Problem Set 3

6.S091: Causality

IAP 2023

Due: Friday, February 2nd at 1pm EST

- Problem sets **must** be done in LaTeX.
- Problem sets are to be submitted at the following link: https://forms.gle/GJWUAhMnwPR8UAxt8, or handed in during class.

Problem 1: Invariance, Causality, and Generalization [10 points]

You are a public health official interested in understanding the factors that contribute to an individual's symptom severity in response to the COVID-19 virus. You collect data from multiple cities across the world. For each city (i.e., environment) $e \in \mathcal{E}$, assume that the data is generated according to a structural equation model over the following exogenous variables:

- Health Risk $H \in \mathbb{R}$: a real-valued health score representing an individual's overall health risk considering many factors such as lifestyle, pre-existing health conditions, etc.
- COVID-19 Severity $Y \in \mathbb{R}$: a real-valued disease severity score.
- Testing Urgency $U \in \mathbb{R}$: a real-valued score representing how quickly an individual tested positive for COVID after initial symptom onset.

The structural causal model M^e for city e is:

$$\begin{aligned} H &= a_e \varepsilon_h & \varepsilon_h &\sim \mathcal{N}(0,1) \\ Y &= H + \varepsilon_y & \varepsilon_y &\sim \mathcal{N}(0,1) \\ U &= Y + b_e \varepsilon_u & \varepsilon_u &\sim \mathcal{N}(0,1) \end{aligned}$$

Notice that the parameters a_e and b_e are environment-dependent. We can think of each environment as coming from a soft-intervention applied to variables H and U.

Since we are interested in understanding what contributes to COVID-19 severity, we will consider Y to be the target variable and H, U to be the covariates. We denote the set of covariates $\mathbf{S} = \{H, U\}$.

Background

You may find the following helpful when completing this problem:

If $\mathbf{X}$ is a N-dimensional centered (i.e., zero-mean) Gaussian random variable, we can compute conditional distributions as follows. Partition $\mathbf{X}$ as $\begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{bmatrix}$ with sizes $\begin{bmatrix} q \\ N-q \end{bmatrix}$. Then the conditional distribution of $\mathbf{X_1}$ given $\mathbf{X_2}$ is a multivariate normal $\mathbb{P}(\mathbf{X_1}|\mathbf{X_2}) \sim \mathcal{N}(\bar{\mu}, \bar{\mathbf{\Sigma}})$ where:

$$\begin{aligned} \bar{\mu} &= \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{X_2} \\ \bar{\mathbf{\Sigma}} &= \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} \end{aligned}$$

The matrix $\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}$ is referred to as the matrix of regression coefficients; it corresponds to the linear regression coefficients obtained when regressing $\mathbf{X_1}$ on $\mathbf{X_2}$.

If $\mathbf{X_1}, \mathbf{X_2}$ are both zero-mean, then $\mathbf{\Sigma}_{12} = Cov(\mathbf{X_1}, \mathbf{X_2}) = \mathbb{E}[\mathbf{X_1}\mathbf{X_2}^T]$.

Invariance and Causality [5 points]

In this problem, we examine the connection between the subset of covariates that have a invariant relationship with the target Y and the subset of covariates that are the causal parents of Y.

(a) Compute the following conditional distributions:

- $\mathbb{P}^e(Y|H)$
- $\mathbb{P}^e(Y|U)$
- $\mathbb{P}^e(Y|H,U)$

(b) Which subsets of the covariates $\mathbf{S}$ have an invariant relationship with Y, i.e., for what $\mathbf{X} \subseteq \mathbf{S}$ does it hold that $\mathbb{P}^e(Y|\mathbf{X})$ is the same for all e? Which covariates are the parents (i.e., direct causes) of Y?

For the next two questions, we will step back from this particular example and consider the extent to which the relationship between invariance and causality holds in general. Assume that, as in the example above, we consider data generated from multiple environments $e \in \mathcal{E}$ over variables $\mathbf{S}$ (covariates) and Y (target). Each environment is generated by performing an intervention on some subset of the covariate variables $\mathbf{X} \subseteq \mathbf{S}$ (but not on variable Y). We assume that there are no hidden variables. We do not assume anything else about the nature of the interventions or the intervention targets.

(c) Does it always hold that the parents of Y satisfy the following invariance condition?

$$\mathbb{P}^e(Y|\mathrm{pa}_{\mathcal{G}}(Y)) = \mathbb{P}^{e'}(Y|\mathrm{pa}_{\mathcal{G}}(Y)) \quad \forall e, e' \in \mathcal{E}$$

If yes, briefly explain why. If no, provide a counter-example.

(d) Consider a subset of covariates $\mathbf{X} \subseteq \mathbf{S}$ that satisfy the invariance condition:

$$\mathbb{P}^e(Y|\mathbf{X}) = \mathbb{P}^{e'}(Y|\mathbf{X}) \quad \forall e, e' \in \mathcal{E}$$

Does it always hold that $\mathbf{X} = \mathrm{pa}_{\mathcal{G}}(Y)$? If yes, briefly explain why. If no, provide a counter-example.

Invariance and Out-of-Distribution Generalization [5 points]

In this problem, we examine the connection between invariant covariates (i.e., invariant features) and predictive models that generalize out-of-distribution. We again consider multi-environment data generated according to SCM M^e. Our goal is to build a model to predict symptom severity Y from a subset of the covariates $\mathbf{X} \subseteq \mathbf{S}$. We seek a model that perfoms well for unseen environments (i.e., cities whose data we **do not** have when training the model).

We assume that when training the model we have access to data from a single city e and that we are interested in finding a model that performs well on an unseen city e' (where $e \neq e'$). We perform ordinary least squares (OLS) regression to fit a model $\hat{Y} = \hat{\beta}^T \mathbf{X}$ using the training data from environment e (since all variables are centered, we don't include an intercept).

(a) What are the estimated parameters $\hat{\beta}$ when using of the following subsets of covariates?

- $\mathbf{X} = \{H\}$
- $\mathbf{X} = \{U\}$
- $\mathbf{X} = \{U, H\}$

Which of the choices of covariates yield a model whose parameters are *not* environment-dependent?

(b) For each subset of covariates, compute the expected error on test environment e', i.e., $\mathbb{E}^e[(Y - \hat{Y})^2]$.

(c) Consider the following variations of the test environment e'. For each, which subset of the covariates should we choose in order to minimize the test error?

- *No distribution shift*: $a_e = a_{e'}$ and $b_e = b_{e'}$.
- *Worst-case distribution shift*: $|a_e - a_{e'}| \approx \infty$ and $|b_e - b_{e'}| \approx \infty$.

Now, we will again step away from the specific example to consider the relationship between invariance, causality, and out-of-distribution performance more generally. We still maintain our assumptions that there are no hidden variables and that there are no interventions on Y. We consider data from two environments: train environment e and test environment e'. Let $f : \mathbf{X} \to Y$ be a model parameterized by θ. Assume we choose θ^* to minimize the loss $\ell(f(\mathbf{X};\theta), Y)$ for samples from train environment e, i.e., $\theta^* = \arg\min_\theta \mathbb{E}^e[\ell(f(\mathbf{X};\theta), Y)]$.

(d) If the covariates $\mathbf{X}$ are chosen such that $\mathbf{X} = \text{pa}_{\mathcal{G}}(Y)$, what will the expected test error be, i.e., $\mathbb{E}^{e'}[\ell(f(\mathbf{X};\theta), Y)]$?

(e) Now consider a set of covariates $\mathbf{X} \subseteq \mathbf{S}$ such that the invariance condition $\mathbb{P}^e(Y|\mathbf{X}) = \mathbb{P}^{e'}(Y|\mathbf{X})$ holds. What will the expected test error be, i.e., $\mathbb{E}^{e'}[\ell(f(\mathbf{X};\theta), Y)]$? Does it matter if $\mathbf{X} \neq \text{pa}_{\mathcal{G}}(Y)$?

(f) What are some possible strategies for finding features that generalize well across distribution shifts? What could we do if we had multiple training environments? Briefly explain.

Problem 2: Central Node Algorithm [10 points]

Consider a rooted tree DAG $\mathcal{G}$ where all edges are oriented away from its root. Let p be the number of nodes in $\mathcal{G}$.

(a) Show that there is no v-structure in $\mathcal{G}$. What is the necessary and sufficient condition for another DAG $\mathcal{G}'$ to be Markov equivalent to $\mathcal{G}$?

(b) Let $\mathcal{I} = \{I\}$ where I is a single-node intervention on some node v in $\mathcal{G}$. How many DAGs are in the $\mathcal{I}$-MEC of $\mathcal{G}$?

(c) Suppose $p > 1$. Show that the verification number $\nu_1(\mathcal{G})$ of $\mathcal{G}$ using single node interventions satisfies $\nu_1(\mathcal{G}) = 1$.

(d) Let $\mathcal{I} = \{I_1, \ldots, I_k\}$ be an arbitrary set of single-node interventions. Show that there exists a single node intervention I_{k+1} such that the number of DAGs in the $\mathcal{I} \cup \{I_{k+1}\}$-MEC of $\mathcal{G}$ is at most half (rounded upwards) of the number of DAGs in the $\mathcal{I}$-MEC of $\mathcal{G}$.

(e) Show that there is an adaptive policy that can orient the full graph in $O(\log_2 p)$ single-node interventions.