нестатистический характер парадокса Симпсона. По традициям байесовской теории принятия решения, они решили ответить на практический вопрос: приходит новый пациент - надо ли нам использовать лекарство или нет? Непосредственный ответ "Если мы знаем пол пациента - то лекарство давать не нужно, а если не знаем - то лекарство давать нужно звучит странно и конечно не верен по сути.

Следующим шагом ребята спросили себя - есть ли какие-то дополнительные статистические данные, которые могли бы нас склонить в пользу раздельной или совместной таблицы для принятия решения. На этот вопрос они ответили негативно: с одними и теми же данными в зависимости от семантики может оказаться правильным выбор как одной таблицы, так и другой: так, если семантика F была в том, что это пол - то надо принимать мнению согласно статистике, обусловленной на пол. Если же F - это кровяное давление, как на картинке 6.2 (b), на которое влияет лекарство, а оно влияет на излечение, то конечно надо ориентироваться на необусловленные на F данные - а обуславливаться на следствие C будет неправильным, ведь мы при этом перекроем один из двух причинных путей.

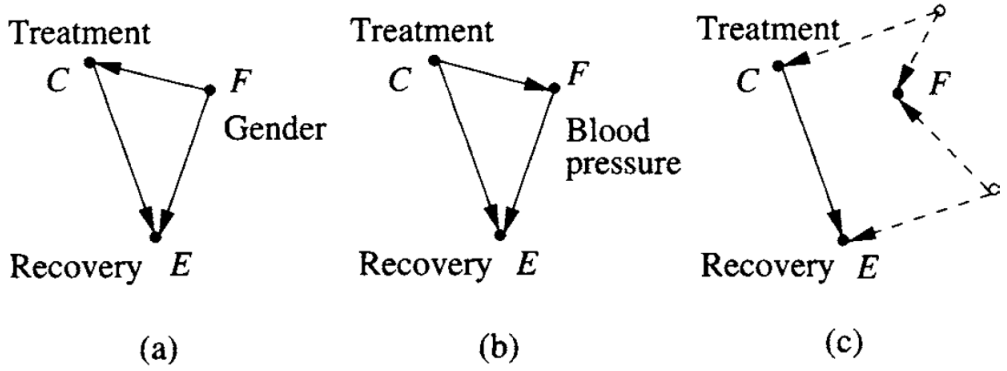


Рис. 6.2: Выбор, согласно каким данным принимать решение - обусловленным на F или нет - зависит от подлежащей причинной структуры

Какова же причина парадокса? Вообще, парадокс возникает, когда наше множество неявных имеющихся предположений оказывается в противоречии друг с другом. В случае парадокса Симпсона, у нас есть два конкурирующих убеждения

1. Предположение, что причинные отношения управляются исчислением вероятностей
2. Множество неявных предположений, которые описывают нашу причинную интуицию

Первое предположение подтверждается явными примерами данных, когда неравенства 6.1-6.3 оказываются согласованными. Второе говорит, что не существует волшебного лекарства, которое вредно и для мужчин, и для женщин - но полезно для человека в целом, если мы не знаем его пол.

Чтобы разрулить парадокс, надо либо отказаться от причинной интуиции, либо от предположения, что причинные связи управляются законами стандартного исчисления вероятностей. Конечно же, мы

выберем второй вариант - причинность действует по своим законам и требуется расширение вероятностного исчисления для её корректного описания.

Собственно, для этого привлечем механику do-исчисления. Покажем, что если

$$P(E|do(C), F) < P(E|do(\neg C), F) \tag{6.4}$$
$$P(E|do(C), \neg F) < P(E|do(\neg C), \neg F) \tag{6.5}$$

то не может оказаться $P(E|do(C)) > P(E|do(\neg C))$, если F - пол.

Для этого докажем теорему

Теорема 6.1 *Если действие C уменьшает вероятность события E в любой подпопуляции, то оно должно уменьшать вероятность события E в популяции в целом при условии, что действие не меняет распределение подпопуляций.*

Докажем для частного случая, когда подпопуляций всего две. Итак, пусть 6.4 и 6.5. Раз действие не влияет на распределение подпопуляций, то

$$P(F|do(C)) = P(F|do(\neg C)) = P(F) \tag{6.6}$$

Разложим $P(E|do(C))$ по подпопуляциям:

$$\begin{aligned} P(E|do(C)) &= P(E|do(C), F)P(F|do(C)) + P(E|do(C), \neg F)P(\neg F|do(C)) \\ &= P(E|do(C), F)P(F) + P(E|do(C), \neg F)P(\neg F) \end{aligned} \tag{6.7}$$

Аналогично, разложим для $P(E|do(\neg C))$:

$$P(E|do(\neg C)) = P(E|do(\neg C), F)P(F) + P(E|do(\neg C), \neg F)P(\neg F) \tag{6.8}$$

Так как оба множителя в первом уравнении меньше обоих множителей во втором, то и сумма в итоге будет в первом случае меньше, чем во втором.

■

Наша причинная интуиция возникает из очевидного предположения, что лекарство не может влиять на пол. В то же время, когда F - давление, мы уже не можем быть уверенными, что лекарство на него не влияет.

Конфаундинг

По смыслу конфаундинг с точки зрения статистики - это ситуация, когда ассоциация между двумя переменными присутствует не только вследствие эффекта, которым одна переменная действует на другую, но и вследствие наличия каких-то общих причин. Таким образом, концептуально ассоциация между X и Y законфаужена, если существует третья переменная Z, которая влияет и на X, и на Y.

Звучит просто, но на деле долгое время не получалось перенести эту концепцию на математический язык: непонятно было, как формализовать *эффект* переменной на другую.

Есть эмпирическое описание эффекта как ассоциации, которая будет *превалировать* в случае контролируемого рандомизированного эксперимента - но такое описание не перенести на язык теории вероятностей, потому что эта теория подразумевает статические условия применения, и не позволяет предсказывать, какие отношения будут доминировать при *изменении* этих условий.

Именно для математического описания эффектов в первую очередь и был разработан используемый нами do-оператор, благодаря которому удается выразить эффект в виде $P(y|do(x))$.

Несмотря на сложности, статистики, экономисты и эпидемиологи долгое время пытались разработать формальные критерии наличия конфаундинга, и далее мы рассмотрим, чем они плохи и как их можно починить.

Начнем с причинного определения конфаундинга:

def Незаканфауженность (причинное определение) пусть M - причинная модель процесса, генерирующего данные. Обозначим через $P(y|do(x))$ вероятность $Y = y$ при совершении интервенции $X = x$. Мы говорим, что X, Y не законфаужены в M если

$$P(y|do(x)) = P(y|x) \tag{6.9}$$

для любых x, y в их области определения, где $P(y|x)$ - условное распределение, генерируемое M.

Именно это определение мы будем использовать для описания незаконфауженности. Когда у нас нет конфаундинга X и Y, говорят, что $P(y|x)$ несмещенная оценка эффекта.

def Незаканфауженность (ассоциационное определение) пусть T - множество переменных, на которые *не влияет* X. Будем говорить, что X и Y незаконфаужены в присутствии T если $\forall Z \in T$ выполняется хотя бы одно из двух:

(U1) Z не ассоциирована с X: $P(x|z) = P(x)$

(U2) Z не ассоциирована с Y при обуславливании на X: $P(y|z, x) = P(y|x)$.

Заметим, что даже такое казалось бы чисто статистическое определение тем не менее содержит загадочный с точки зрения теорвера концепт *влияние одной переменной на другую*. На самом деле, это причинная концепция, но без неё никуда не деться при изучении эффектов, поэтому везде далее в статистическом подходе к конфаундингу мы тем не менее будем считать что исследователь как-то знает, какие переменные на какие влияют.

Когда фейлится ассоциативный критерий конфаундинга

Фейлим достаточность ассоциативного критерия

Рассмотрим такой пример: Z_1, Z_2 - исход бросания двух честных монеток. $X = Z_1 == Z_2$, $Y = Z_1! = Z_2$. Ясно что по отдельности Z_1, Z_2 не конфаундят X, Y: $P(x|z_1) = P(x|z_2) = 0.5 \forall x, z_1, z_2$, то есть выполняется U_1. Однако, пара (Z_1, Z_2) конфаундит X, Y: несмотря на отсутствие причинной связи, они оказываются отрицательно скоррелированы в таком контексте.

Другая проблема с ассоциативным определением - нам надо перечислить *все* переменные, которые не аффектятся X. Если мы какую-то не измерим, то мы можем счесть, что конфаундинга нет, когда он есть - и вообще говоря нет никакого способа проверить что мы учли все переменные.

Фейлим необходимость ассоциативного критерия

Не сложно привести пример, когда переменная Z фейлит оба условия $U1, U2$ но при этом конфаундинга между X, Y нет. Например, рассмотрим 6.3.

Стабильная и нестабильная несмещенность

def Стабильная несмещенность пусть A - множество предположений/ограничений, наложенных на процесс генерации данных, C_A - класс причинных моделей, удовлетворяющих A. Оценка эффекта X на Y стабильно несмещенная, если $P(y|do(x)) = P(y|x)$ соблюдается $\forall M \in C_A$. Соответственно, пара переменных X, Y называется стабильно незаконфауженной.

Предположения, которые обычно ограничивают класс рассматриваемых причинных моделей, представляются либо параметрически, либо топологически. Например, в случае моделей, заданных структурными уравнениями, ограничения могут состоять в том, что мы рассматриваем только линейные модели с нормальным шумом.

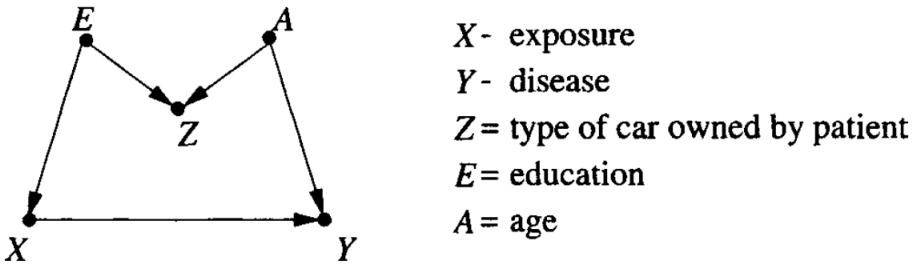


Рис. 6.3: Переменная Z фейлит оба условия ассоциативного критерия незаконфауженности, но при этом не приводит к конфаундингу

Более слабые ограничения накладываются топологической структурой причинной диаграммы - при этом мы не задаем распределения шума и функциональную форму связи переменных. Рассмотрим статистические следствия таких непараметрических ограничений.

Как мы знаем, когда диаграмма ациклична, бэкдор-критерий задает необходимый и достаточный тест на незаконфауженность. Таким образом, получаем теорему

Теорема 6.2 ***Принцип общей причины*** *Пусть A_D - множество ограничений, задаваемых ациклической причинной диаграммой D. Переменные X, Y стабильно незаконфаужены при заданном A_D тогда и только тогда, когда у X, Y нет общего предка в D.*

Если нет общего предка ($\Leftarrow$), то законфауженности нет по бэкдор критерию. То что предка нет, если нет законфауженности ($\Longrightarrow$) доказывается предоставлением конкретного примера распределения, которое это условие нарушает (вроде это несложно).

■

Допустим однако, что у нас нет под рукой причинной диаграммы. Пусть все что мы знаем про любую переменную Z - точно ли она не имеет эффект на Y, и имеет ли X на эффект на неё. Вопрос - хватит ли такой, урезанной информации для выявления наличия или отсутствия конфаундинга?

Теорема 6.3 ***Критерий стабильного отсутствия конфаундинга*** *Пусть A_Z - предположения, состоящие в том, что данные сгенерированы*

(i) ацикличной (хоть и неизвестной) моделью M

(ii) Z - переменная в M, которая не аффектится X (нет направленного пути $X \rightsquigarrow Z$) и возможно аффектит Y (это значит, что в подлежащей диаграмме есть направленный путь $Z \rightsquigarrow Y$).

Тогда, если оба условия U1, U2 ассоциатичного определения законфауженности зафейлены, то есть если (U1) $X \not\perp\!\!\!\perp Z$ и (U2) $Z \not\perp\!\!\!\perp Y|X$, то X, Y стабильно законфаужены.

Ну действительно, по предыдущей теореме, если переменные незаконфаужены, у них нет общего предка. Но раз у них нет общего предка, то должно выполняться либо U1, либо U2 для любой Z, удовлетворяющей A_Z.

■

Таким образом, нашли достаточный, но не необходимый статистический критерий законфауженности.

7 Логика структурных контрфактов