Due Date: Oct 8, 3pm

Please typeset all responses in LaTeX and submit a PDF through Markus (instructions to follow). In responding to questions in Section 1, please provide *both* written responses *and* code[1] in your submission.

1 Coding Exercises

Remark: Helper code and data can be found at the following github repo:
https://github.com/ecreager/csc2541-f19/tree/master/assignment1

1.1 Objectives and Preliminaries

We consider fairness implications of training binary classifiers on a dataset where a binary sensitive attribute is observed. Let Y_i be the class of the ith element on which we evaluate our classifier, let A_i be its sensitive attribute for which we are interested in ensuring fairness, and let $\hat{Y}_i$ be the classifier prediction. All of Y_i, A_i, and $\hat{Y}_i$ are binary (either 0 or 1).

We will need the following three metrics to analyze our classifier performance. Let n be the number of examples we evaluate our classifier on, and $n_{A=0}, n_{A=1}$ are the number of examples with $A_i = 0$ or 1 respectively. First, define the *accuracy* $\mathcal{A}$ as:

$$\mathcal{A} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[\hat{Y}_i = Y_i], \tag{1}$$

where $\mathbb{1}$ is the indicator function (equal to 1 if the statement inside is true, 0 otherwise). $for readability$.

Next, define the *reweighted accuracy* ($\mathcal{R}_Y$) of a classifier as the mean accuracy normalized by the size of the two groups:

$$\mathcal{R}_Y = \frac{1}{2}\Big(\frac{1}{n_{A=0}} \sum_{i=1}^{n} \mathbb{1}[\hat{Y}_i = Y_i, A_i = 0] + \frac{1}{n_{A=1}} \sum_{i=1}^{n} \mathbb{1}[\hat{Y}_i = Y_i, A_i = 1]\Big) \tag{2}$$

If we are interested in predicting A rather than Y, we can define this metric analogously as

$$\mathcal{R}_A = \frac{1}{2}\Big(\frac{1}{n_{A=0}} \sum_{i=1}^{n} \mathbb{1}[\hat{A}_i = A_i, A_i = 0] + \frac{1}{n_{A=1}} \sum_{i=1}^{n} \mathbb{1}[\hat{A}_i = A_i, A_i = 1]\Big) \tag{3}$$

Finally, a fairness metric Δ_{DP}, which measures the (lack of) demographic parity (DP) of the classifier:

$$\Delta_{DP} = \Big|\frac{1}{n_{A=0}} \sum_{i=1}^{n} \hat{Y}_i \cdot (1 - A_i) - \frac{1}{n_{A=1}} \sum_{i=1}^{n} \hat{Y}_i \cdot A_i\Big| \tag{4}$$

[1] Organize the code for each coding question in its own (clearly named) script.

This is a lot of notation, but the concepts are fairly simple: accuracy measures how often the classifier is correct; reweighted accuracy measures how often the classifier is correct if we weight each group equally, and Δ_{DP} measures the absolute difference in predictions between the two groups.

1.2 Fair Classification

1. Let's start by looking at our data. In this assignment, we'll use the Adult dataset[2], which is a classic machine learning dataset, using data from a US census. The label (Y) we are trying to predict is income, which is binarized to two categories. The sensitive attribute (A) we are concerned about fairness with respect to is gender (male or female). You can find more info about this data in a README in the assignment folder.

   If memory issues arise on your computer, you may sample a half or a quarter of the dataset; if you do so please make an note of it in your submission.

   Let's look just at the training set for now. Name the 10 features which are most correlated with Y_i, and the 10 which are most correlated with A, as measured by (absolute) Pearson correlation (ignore any NaN correlations you see).

2. Now let's train a binary classifier to predict Y. There is a training set and test set specified in the assignment folder. We'll use $\hat{Y}$ to denote the binary prediction of the classifier (0 or 1). There is some code for training a neural network in the assignment folder as well. You will have to modify this file to implement the three metrics we care about (overall accuracy, reweighted accuracy, Δ_{DP}).

   Train the classifier on the training set and report its overall accuracy, reweighted accuracy and Δ_{DP} on the test set. Which sensitive group has higher values of $\hat{Y}_i$, on average?

   Remark: Several students noted that the sensitive attribute A is contained in two of the feature columns of X, and asked whether these columns should be removed prior to computing the correlations. *Either removing these columns from X or retaining them in them is acceptable for the purposes of this assignment.* The choice of including or excluding the features should not affect the classifier's fairness-accuracy tradeoff in the subsequent questions too much.

3. Let's take a look at how the features in our data correlate with the learned predictor $\hat{Y}$. Which three features in the data are most correlated with $\hat{Y}_i$? Which three features are most correlated with $\hat{Y}$, only looking at examples where $A = 0$? Which three features are most correlated with $\hat{Y}$, only looking at examples where $A = 1$?

4. Change the training objective to include a term that estimates the lack of statistical parity on the training set (Δ_{DP}), with a hyperparameter α to control the weighting of this term relative to accuracy. Determine a range of α that produces reasonable

[2] https://archive.ics.uci.edu/ml/datasets/adult

results when training the neural network and report this range in your writeup. Carry out a sequence of network trainings with α varying over this range, and produce a plot (similar to the one described above) where accuracy (overall and reweighted) and Δ_{DP} are shown as functions of α.

Remark: Note that the new regularizer term must be differentiable w.r.t. the network parameters, so functions of the hard predictions $\hat{Y} \in \{0, 1\}$ (computed by thresholding network output, for example) are not permissible. Discuss your the functional form of your regularizer in the writeup, as well as any design considerations that went into its formulation.

2 Written Exercises

1. Construct three random variables X, R and Y such that X is independent of R, but. where X is dependent of R given Y.

2. Exercise 1 from Chapter 2, Fairness and Machine Learning. Give an example of a classification problem where the target variable Y assumes three distinct values, and such that independence and separation are simultaneously achievable (in the nondegenerate case where A, Y are not independent.)

3. In many situations we can predict both Y and A from the same data with reasonable accuracy. Suppose you've trained some classifier g to predict Y from X. Show that there exists a classifier h which predicts A from X with reweighted accuracy $\mathcal{R}_A$ greater than or equal to $\frac{1}{2}\Delta_{DP} + \frac{1}{2}$, where Δ_{DP} is measured with respect to g.