DSCI 6001P 数据科学基础

作业 5

提交截止日期：2025.1.5 号晚上 24 点之前

提交方式：电子版发送至 汪远 (wy1001@mail.ustc.edu.cn)

邮件主题和 PDF 命名格式：HW5-姓名-学号，如 HW5-张三-SA24123123

1. Graph. (You should already know the basic concepts of node, edge and path in graph) A graph in which all of the edges are directed is a directed graph, such as Figure 1. And there are some basic concepts for it.

   (1) The node that a directed edge starts from is called the **parent** of the node that the edge goes into. (such as X is the parent of Y)

   (2) The node that the edge goes into is the **child** of the node it comes from. (such as Y is the child of X)

   (3) A path between two nodes is a **directed path** if it can be traced along the arrows. (such as X->Y, and X->Y->Z)

   (4) If two nodes are connected by a directed path, then the first node is the **ancestor** of every node on the path, and every node on the path is the **descendant** of the first node. (such as X is the ancestor of both Y and Z, meanwhile, both Y and Z are descendants of X)

   (5) When a directed path exists from a node to itself, the path (and graph) is called **cyclic**. A directed graph with no cycles is **acyclic.**

   Consider the graph shown in Figure 1,

   (a) Name all of the parents of Z
   (b) Name all the ancestors of Z.
   (c) Name all the children of W.
   (d) Name all the descendants of W.
   (e) Draw all the directed paths between X and T.
   (f) Is this graph acyclic?

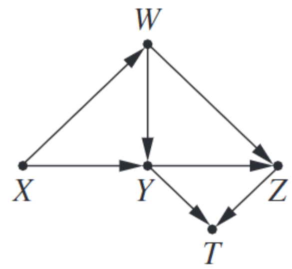

Figure 1

2. Structural Causal Model (SCM) and Graph.
   SCM is a way of formally setting down our assumptions about the causal story behind a dataset. Such as SCM 2.1, a SCM consists of two sets of variables U and V, and a set of functions F that assign each variable in V a value based on the values of the other variables in the model. Every SCM is associated with a graphical causal model, such as SCM 2.1 and Figure 2. Definition of causation in SCM: A variable X is a direct cause of a variable Y if X appears in the function that assigns Y's value. X is a cause of Y if it is a direct cause of Y, or of any cause of Y.

   The variables in U are called **exogenous** variables ("error terms" or "omitted factors."), meaning, roughly, that they are external to the model; we choose, for whatever reason, not to explain how they are caused. The variables in V are

endogenous. Every endogenous variable in a model is a descendant of at least one exogenous variable. Exogenous variables cannot be descendants of any other variables, and in particular, cannot be a descendant of an endogenous variable;

Assume all exogenous variables are independent and that the expected value of each is 0. Answer the following questions:

SCM 2.1

$$V = \{X, Y, Z\}, \quad U = \{U_X, U_Y, U_Z\}, \quad F = \{f_X, f_Y, f_Z\}$$

$$f_X : X = U_X$$

$$f_Y : Y = \frac{X}{3} + U_Y$$

$$f_Z : Z = \frac{Y}{16} + U_Z$$

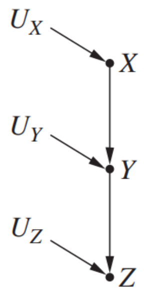

Figure 2.

(a) Describe the causal relationship between X, Y, and Z. (using "cause, caused by (depends on), direct cause")

(b) Determine the best guess of the value (expected value) of Z, given that we observe Y = 3.

(c) Determine the best guess of the value of Z, given that we observe X = 3.

(d) Determine the best guess of the value of Z, given that we observe X = 1 and Y = 3.

(e) Assume that all exogenous variables are normally distributed with zero means and unit variance, that is, σ= 1.

  i. Determine the best guess of X, given that we observed Y = 2.

  ii. Determine the best guess of Y, given that we observed X = 1 and Z = 3. [Hint: You may wish to use the technique of multiple regression, together with the fact that, for every three normally distributed variables, say X, Y , and Z, we have $E[Y|X = x, Z = z] = R_{YX \cdot Z}\, x + R_{YZ \cdot X}\, z.$]

(f) Because of the relationship between SCM and Graph, a graphical definition of causation can be defined as following: If, in a graphical model, a variable X is the child of another variable Y, then Y is a direct cause of X; if X is a descendant of Y, then Y is a **potential** cause of X.

  i. Can we find the **qualitative** causal relationship between X, Y, Z with only the graph of Figure 2? (e.g., answer --- which variable is the cause of Y?)

  ii. Can we find the **quantitative** causal relationship between X, Y, Z with only the graph of Figure 2? (e.g., answer --- how does X affect Y?)

  iii. For causal analysis, what do you think is the potential advantage of causal graph?

  iv. (Optional) When X is a descendant of Y, we only claim that Y is a **potential** cause of X. It means that there are rare cases in which Y will not be a cause of X. Can you find an example for the graph: X->Y->Z? (Hint: For a graph, there are many SCMs associated with it, so the assign functions can be any form)

3. Chains, Forks, Colliders, and d-separation

SCM 3.1 **(Work Hours, Training, and Race Time)**

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$
$$f_X : X = U_X$$
$$f_Y : Y = 84 - x + U_Y$$
$$f_Z : Z = \frac{100}{y} + U_Z$$

SCM 3.2 **(Temperature, Ice Cream Sales, and Crime)**

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$
$$f_X : X = U_X$$
$$f_Y : Y = 4x + U_Y$$
$$f_Z : Z = \frac{x}{10} + U_Z$$

SCM 3.3 (coin 1-gain, coin 2-gain, total-gain)

$$\mathrm{V} = \{\mathrm{X}, \mathrm{Y}, \mathrm{Z}\}, \mathrm{U} = \{\mathrm{U_X}, \mathrm{U_Y}, \mathrm{U_Z}\}, F = \{f_X, f_Y, f_z\}$$
$$\mathrm{f_X}: \quad \mathrm{X} = \mathrm{U_X}$$
$$\mathrm{f_Y}: \quad Y = U_Y$$
$$\mathrm{f_Z}: \quad Z = (X + Y + \mathrm{U_Z})\%2$$

(1) Draw causal graphs that complies with the above SCMs. And denote with them with: Chains, or Forks, or Colliders.

(2) For the three graphs, please judge the independence between variables ('|A' denotes giving A), and prove your results with the above SCMs:
   - i. Chains. (X; Z), (X; Z | Y).
   - ii. Forks. (Y; Z), (Y; Z | X). (Note: X is a **common cause** of variables Y and Z)
   - iii. Colliders. (X; Y), (X; Y | Z). (Note: Z is the **collision** node between two variables X and Y)

(3) The above conclusions are always suitable for graph structures (Chains, Forks, and Colliders) **almost** regardless of the form of F in SCM. Based on these results, there comes **d-separation.**

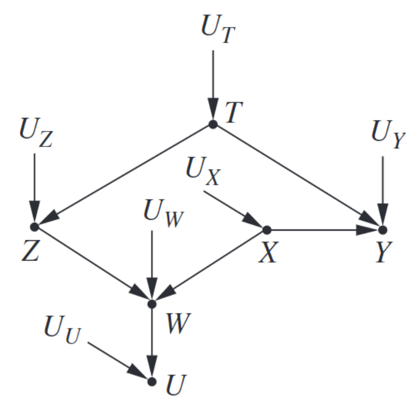

Figure 3.

For graph in Figure 3, answer the questions and give reasons:

i. Condition on {T, W}, whether Z and Y are d-separated.
ii. Condition on {T, U}, whether Z and Y are d-separated.
iii. Condition on {X, T}, whether Z and Y are d-separated.
iv. Find sets that make Z and U be d-separated.

4. Intervention – the backdoor adjustment.
Consider the graph in Figure 4:

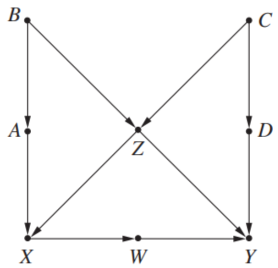

Figure 4

(1) List all the backdoor paths when considering the causal effect of X on Y (i.e. P(Y|do(X))). And try to block each path.
(2) List all of the sets of variables that satisfy the backdoor criterion to determine the causal effect of X on Y.
(3) List all of the minimal sets of variables that satisfy the backdoor criterion to determine the causal effect of X on Y. And write down the causal effect based on one minimal set.

5. Intervention. Consider the causal graph in Figure 5.

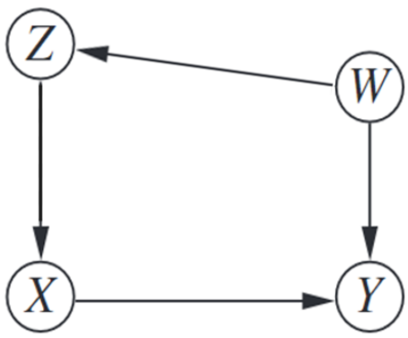

Figure 5

According to the backdoor adjustment, we have:

$$1):\ \mathrm{P}(\mathrm{Y}|\mathrm{do}(\mathrm{X}=\mathrm{x})) = \sum_z P(Y|X=x, Z=z)P(Z=z)$$

$$2):\ \mathrm{P}(\mathrm{Y}|\mathrm{do}(\mathrm{X}=\mathrm{x})) = \sum_w P(Y|X=x, W=w)P(W=w)$$

Try to directly show that:
$\sum_z P(Y|X=x, Z=z)P(Z=z) = \sum_w P(Y|X=x, W=w)P(W=w)$.

6. Gaussian Linear system. Consider the model shown in Figure 6.

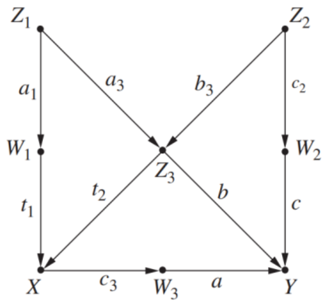

Figure 6

We assume each variable (e.g., Y) corresponds to an omitted **exogenous** variable (e.g. U_Y), and all **exogenous** variables are independent. The functions set is also omitted, but it is easy to recovery with the coefficient on different edges, such as $f_Y: Y = a\,W_3 + bZ_3 + cW_2 + U_Y$.

(1) Compute $E(Y|do(Z_1 = z_1 + 1)) - E(Y|do(Z_1 = z_1))$

(2) Compute $E(Y|do(X = x + 1)) - E(Y|do(X = x))$

(3) How to estimate the causal effect of X on Y with regression.

7. Counterfactual. Given the following SCM, consider now the counterfactual sentence, "Y would be y had X been x, in situation U = u," denoted Y x (u)= y.

$$X = U_X$$
$$H = a \cdot X + U_H$$
$$Y = b \cdot X + c \cdot H + U_Y$$
$$\sigma_{U_i U_j} = 0 \quad \text{for all } i, j \in \{X, H, Y\}$$

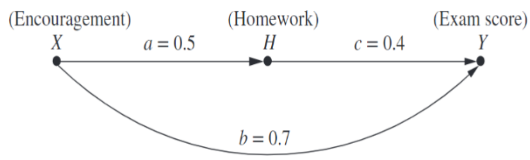

Figure 7

Meanwhile we assume U_H~Bernoulli(p) and $p = 0.5$. Let us consider a student named Joe, for whom we measure H = 2, and Y = 2.5. Suppose we wish to answer the following question: What would Joe's score have been had he doubled his study time? (Hit: the result is not a determined value)

8. Intervention – the front-door adjustment. Consider the graph G in Figure 8,

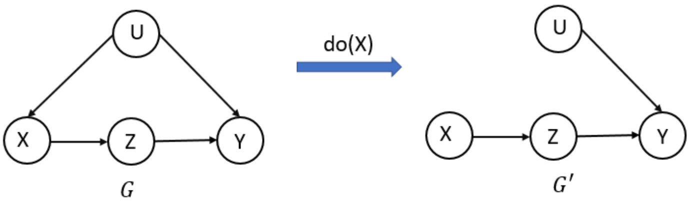

Figure 8

(1) List different expressions of $P(Y|do(Z=z))$, i.e., compute it with different sets that satisfy the backdoor criterion.

(2) Try to prove that:

$$\mathrm{P}(\mathrm{Y}|\mathrm{do}(\mathrm{X}=\mathrm{x})) = \sum_z P(z|x) \sum_{x'} P(Y|X=x', Z=z) P(X=x').$$

Indeed, this adjustment is called the front-door adjustment, which can be used to estimate the causal effect of X on Y when U is not observed.

(Hint: (a) refer to pages 49-52 of the PPT. (b) $P_G(Z=z|X=x) = P_{G'}(Z=z|X=x)$. (c) utilize different expressions of $\mathrm{P}(\mathrm{Y}|\mathrm{do}(\mathrm{Z}=\mathrm{z}))$.)