

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

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

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

Можно показать, что при заданном упорядочивании переменных марковские родители для каждой переменной определяются однозначно, если распределение $P(X_1, \dots, X_N)$ 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 $P(x_n | x_1 \dots x_{n-1}) = \frac{P(x_1, \dots, x_n)}{P(x_1, \dots, x_{n-1})}$, так как в этом случае знаменатель не будет нигде обращаться в 0 на области определения $P(x_1, \dots, x_N)$.

def Марковская согласованность (Markov Compatibility) - говорят что распределение P марковски согласовано 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 $a \rightarrow b \rightarrow c : b \in Z$
2. Он содержит вилку $b \rightarrow a, b \rightarrow c : b \in Z$
3. Он содержит v -структуру с вершиной, которая не в Z и все наследники которой тоже не в Z : $a \rightarrow b, c \rightarrow b, b \notin Z, de(b) \cap Z = \emptyset$

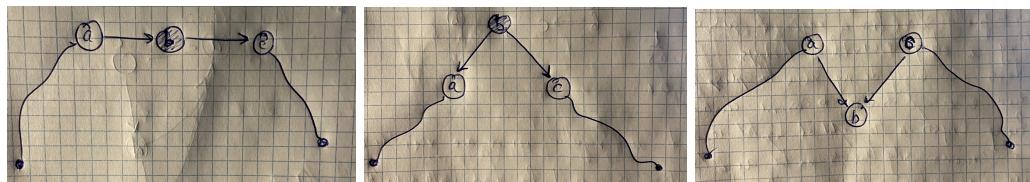


Рис. 1: Различные причины d -сепарации, заштрихованные вершины $\in Z$

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: вероятностные следствия d -сепарации

Если множества X и Y d -разделены множеством Z , то $X \perp\!\!\!\perp Y | Z$ 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 : $X \not\perp\!\!\!\perp Y | Z$ 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 - неперескающиеся 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: $I(X, Y, Z) \iff X \perp\!\!\!\perp Y | Z$

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 $\sum_w P(X, YW|Z) = P(X, Y|Z)$.

Для правой части аналогично

$$\sum_w P(X|Z)P(YW|Z) = P(X|Z) \sum_w P(YW|Z) = P(X|Z)P(Y|Z) \quad (1)$$

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

$$P(X, W|Z) = P(X|Z)P(W|Z) \quad (2)$$

Запишем факторизацию

$$P(X, Y, Z, W) = P(X, YW|Z)P(Z) = P(X|Z)P(YW|Z)P(Z) = P(X|Z)P(Y|ZW)P(W|Z)P(Z) \quad (3)$$

$$\begin{aligned} P(X, Y|ZW) &= \frac{P(X, Y, Z, W)}{P(Z, W)} = \frac{P(X|Z)P(Y|ZW)P(W|Z)P(Z)}{P(ZW)} = \frac{P(X, W|Z)P(Y|ZW)P(Z)}{P(ZW)} \\ &= \frac{P(X, Z, W)P(Y|ZW)}{P(ZW)} = P(X|ZW)P(Y|ZW) \end{aligned} \quad (4)$$

4. Пусть

$$P(X, W|ZY) = P(X|ZY)P(W|ZY) \quad (5)$$

$$P(X, Y|Z) = P(X|Z)P(Y|Z) \quad (6)$$

Рассмотрим $P(X, YW|Z)$:

$$\begin{aligned} P(X, YW|Z) &= \frac{P(X, Y, Z, W)}{P(Z)} = \frac{P(X, W|ZY)P(ZY)}{P(Z)} = \frac{P(X|ZY)P(W|ZY)P(ZY)}{P(Z)} \\ &= P(X|ZY)P(W|ZY)P(Y|Z) = P(X|ZY)P(Y, W|Z) = \frac{P(X, Y|Z)}{P(Y|Z)}P(Y, W|Z) \quad (7) \\ &= \frac{P(X|Z)P(Y|Z)}{P(Y|Z)}P(Y, W|Z) = P(X|Z)P(YW|Z) \end{aligned}$$

Ну в общем, вроде всё верно :)

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 экспоненциально 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: $I(X, Z, Y \cup W) \implies I(X, Z, Y)$
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: $I(X, Z, Y) \implies I(X, Z, W) \vee I(Y, Z, W) \forall W \not\supseteq X \cup Y \cup Z$

Инференс в байесовских сетях

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

Теорема: наблюдаемая эквивалентность

Два графа наблюдаемо эквивалентны тогда и только тогда, когда они имеют один и тот же скелет и набор v -структур.

Таким образом, наблюдаемая эквивалентность определяет границы, в рамках которых возможно определение ориентаций в байесовской сети.

A Theory of Inferred Causation

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 зависит от B , B зависит от C , но A и C независимы. В таком случае, если немного подумать, выходит, что наиболее простой граф, описывающий такую конфигурацию, выглядит как на 2

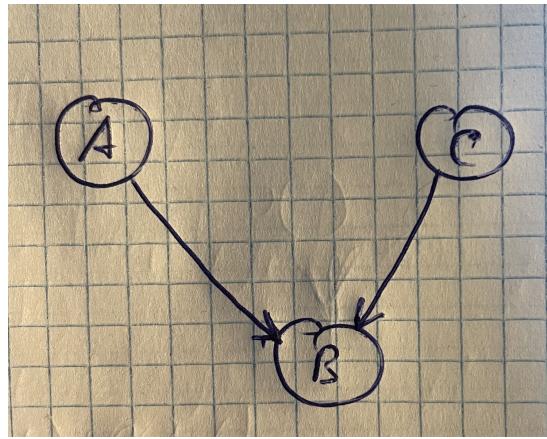


Рис. 2: A, C безусловно независимы, но зависимы при наблюдаемом следствии B

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 ненаблюдаются.

Фреймворк

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 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 .

Шум, влияющий на значение переменных, можно рассматривать например как следствие ненаблюдаемости некоторых переменных, и считается взаимонезависимым: $(U_i \perp U_j)$.

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

Вообще говоря, так как V неизвестно, можно придумать сколь угодно много разных моделей, которые смогут 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 наблюдаемыми переменными в такой модели не будет, причинная структура для такой модели представлена на 3.

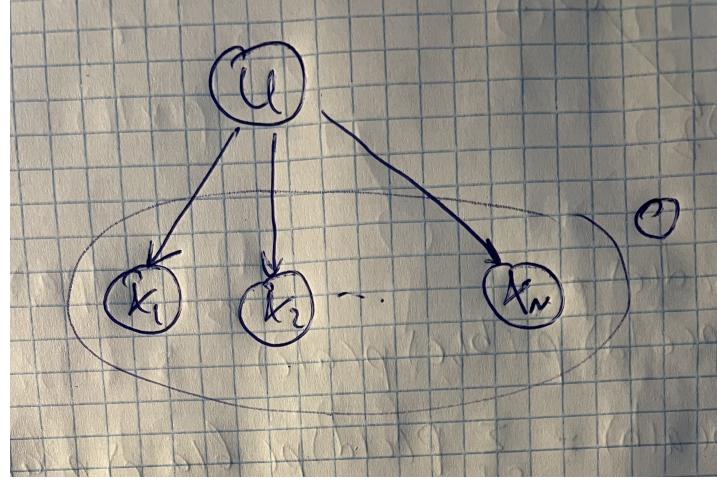


Рис. 3: Довольно бесполезная причинная структура

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

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

def Выведенная причинность (predv.) Переменная X имеет причинное влияние на переменную Y , если существует направленный путь из X в Y в любом минимальной причинной структуре.

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 $L' \preceq L$.

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 за-кодировано 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 4 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.

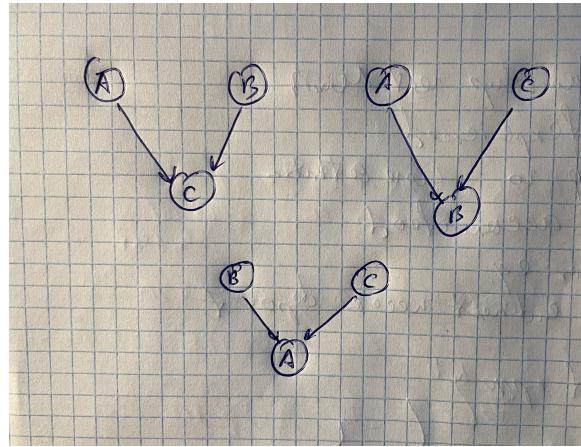


Рис. 4: Какую из трех причинных структур выбрать?

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

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.

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 (экви-

валентные 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 неориентированный граф D на вершинах V . Ребром соединяются любые две вершины a, b такие, что $\nexists S_{ab} \subset V \setminus \{a, b\} : a \perp\!\!\!\perp b | S_{ab}$

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