ML and Al part 1(DATA37711)- Homework #3

Due on Dec 2 2024

Professor Victor Veitch Autumn 2024

Hongyan Wang

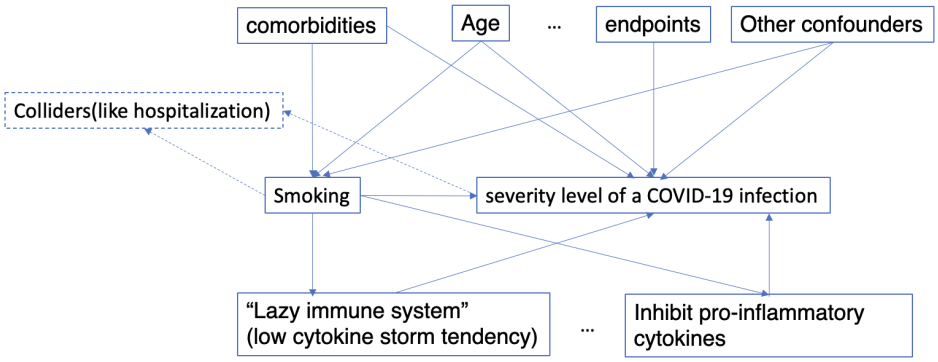
Problem 1

(a) The European Commission review [1] finds that there is an observed negative association between the severity level of a COVID-19 infection and smoking cigarettes. What might explain this observed association?

A:

Let's summarize the reviewed exploratory analysis to get a startpoint for explaination. The relationship betweeen smoking and the severity level of a COVID-19 infection was approached by comparing data on comorbidities reported in the cohort studies with the related data of the underlying populations. Information on the smoking status and comorbidities of the investigated cohorts were compiled from the respective publications. Then smoking prevalence ratios were calculated based on available cohort data, while expected prevalence ratios were adjusted for gender and country-specific differences. This analysis showed lower values of the observed number of smokers than the number of smokers expected for the cohort size in the general population (except two studies), which form the basis for the hypothesis of negative association between the severity level of a COVID-19 infection and smoking cigarettes.

Many factors could contribute to this association. First, it might be backed by biochemical machenisms, like thoese listed in the reference. For instance, smoker's immune system might be more tolerant and less prone to the over- production of immune cells and cytokines, reducing the likelihood of the development of acute respiratory distress syndrome. Or, nicotine inhibits the production of pro-inflammatory cytokines.



Second, a common limitation of all the presented meta-analyses is that most of the data come from retrospective case studies, which might not account for critical confounding variables relevant to the smoking-COVID relationship. In fact, the referenced studies considered only a limited set of confounders and analyzed them using univariate approaches. Without specifically designed studies, the representativeness of these data and the adequacy of the adjustments they allow remain highly speculative.

The final and quite important factor to consider is collider bias. An example is illustrated in the graph: hospitalization or testing can introduce collider bias if both smoking and COVID-19 independently increase the likelihood of being tested or hospitalized. Similar to the "talent-beauty-celebrity" example Victor mentioned in class, where focusing on famous actors creates a negative correlation between talent and beauty, collider bias can distort relationships. In the context of smoking and COVID-19 studies the European Commission reviewed, it is possible that some colliders were inadvertently controlled during data collection.

(b) Bickel et al. [2] presented a famous apparent paradox that relates to sex dis- crimination in admissions at the University of California Berkeley. Based on an analysis of acceptance results across several university

Problem 1 continued on next page...

departments, female applicants had lower overall acceptance rates than male applicants. However, it was also observed that in every department, female applicants had higher acceptance rates than male applicants. What might explain this?

A:

This case is a form of Simpson's paradox. Prof. Peter Bickel stratified data by departments which were the decision-making units. He found that women tended to apply to harder departments to get into, causing greater rejection numbers. Specifically, a higher proportion of females than males applied to departments in the humanities and social sciences where they faced more severe competition and the number of admmisions was smaller. On the other hand, females did not apply as often to departments like mechanical engineering, which had higher acceptance rates.

If we formalize the problem within a causal framework, the phenomenon in which conditioning on departments alters the effect suggests the existence of a pathway through departments that mediates the relationship between sex and admission decisions.

When investigating sex discrimination, the key question is whether the admissions office would have made the same decision if the student had been of a different gender, assuming all other factors remained the same. Addressing this requires to find out the direct effect of gender on the decision while blocking any suspicious indirect pathways.

However, Prof. Peter Bickel's way of estimating the direct effect by conditioning on department do have a crucial assumption – no other varibles, such as unobserved confounders linking the relation between department and admission decision, need to be controlled.

Problem 2

Under the assumption of unconfounderness and overlap, ATE, the average treatment effect:

$$\tau^{ATE} = E_X[E[Y|A = 1, X] - E[Y|A = 0, X]],$$

and ATT, the average treatment effect on the treated:

$$\tau^{ATT} = E_X[E[Y|A = 1, X] - E[Y|A = 0, X]|A = 1].$$

(a) Why might this estimand be preferred in the case where some units are very unlikely to be treated? Does this estimand still do a good job of capturing the qualitative goal of "the effect of the job training program"?

A:

If a design have some kinds units in the population who only very rarely take the treatment. It encounters the problem of poor overlap. In this situation, there isn't valid comparisons for both treated and untreated units across the entire population if trying to estimate the Average Treatment Effect (ATE). ATT estimates

the effect of the treatment only for the treated units. It does not attempt to extrapolate the treatment effect to units that are highly unlikely to receive the treatment, which reduces the reliance on assumptions about those units.

This estimand doesn't estimate "the effect of job training program". It gives "the effect of the program on thoese who participate in the program". Acctually in application, we sometimes should decide wheter the ATE or ATT is more solid and practical in the particular research context.

(b) Propose a simple estimator for τ^{ATT} based on the plug-in estimator for τ that we saw in class.

A:

Let $g(X) = Pr(A = 1|X)$ be the propensity score (assumption, $A \perp\!\!\!\perp Y|X$ and $g(X) < 1$), ATT is

$$\tau^{ATT} = E(Y|A = 1) - E\{E(Y|A = 0, X)|A = 1\} \tag{1}$$

$$= E(Y|A = 1) - E(\frac{g(X)}{g}\frac{(1 - A)Y}{1 - g(X)}) \tag{2}$$

where $g = Pr(A = 1)$ is the marginal probability of the treatment.

The inverse propensity weighting (IPW) estimator of ATT can be given by:

$$\hat{\tau}^{ATT} = \frac{1}{N}\sum_{i=1}^{N} A_i * Y_i/\hat{g} - \frac{1}{N}\sum_{i=1}^{N}\frac{g(X_i)}{1 - g(X_i)} * (1 - A_i)Y_i/\hat{g}$$

where $\hat{g} = \sum I_{A=1}/N$ is the propotion of the treatment in observations.

(c) Using the same data and the same random forest model for the outcome, estimate τ^{ATT} using the plug-in estimator you proposed in part (b). How does your estimator compare to the one in the notebook?

A:

The estimation using in the tutorial is doubly robust estimator(also called also called the augmented inverse propensity score weighting) by James Robins. It combines the outcome-regrssion and the IPW. Doubly robust(DR) estimator is a consistent estimator of the ATE if either the propensity score model or the potential outcome model is, but not necessary both are correctly specified, offerring protection against model mis-specification.

The IPW estimate is $\hat{\tau}^{ATT} = 2232.74$. It is higher then the one in the notebook. Please refer to the ATT_bootstrap.ipynb for details.

(d) Create a 95% confidence interval for your estimate of τ^{ATT} using the bootstrap. How does it compare to the confidence interval in the notebook?

A:

Theoretically, if propensity sccore and outcomes are modeled correctly, DR estimator will have smaller variance than the IPW estimator (in large samples).

The 95% CI for point estimation using bootstrap is: (1463.64, 4261.59). The confidence interval in the notebook is "1296.27 pm 1619.66 = (-323, 2915)", which is slightly wider than the bootstrap result. In total, the estimation of IPW we used here systmatically shifts to the right relative to the result in the tutrial notebook. Please refer to the ATT_bootstrap.ipynb for details.

Problem 3

(a) What is the relationship between these statistics and the distributional fairness criteria discussed in class?

A:

$$DP := average(\{\sigma(\hat{f}(x_i)) : z_i = 1\}) - average(\{\sigma(\hat{f}(x_i)) : z_i = 0\})$$

This statistic checks if Demographic parity holds. If this notion hold, classification score $\hat{f}(X)$ is independent of the sensitive attribute Z and $DP = 0$.

$$EO := average(\{\sigma(\hat{f}(x_i) : z_i = 1, y_i = 1\}) - average(\{\sigma(\hat{f}(x_i) : z_i = 0, y_i = 1\})$$

EO checks the independence of the classification score $\hat{f}(X)$ and the sensitive variable conditional on the value of the target variable Y, targeting the Equalized Odds notion. When Equalized Odds(equal TPR) holds, this statistic is zero.

$$PP := average(\{|y_i - \sigma(\hat{f}(x_i))| : z_i = 1\}) - average(\{|y_i - \sigma(\hat{f}(x_i))| : z_i = 0\})$$

PP meansures the overall performance of model across groups. It is related with the acoring accuracy. If the accuracy parity holds, the PP statistic is expeted to be 0.

(b) Train a fairness through unawareness classifier $\hat{f}_{FTU}$ that predicts Y using all the non-sensitive features. Use a gradient boosting classifier.

A:

The test accuracy is 84.66%. The value of statistics in (a) : DP= -0.1673, EO= -0.1816, PP= -0.2399.

Please refer to the Fairness.ipynb for details.

(c) Train a classifier $\hat{f}(x, z)$ using all the features, including the sensitive attribute Z. And then define a new classifier by marginalizing out the sensitive attribute.

A:

The test accuracy is 84.08%, lower than the model in (b). However, The value of statistics is: DP= -0.0525, EO= -0.0173, PP= -0.1875. These value of statistics are closer to zero than those in the setting of (b), showing some mitigation of unfairness.

(d)

A:

According to the causal graph, the $Y \perp\!\!\!\perp W|Z$, so,

$$\hat{f}^{inv}(x) = \frac{1}{|Z|}\sum_{z\in Z} f(x,z) \tag{3}$$

$$= \frac{1}{|Z|}\sum_{z\in Z} P(Y=1|X=x,Z=z) \tag{4}$$

$$= \frac{1}{|Z|}\sum_{z\in Z} P(Y=1|P=p,W=w,Z=z) \tag{5}$$

$$= \frac{1}{|Z|}\sum_{z\in Z} P(Y=1|P=p,Z=z) \tag{6}$$

$$= \frac{1}{|Z|}\sum_{z\in Z} f(p,z) \tag{7}$$

is a function of P only.

(e) Next we will just use the real data, dropping the synthetic sensitive attribute Z and instead treating sex as the sensitive attribute. Run the same analysis as in parts (b) and (c) using the real sensitive attribute. How do the results compare to the synthetic sensitive attribute?

A:

For the fairness through unawareness classifier, prediction accuracy on test set is 84.62%, values of the three statistics in (a) are: DP= -0.0585, EO= -0.0781, PP= -0.0664.

For the classifier including sensitive feature, the prediction accuracy on test set is 84.61%, values of the statistics are: DP= -0.0539, EO= -0.0670, PP= -0.0655.

First, all these statistics are negative, indicating that females (corresponding to Z=1)are less likely to be predicted as high earners. When the sex feature is included to train a fairness-aware model, these unfairness metrics only shift slightly closer to zero. Compared with the synthetic case, synthetic sensitive attribute are more critcal for model fairness. But like sex, adding sensitive features to a machine learning model does not inherently decrease unfairness.