CS7290 Causal Modeling in Machine Learning: Homework 4

Submission guidelines

Use a Jupyter notebook and/or R Markdown file to combine code and text answers. Compile your solution to a static PDF document(s). Submit both the compiled PDF and source files. The TA's will recompile your solutions, and a failing grade will be assigned if the document fails to recompile due to bugs in the code. If you use Google Collab, send the link as well as downloaded PDF and source files.

Background/Reference

Causal Inference in Statistics - A Primer, Chapter 4

Question 1 - Section 4.2.4, 4.3.1

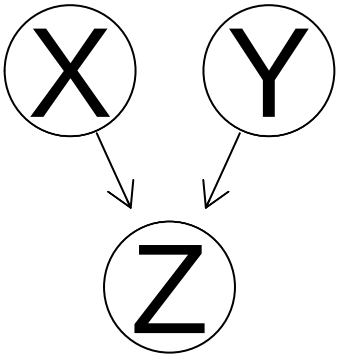
Question 2 - Section 4.4.3, 4.5.1

Question 3 - Section 4.5.2, 4.4.5

Question 4 - Section 4.3.4, 4.3.2

Question 1: Counterfactual inference algorithm (14 points)

X and Y are causes of Z.



The causal mechanism is either an AND gate or and OR gate depending on initial conditions.

AND Gate

X	Y	Z
0	0	0
0	1	0
1	0	0
1	1	1

OR Gate

X	Y	Z
0	0	0
0	1	1

OR Gate		
1	0	1
1	1	1

There is a 50% probability it is an AND gate and a 50% probability it is an OR gate. In both case of the AND gate and the OR gate, X and Y have 50% probability of being 1, and 50% probability of being 0.

The following code represents the structural assignments in a structural causal model of this system.

```
def fx(Nx):
  X = Nx
  return X

def fy(Ny):
  Y = Ny
  return Y

def fz(X, Y, Nz):
  # Mixture of AND gate and OR gate
  Z = Nz * min((X + Y), tensor(1.)) + (tensor(1.) - Nz) * (X * Y)
  return Z
```

1.1

Suppose we observe that X is 1 and Z is 1. What is the probability it is an OR gate? Calculate by hand. (1 point)

1.2

What is $P(Y = 1|X = 1, Z = 1)$? (1 point)

1.3

Suppose we observe that X is 1 and Z is 1. What would Z have been if X were 0? Express this as a probability distribution (assign a probabilities to Z = 1 and Z = 0). Calculate by hand. (1 point)

1.4

Fill in the “. . . ” in the following SCM. (2 points)

Hint: Pyro has a distribution called `Delta`. Its constructor takes only one parameter (e.g. `Delta(a)`), and when you sample from it, you always get a value equal to that parameter. In other words all of the probability in the distribution is concentrated on the parameter. For example, if you write `A = sample("A", Delta(a))`, then the value you sample for A will always be `a`. Why would you want `A = sample("A", Delta(a))` instead of just `A = a`? The reason the `sample` function has you name a variable (e.g. `"A"` in `sample("A", ...)`) is so you can store it by name in the trace object, and refer to that item later with expressions like `condition(model, {"A": a})`. When you have a deterministically set variable and you want to apply `condition` or `do` to it, you can sample it from a `Delta` distribution.

```
def model():
  Nx = sample('Nx', Bernoulli(tensor(.5)))
  Ny = sample('Ny', Bernoulli(tensor(.5)))
  Nz = sample('Nz', Bernoulli(tensor(.5)))
  ...
  return X, Y, Z
```

1.5

Condition the model on X = 1 and Z = 1. Infer the posterior of Nz and the posterior of Y given X = 1 and Z = 1 using importance sampling. Do this by passing the conditioned model to `pyro.infer.Importance`, and naming the resulting object `posterior`. You know it worked if `type(posterior)` returns an object of the class `pyro.infer.importance.Importance`, and `type(posterior())` returns and object of the class `pyro.poutine.trace_struct.Trace`. (4 points)

1.6

Compute the counterfactual probability $P(Z_{X=0} = 0|X = 1, Z = 1)$ and $P(Z_{X=0} = 1|X = 1, Z = 1)$ using Pyro, and compare the results with Question 1.3:

1.6.a

Abduction Step: Generate one sample from the posterior distribution given X=1 and Z=1 using trace (as in HW3 Question 3). Print the values of Nx, Ny, and Nz from the trace. (1 point)

1.6.b

Action Step: Create an intervention model using the intervention $do(X = 0)$. (1 point)

1.6.c

Prediction Step: Condition the intervention model from 1.6.b on the values of the noise variables Nx, Ny, Nz from 1.6.a. In doing this, you are creating a counterfactual model by conditioning on the values of noise variables (Nx, Ny, and Nz) obtained from real evidence (X=1, Z=1) and combining it with an intervention (X=0) that conflicts with that evidence. Generate one sample of Z using this counterfactual model. (2 points)

1.6.d

Repeat the above three steps 1000 times to obtain a sample of Z of size 1000 from the couterfactual distribution. Compute $P(Z_{X=0} = 0|X = 1, Z = 1)$ and $P(Z_{X=0} = 1|X = 1, Z = 1)$ from this sample. (1 point)

Question 2: Necessity and Sufficiency (8 points)

Consider the following data comparing purchases on an e-commerce website with high and low exposure to promotions. X represents promotion exposure (X=0 means low exposure, and X=1 means high exposure). Y represents purchases (Y=1 means purchase, and Y=0 means no purchase).

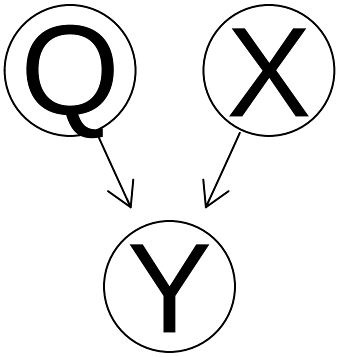
		Promotional Exposure	
		High (X=1)	Low (X=0)
Purchase	Yes (Y = 1)	930	201
	No (Y = 0)	81	808

		Promotional Exposure		
		High (X=1)	Low (X=0)	
Purchase	Yes (Y = 1)	P(X=1,Y=1)=0.4604	P(X=0,Y=1)=0.0995	P(Y=1) = 0.5599
	No (Y = 0)	P(X=1,Y=0)=0.0401	P(X=0,Y=0)=0.4	P(Y=0) = 0.4401
		P(X=1) =0.5005	P(X=0) =0.4995	

Conditional probabilities	
P(Y = 1 | X = 0)	0.1992071
P(Y = 1 | X = 1)	0.9198813
P(X = 1 | Y = 0)	0.0911136
P(X = 1 | Y = 1)	0.8222812

Given these data, we want to estimate the probabilities that high exposure to promotion (X=1) was a necessary cause of purchase (Y=1). We also want to estimate the probabilities that high exposure to promotion (X=1) was a sufficient cause of purchase (Y=1).

We assume the following model, where X is exposure to promotion, Y is purchase, Q is an enabling factor, and Ny is another cause of Y. $\wedge$ means logical AND. $\vee$ means logical OR.



$$\begin{array}{ccc} N_x & \sim & \text{Bernoulli}(p = 0.5) \\ N_q & \sim & \text{Bernoulli}(p = 0.9) \\ N_y & \sim & \text{Bernoulli}(p = 0.2) \\ X & = & N_x \\ Q & = & N_q \\ Y & = & (X \wedge Q) \vee N_y \end{array}$$

2.1 Probability of Neccessity (PN) and Probability of Sufficiency (PS)

Calculate the Probability of Necessity (PN) and the Probability of Sufficiency (PS) using the structural model above.

2.1.a

Calculate PN. PN=$P(Y_{X=0} = 0|X = 1, Y = 1)$. In this scenario, PN means the probability that the purchase would not have occurred (Y=0) in the absence of high exposure to promotion (X=0), given that purchase and high exposure to promotion did in fact occur (X=1, Y=1). (2 points)

2.1.b

Calculate PS. PS=$P(Y_{X=1} = 1|X = 0, Y = 0)$. In this scenario, PS means the probability that the purchase would have occured (Y=1) udner high exposure to promotion (X=1), given that neither high exposure to promotion or purchase did occur (X=0, Y=0). (2 points)

2.2 Probability of Neccessity and Sufficiency (PNS) and Identifiability

Probability of Neccessity and Sufficiency (PNS) is the probability that X=1 is both necessary and sufficient cause of Y=1, represented by $P(Y_{X=0} = 0, Y_{X=1} = 1)$. PNS=$P(X = 1, Y = 1)$PN + $P(X = 0, Y = 0)$PS.

In general, PN, PS, PNS are nonidentifiable. Identifiablility means can be estimated from emperical data, be it observational, experimental, or a combination of both. However, they can be identified under certain assumptions.

Monotonicity: Y is monotonic relative to X, if $Y_{X=1}(u) \geq Y_{X=0}(u)$ for all u. Here u is the values assigned to exgenous variables, which corresponds to a particular member in a population or a specific situation in "nature". In binary case, this means a change of X from 0 to 1 cannot, under any circumstance, make Y change from 1 to 0. In the scenario of this question, it means promotion cannot cause anyone to NOT make a purchase.

Exogeneity: X is exogenous relative to Y if $Y_{X=x} \perp X$. This means the way Y would potentially respond to conditions of X is independent of the actual value of X.

Under the assumptions of both monotonicity and exogeneity, PN, PS, and PNS are all identifiable and are given by the following formulas (Proof can be found in Pearl's book Causality, Chapter 9):

$$\begin{aligned} \text{PNS} &= P(Y = 1|X = 1) - P(Y = 1|X = 0) \\ \text{PN} &= \frac{PNS}{P(Y = 1|X = 1)} \\ \text{PS} &= \frac{PNS}{P(Y = 0|X = 0)} \end{aligned}$$

Typically we don't know the whole structural model. Suppose we only have the statistical tables above, and the assignment formula of $Y = (X \wedge Q) \vee N_y$

2.2.a

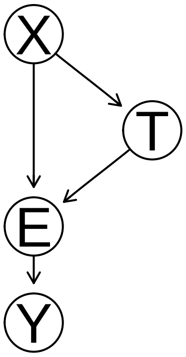
Calculate PNS. (2 points)

2.2.b

Calculate PN and PS using just PNS and the conditional probabilities. (2 points)

Question 3: Mediation (12 points)

Suppose you are a developer for a freemium subscription content platform. Your company did an A/B test for a new feature, designed to increase conversions to a paid premium subscription. Based on some analysis and domain knowledge, you come up with the following model.



$$\begin{aligned} X &= N_X \\ T &= 3 * X + N_T \\ E &= 2 * T + 8 * X + N_E \\ Y &= I(E > 10 + N_C) \end{aligned}$$

$X \in \{0, 1\}$ is whether or not a user was exposed to the feature (X=1 means exposed), and $Y \in \{0, 1\}$ is whether the user converted to a paid premium subscription (Y=1 means converted). T is "thrash". Since the new feature changes the website's user interface, "thrash" quantifies the time and effort the user has to spend familiarizing themselves with the new user interface. E is engagement. The model assumes that the more the user engages with the site the more likely they are to convert. $I()$ is an indicator function, it returns 1 if engagement E is larger than $10 + Nc$, 0 otherwise. Though the A/B test tries to estimate the causal effect of X on Y, T and E are mediators of that effect. You want to know how much the feature drives conversions directly through engagement E, and how much is just due to thrash T (which might have negative consequences on other outcomes not explicitly included in this model).

N_X comes from a fair-coin flip. N_T, N_E and N_C are normal distributions with mean 0. However, for simplicity, we are going to assume noise variables all have a variance/standard deviation of 0. In other words, for our purposes you can assign a value of 0 to all the noise terms.

3.1

Calculate the Total Effect (TE) of X on Y. TE measures the expected increase in Y as the treatment changes from X = 0 to X = 1. (2 points)

3.2

Calculate the Controlled Direct Effect (CDE) when thrash T is 0. CDE is the effect you get when holding a mediator at a fixed value. (2 points)

3.3

Calculate the Natural Indirect Effect (NIE). $NIE = E[Y|do(X = 0, M = M_{X=1})] - E[Y|do(X = 0, M = M_{X=0})]$, where M is mediator. In this scenario, NIE is the expected difference in conversion, given no exposure to the feature (X=0), and thrash T changes from the level of no exposure to the feature (X=0) to the level it would have taken if one was exposed to the feature (X=1) . (2 points)

3.4

Compute the reverse Natural Indirect Effect (NIEr). $NIEr = E[Y|do(X = 1, M = M_{X=0})] - E[Y|do(X = 1, M = M_{X=1})]$, where M is mediator. NIEr is the NIE under reverse transition (from X=1 to X=0). In this scenario, NIEr is the expected difference in conversion, given exposure to the feature (X=1), and thrash T changes from the level of exposure (X=1) to the level it would have taken if one was not exposed to the feature (X=0). (2 points)

3.5

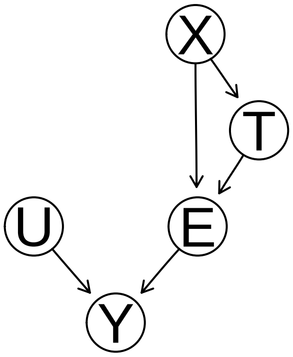
Compute the Natural Direct Effect (NDE) using the following formula: TE = NDE - NIEr. Explain what the implications of this is to the analysis of this feature? (1 point)

3.6

If the noise variables were not degenerate (meaning the didn't have non-zero variance), how would this have affected the calculations and the conclusion about the NDE and NIE? (1 point)

3.7

Suppose instead we used the following model.



$$\begin{aligned} X &= N_X \\ U &= N_U \\ T &= 3 * X + N_T \\ E &= 2 * T + 8 * X + N_E \\ Y &= I(g(E, U, N_C) > \epsilon) \end{aligned}$$

Here, U is a vector of user-related features. $g()$ is a deep neural network that takes as input engagement E as well as these other user features U and the noise term N_C, and outputs a value. ϵ is a threshold. Describe how this would change the overall approach of computing NDE and NIE, if at all. (2 points)

Question 4: Effect of the Treatment on the Treated (9 points)

Suppose you work for a car-sharing service company like Uber. You find that many drivers are making decisions in ways that are sub-optimal for the drivers, often missing low-hanging fruit (e.g. picking up riders closer to where they live, or choosing to drive in areas that have less demand and yet more traffic than others). If the drivers made better decisions about when and where to drive, they could make more money with a similar amount of effort.

The company hires a statistical consulting company that samples some drivers for a training study. The goal of the study is to test whether a driver training program will lead drivers to make better decisions. Drivers in the study are randomly assigned to $X = 1$ (recieved optimal driving training) or $X = 0$ (recieved basic training that doesn't encourage optimal descision-making). The outcome variable Y is the amount of revenue the drivers earn in the study period. Let $Y_{X=1}$ be the revenue earned under exposure to the optimal training and $Y_{X=0}$ be revenue earned under exposure to baseline training. The study showed that the training is highly effected ($E(Y_{X=1} - Y_{X=0}) > \epsilon$) where ϵ is some stastical significance threshold.

Your team is debating whether or not you should build that training program into the mobile app that drivers would opt-in to recieve training and guidance while driving. Your colleagues say that the expected revenue $E(Y_{X=1} - Y_{X=0})$ would more than make up for the cost of developing the app. You argue that most drivers who would opt-in are already highly motivated drivers. You think they would go on to drive more optimally by learning from their own experience, research, seeking out successful drivers, etc.

To demonstrate this, you will estimate the effect of the treatment on the treated (ETT) $E(Y_{X=1} - Y_{X=0}|X = 1)$. ETT in this scenario is the expected difference in earned revenue from those who recieved training relative to what the revenue would have been had they not recieved training. The terms $Y_{X=1}$ and $Y_{X=0}$ here are causal variables, in order to estimate them, you need to convert them into variables than can be estimated directly from data. The following mathematical derivations show you how to calculate ETT given Z. Z is a set of valid adjustment variables that satisfy the backdoor criterion with respect to X and Y. We assume Z is motivation here.

$$\begin{aligned}
P(Y_x = y|X = x') &= \sum_z P(Y_x = y, Z = z|X = x') \\
&= \sum_z P(Y_x = y|Z = z, X = x')P(Z = z|X = x') \\
&= \sum_z P(Y_x = y|Z = z, X = x)P(Z = z|X = x') \\
&= \sum_z P(Y = y|Z = z, X = x)P(Z = z|X = x')
\end{aligned}$$

The first line is summing over all the values of Z. The second line is based on conditional probability. The third line is because the counterfactual interpretation of the backdoor criterion is $X \perp Y_x|Z$ (Theorem 4.3.1 in the Primer book). So conditional on Z, the probability of Y_x doesn't change based on X. This means $P(Y_x = y|Z = z, X = x') = P(Y_x = y|Z = z, X = x)$. The last line is because of consistency rule (Equation 4.6 in the Primer book).

All of the terms in the last line are estimable from data. The data for this question is "driver.csv". Show your calculation of $E(Y_{X=0}|X = 1)$, ETT, and $E(Y_{X=1} - Y_{X=0})$. Then write a short paragraph to explain your results.