10-708: Probabilistic Graphical Models, Spring 2020

17: Causality 1

Lecturer: Kun Zhang *Scribe: Ni Zhan, Tianqin Li, Carmel Fiscko, Xuecong Fu*

1 Intro

We want to learn why causality is important, how causal models relate to graph models, how to discover causal info, infer causal effects given causal structure and data, and benefits of knowing causal info.

Causality vs. Dependence

In daily life and scientific discovery, we have to distinguish between causal connections and associations. For example "couples who share housework are more likely to divorce" seems astonishing. However, association doesn't mean causality. In many real world application, we want to see causal connection, not just association.

Causality implies dependence. The idea "causality implies correlation but correlation does not imply causality" is basically correct. Fortunately, we can make use of dependence structure or pattern, and using multiple distributions/variables, we can discover causation from correlation under some assumptions.

Association - if two variables are not independent, one is useful to predict another.

X and Y are **associated** iff $\exists x_1 \neq x_2, P(Y|X = x_1) \neq P(Y|X = x_2)$

Causality

To say x is cause of y, need to go another level. We define causality based on intervention. Let's say you do intervention on x, and for different values of x, the corresponding distribution for y is different, then x causes y.

Def. Intervention (on x) - only change x without changing any other variable.

X is a **cause** of Y iff $\exists x_1 \neq x_2, P(Y|do(X = x_1)) \neq P(Y|do(X = x_2))$

Since all other variables have same value, we can say the difference in y is because of x.

Examples

- Causal relation between hot weather and high sales of ice cream, are they causally related? If you can make it very hot, you can see high sales of ice cream. However you can intervene on B (ice cream sales) without changing A (hot weather), by hiring a lot of people to buy ice cream. We can have noise in the variables, and A and B not necessarily deterministically related in this case.
- Ex 2. A guy goes to work by bus around 8:30. At the same time, the bus is coming. Variable-1 is guy leaving home 0 or 1, variable-2 is bus coming. Variables are clearly correlated, but are they causally related? Intuitively, no because the man does not see the bus coming. Now we do intervention by changing target variable, jump in front of bus while it is incoming to make bus stop, action will not change anything else in system. In this case, man would probably still leave. There could be a common cause such as bus timetable. This example shows we really have to pay attention to definition of intervention. If we change bus timetable to change if bus is coming, it is not intervention because it will probably also change if guy leaves home at that time.

We can use directed graphs for causal relations. In a DAG, direction from a to b means a causes b.

1.1 Outline

- why benefit from causal thinking.
- difference between causal graph model and graph models.
- identification of causal effects, counterfactual reasoning.

2 Causal thinking

We have to distinguish between dependence and causality. If you want to make prediction, dependence is enough. If you want to change the system, then must care about causality.

Examples

- If we want to change incidence of cough, knowing dependence with yellow fingers does not help. We need to go to cause, which is smoking.
- **Simpson's paradox:** (based on a real dataset published in 1970s, Figure 1) There is a hospital with two groups of patients with kidney stones: large and small, and two treatments A and B. For each group, treatment A is better. If you merge the data, treatment B seems better. If you are doctor and want to maximize recovery, would you recommend A or B? In principle, even if you don't know the size of kidney stone, should still pick A. Stone size influences treatment and recovery, stone size is a common cause. Because common cause is there, you cannot correctly see causal inference of treatment on recovery from the dependence pattern. We will come back to this example later.

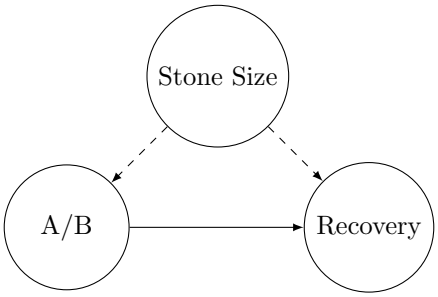

Figure 1: Simpson's Paradox.

- Cholesterol vs. exercise per age group. Separating by age, cholesterol decreases when exercise increases. If you ignore the age, more exercise = more cholesterol, different conclusion than when you observe ages. If you want to make recommendation about exercise and cholesterol, have to go to causal level.
- **Dependencies in data.** Going back 60 years in data, female college students smarter than male. This happened essentially because of selection. Admittance into college was dependent on gender and IQ. If you look at whole population, the "female smarter than male" pattern will disappear. Need to look for and correct selection bias.
- **Survivorship bias.** In world war 2, we put armor on airplanes to increase chance of survival, and looked at holes on returning flights. Where should I put armor? "Being hit" and "where holes are" are causes of survival. We can observe $probability(holes|survived)$. We want to infer pattern of holes that did not survive.
- **Monty hall problem.** We want to increase chance of winning money. Location of money is independent of your initial choice. "Which door is opened by host" is common effect. Two originally

independent variables ("initial choice" and "where money is") are now dependent. Now initial choice has some info about location of money, so I should change my mind of door selection.

Causal thinking makes a difference. If we know causality, we can do:

- active manipulation
- generalization / adaptation in new environments
- info integration. x can cause z through y. Contrast this with dependence: If you only know dependence, you cannot integrate information in this way. (For example, x and y are dependent, and y and z dependent, x and z could be correlated, identical, independent, etc. Only knowing dependence, you do not know relationship between x and z.)

Causality is creativity problem, we need to ask "what if", and integrate information from different aspects.

Ex. Causality for prediction

- **Figure and shadow.** In first scenario, you only see shadow, can you say anything about figure? Second scenario: you only see figure, can you say anything about shadow? Which one is easier?
- Suppose you write down 5, and ask someone to predict if it is a 5? Easy. Then ask someone to predict how I handwrite 5? Hard.
- Prediction is not symmetric, because there is an environment change.
- The figure is cause of shadow.
- Effect contains information about cause.
- If I have no idea about environment, I cannot predict effect from cause, but it is easier to predict cause from effect (This is not rigorous).
- We see that it is easier to predict the figure given shadow, than the shadow given figure, if we do not know about the environment (light source, time of day, etc.).

We can make use of some dependence pattern in data to make predictions.

Change of environment is related to a distribution shift problem. Using causality, we can understand system and causal process, identify source of change, then make prediction.

Problem with google photos

Google photos classified African people as gorillas. Two and half years later, google claimed they solved problem by removing "gorillas" from set of possible labels.

Adversarial Attack

Add a little noise, the prediction changes from panda to gibbon. Machines use a different process than humans to make prediction. If human and machine processes are different, then it is always possible to change something so that the machine decision changes but ours stays the same, or vice versa. This is called adversarial attack. Are human and machine processes consistent in some way? Human process does not look at pixels, we look at high level features and our experience. We want to enable machines to think in a human-like manner.

Artificial Intelligence

We usually assume a fixed distribution of data. We train then apply same model to test. In most cases distribution is not fixed. If you play badminton well, you can play ping pong well. A human driver can make right decision in new scenario. But this is hard for machines.

Desribe intelligence:

- able to understand system
- because of understanding, can do control, intervention

- decompose complex task into small task
- information theory: combine small pieces into big picture
- learn from a small number of examples because of understanding
- interpolation and extrapolation

How can we achieve these things for machines?

How to achieve intelligence?

One way is to examine how brain works, and develop some system. We want to find principles underlying intelligence. From evolution perspective, there are two properties: humans survived and prospered. We have a good ability to make prediction across scenarios (otherwise hard to survive). We have an inner compact representation, allowed us to explain observations, explain connection between scenarios, tasks, etc. We are creative, we can change the world (intervention). We used causal info about world.

3 Causal Graphical Models

D-separation

review on your own.

What does d mean in d-separation? originally means directional. Separation criteria is different for undirected graphs.

Local and global markov condition: They are same for DAGs.

Causal Bayesian Networks Interpreting causality using graph models and conditional distributions.

The causal DAG describes effects of intervention, and you can get the distribution of variables resulting from intervention. Given a DAG, you may modify DAG to represent an intervention. How can you know when a DAG represents causal relationship? DAG is causal if 3 conditions are true:

- 1. Is markov: intervention does not change conditional independence relations
- 2. If you do intervention, then you set the value. Ex. If you set x3 to "on", then p(x3 is on) = 1
- 3. If you do intervention on x3, conditional distribution of other causal modules does not change because of intervention. To only change this module, you cut off edges going into x3 and set x3 to on, without changing anything else.

Formally, Let $P_x(V)$ be distribution of V resulting from intervention $do(X = x)$. A DAG G is a causal bayesian network if:

- 1. $P_x(V)$ is Markov relative to G
- 2. $P_x(V_i = v_i) = 1$ for all $V_i \in V$ and v_i consistent with $X = x$
- 3. $P_x(V_i|PA_i) = P(V_i|PA_i)$ for all $V_i \notin X$ (PA_i means parents of V_i)

If the three conditions are met, then the graph representation is causal, and you can do inference.

Structural Causal Models Alternatively, we can use structural causal models, which is a set of equations. The value of xi is a function of parents and ei (exogenous / error).

$$X_i = f_i(PA_i, E_i)$$

If system is causal, then each equation represents autonomous mechanism, i.e. each equation is not relevant to other equations. The variables can be dependent, but equations are not related, which is known as **modularity**. Modularity is essential property of causal system. When modularity holds, I can locate where

changes are. Each equation describes how you assign value to left hand side, and describes structure that underlies the variables.

Three types of problems in AI

- Prediction: Do not need causal representation, only care about conditional distribution $p(x3|x2 = 1)$
- Intervention: Predict effect of intervention. "Would person cough if we make sure he has yellow fingers?" We want to see $p(x3|do(x2 = 1))$ (probablility of x3 given intervention on x2). We need the causal picture. If x2 is cause of x3, we need $p(x3|do(x2 = 1))$. If x3 is cause of x2, $p(x3|do(x2 = 1)) = p(x3)$.
- Counterfactual: "Would George cough had he had yellow fingers, given that he does not have yellow fingers and coughs?" This question has more information, we want to see effect of intervention given observation that he does not have yellow fingers and coughs. $p(x3_{x2=1}|x2 = 0, x3 = 1)$ Couterfactuals relate to a particular person, day, etc.

4 Identify causal effects

Randomized control experiments Ex. There are two groups of patients and two treatments. Apart from treatment, other variables need to have same value or distribution. If I observe difference in recovery, it must be caused by treatment. It is hard to do this experiment, hard to constrain patients to have the same value/distribution of other features. Sometimes you do not know what other variables need to be considered. Instead of randomized control experiments, we can use causal understanding and observation data, called causal inference.

Def. Causal discovery: aims to discover causal info by analyzing data

Causal inference and causal discovery are two different problems in causality.

Causal effect Ex. Simpson's paradox. We want to see the effect of treatment on recovery.

$$P(R|T) = \sum_S P(R|T, S)P(S|T)$$

$$P(R|do(T)) = \sum_S P(R|T, S)P(S)$$

Top equation is conditional distribution, and from chain and sum rule in probability theory. If we want to see effect of treatment on recovery, need to find $P(R|do(T))$ (the probability distribution of recovery given intervention of treatment), this describes causal link (we do not care about other factors). How are two equations different? $P(S|T)$ (stone size given treatment (conditional distribution)) becomes $P(S)$ (marginal of stone size), because when you do intervention, you cut off arrow from stone size to treatment, then stone size and treatment become independent.

Origin of paradox: A is better than B. Blue line is for small stones. Many patients with small stone received treatment B. Many patients with large stone received teatment A. For conditional distribution, take average across all patients, and mean is close to where more patients are. Then you get "B better than A" for all patients. When use intervention, you get "A better than B". Intervention is manipulation, we cut some edges off, so the calculated expression is different.

Def. Causal effect probability of target variable given intervention on variable you want to change. To understand this quantity intuitively, in an intervention, we cut off edges into x (the variable we are changing),

set x value, infer probability distribution of target variable y. The average effect is expectation of this distribution. We can also derive other measures based on this distribution.

Identifiability of causal effects A fundamental question in causuality. Statisticians and computer scientists do research in this direction. What is meant by identifiable? Given a causal graph and the same distribution for observed variables, if $P(y|do(x))$ is same for different models, then causal effect of X on Y is identifiable. Identifiable implies unique solution. Ex. We do not observe stone size data, but do observe treatment, recovery data, can we recover influence of treatment on recovery?

Key issue is to control confounding bias. Ex. If you do not observe age, then cholesterol-exercise casual relation is not identifiable.

Two criteria for identifiability:

- **Back door criterion** For a set of variables Z, Z satisfies back door criterion relative to ordered pair (Xi, Xj) if:
  - No node in Z is descendant of Xi.
  - Z blocks every path between Xi and Xj that contains arrow into Xi. Here the back door paths are paths into Xi.

  Z = {X4, X3} satisfies.
  Z = {X4} does not. If you condition on X4, the M-shaped back door path is open (according to d-separation).
  If Z satisfies back door criterion relative to (X,Y), then

$$P(y|do(x)) = \sum_z P(y|x,z)P(z)$$

- **Front door criterion** A set of variables Z satisfies front door criterion relative to ordered pair (X, Y) if:
  - Z intercepts all directed paths from X to Y.
  - There is no back door path from X to Z.
  - All back door paths from Z to Y are blocked by X.

  If Z satisfies front door criterion relative to (X,Y), then

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

Ignorability Compare backdoor criterion with ignorability. Ignorability condition is

$$Y(x) \perp\!\!\!\perp X|Z.$$

What is small x random variable according to potential outcome framework? Y is function of x, u. x is value of some r.v., u is specification of unit (person, day, scenario). $Y(x,u)$ is value of y for unit u on intervention x. We can ignore the unit u, and then $Y(x)$ is counterfactual random variable. Y(x) is description of causal mechanism from x to y. The ignorability condition means: how Y is generated from X will be independent from X, given Z. If back door (graphical condition) is satisfied, then ignorability is satisfied (mostly, we will not discuss the subtle differences.)

Next lecture We will discuss unification of criteria, why do counterfactual reasoning, and how discover causal information from observational data.