Midterm 2 Practice for COMP 6321 Fall 2019

– With my answers

Q1. Suppose you want to build a predictor $\hat{Y}$ that predicts whether a future PhD applicant would pass a comprehensive exam. You have historical training cases containing three attributes:

- G (their undergraduate grades),
- S (their biological sex), and
- Y (whether they passed or failed).

For each of the following, give your answer in words, **without** using any probabilistic notation $P(\cdot \mid \cdot)$.

- 1. Describe the **fairness through unawareness** criterion applied to your predictor.
- 2. Describe the **individual fairness** criterion applied to your predictor.
- 3. Describe the **demographic parity** criterion applied to your predictor.
- 4. Describe the **equality of opportunity** criterion applied to your predictor.

Answer:

- 1. **Fairness through unawareness** (FTU) definition: *An algorithm is fair so long as any protected attributes A are not explicitly used in the decision-making process.*

  In this example, we should treat S (their biological sex) as protected attributes A, so if we apply FTU to our predictor, we should predict Y only on G (their undergraduate grades) discarding the protected attribute (their biological sex S)

  2. **Individual fairness** (IF) definition: *An algorithm is fair if it gives similar predictions to similar individuals. Formally, given a metric d(· , ·), if individuals i and j are similar under this metric (i.e. d(i, j) is small), then their predictions should be similar:* $\hat{Y}\left(X^{(i)}, A^{(i)}\right) \approx \hat{Y}\left(X^{(j)}, A^{(j)}\right)$.

  In this example, if we apply IF to our predictor, it means any 2 PhD applicants with similar undergraduate grades and similar biological sex should have similar prediction on whether they passed or failed.

- 3. **Demographic parity** (DP) definition: *A predictor $\hat{Y}$ satisfies demographic parity if* $P(\hat{Y} \mid A = 0) = P(\hat{Y} \mid A = 1)$

  In this example, we should treat S (their biological sex) as protected attributes A. So, if we apply DP to our predictor, we should have same probability saying that an applicant is passed whether the applicant is male or female, and same for the probability saying the applicant is failed.

  (From website: *http://blog.mrtz.org/2016/09/06/approaching-fairness.html*)
  DP requires that a decision of passing or failing a PhD applicant should be independent of the protected attribute (in this example, is biological sex S)

- 4. **Equality of Opportunity** (EO) definition: *A predictor $\hat{Y}$ satisfies equality of opportunity if* $P(\hat{Y} = 1 \mid A = 0, Y = 1) = P(\hat{Y} = 1 \mid A = 1, Y = 1)$

  In this example, we should treat S (their biological sex) as protected attributes A. So if we apply EO to our predictor, while training, if the applicant is actually passed (i.e. $Y = 1$), the probability of the applicant is passed if the applicant is male should be equivalent to the probability of the applicant is passed if the applicant is female.

  (From website: https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb)
  In this example, EO is to say that we should have equal proportion of individuals from the qualified fraction of each group (male, female)

Q2. Does the *fairness* through unawareness criterion guarantee that protected attributes will not influence the prediction? **Explain.**

Answer:

The answer is no. Fairness Through Unawareness (FTU) criterion does not guarantee that protected attributes will not influence the prediction.

Reason: Because FTU doesn't use protected attributes in decision-making process. It is not equivalent to protected attributes will not influence the prediction if we use them in decision-making process.
In addition, FTU has a shortcoming as elements of X (observable attributes) can contain discriminatory information analogous to A (protected attributes) that may not be obvious at first.

Q3. The paper describes a "red car" scenario (fairness in insurance prices) and a "high crime" scenario (fairness in law enforcement). **Why** in the "red car" scenario did the postulated (假设，假定) causal graph (因果图) have **no arc from X (car is red) to Y (accident rate),** whereas in the "high crime" scenario the postulate causal graph did have an arc from X (resident location) to Y (arrest rate)?

Answer:

Because, **in the "red car" scenario,**

A: human race

X: whether individual prefers red cars

U: unobserved factor corresponding to aggressive driving

Y: probability of accident rate

The changes on car color should not influence the probability of accident rate, so there is no arc from X to Y.

Whereas, **in the "high crime" scenario,**

A: human race

X: resident location

U: totality of socioeconomic factors and policing practices which both influence where an individual may live and how likely they are to be arrested and charged

Y: a binary label with $Y = 1$ indicating a criminal arrest record

Due to the fact that higher observed arrest rates in some location X are due to greater policing in that location, and $Y = 0$ does not mean someone has not committed a crime but rather they have not been caught and charged**.** So, the change on location X has an effect on our prediction (criminal arrest record). So we have an arc from X to Y.

Q4. In the "high crime" scenario:

- 1. Why did the postulated causal graph have **an arc from A (race) to X (residential location)?**
- 2. If a resident of neighborhood X is observed to **have $Y = 0$ arrests**, **what can we say** about whether they have committed a crime?

Answer:

- 1. In the “high crime” scenario,

  A: human race

  X: resident location

  U: totality of socioeconomic factors and policing practices which both influence where an individual may live and how likely they are to be arrested and charged

  Y: a binary label with $Y = 1$ indicating a criminal arrest record

  Due to historically segregated housing, the resident location X depends on human race A. So the graph has an arc from A to X.

- 2. In the “high crime” scenario, the label $Y = 0$ does not guarantee someone did not commit a crime but rather they have not been caught in location X. So if we observed it has $Y = 0$ arrests in location X, we can only say people living there who may not commit a crime or may commit a crime and didn’t get caught.

Q5. In the "law school success" scenario:

- 1. What were the specific unprotected features and protected features used to make a prediction? List them separately.
- 2. What specific outcome was being predicted, and how was prediction error measured?
- 3. Which model had **lowest prediction error**, and why?
- 4. Which model had the **highest prediction error**, and why?
- 5. The authors propose a "level 2" model that postulates a certain unobserved variable as having causal influence on the observable features. **What was this latent variable**, and **how did the authors choose to model its influence on the observed features**? Be precise.

Answer:

- 1. Unprotected features:
  - LSAT (entrance exam scores)
  - GPA (grade-point average)
  - FYA (first year average grade)

  Protected features:
  - Race
  - Sex

- 2. Outcome: predict if an applicant will have a high FYA

  Error measurement: using RMSE (Root Mean Square Error) to measure the error

- 3. Full model has the lowest prediction error.

  Because it uses Race and Sex to more accurately reconstruct FYA. But the model is unfair.

- 4. Fair K model has the highest prediction error.

  Because Fair K model doesn't use any one of the unprotected features and protected features to predict, since even the unprotected features may be biased due to social factors. What the model does is that it postulates *a weak assumption – a student's knowledge (K) affects GPA, LSAT and FYA scores.* And it uses

observed training set to estimate the posterior distribution of K, then use K to construct the predictor. The advantage of this is that K is independent of Race and Sex.

- 5. The latent 'far' variables are parents of observed variables which are independent of both Race and Sex.

In Level 2, Author postulates that a latent variable: a student's knowledge (K) affects LSAT, GPA, and FYA scores. They perform inference on the model using an observed training set to estimate the posterior distribution of K. Then they construct the predictor using K.

In Level 3, Author models GPA, LSAT, and FYA as continuous variables with additive error terms ε_G, ε_L, ε_F which are independent of race and sex. Then use these residual estimates of ε_G, ε_L to predict FYA.