Assignment 12

Statistical Machine Learning

Moritz Haas / Prof. Ulrike von Luxburg
written by Michael Lohaus

Summer term 2022 — due on **July 18th at 12:00**

Exercise 1 (COMPAS recidivism scores, 3+3+1+3 points)
From the lecture, you are familiar with the COMPAS dataset. In this exercise we will take a closer look. The COMPAS dataset provides the `decile_score` attribute ranging from 1 to 10, where 10 indicates the highest risk. To transform the score into a binary class label (is the defendant high-risk or not?), we introduce a cut-off score s_0. Our binary prediction is 1 if the score is above s_0 and 0 otherwise, so $\hat{Y} = 1(s > s_0)$. The dataset also contains a binary outcome Y and a binary sensitive attribute A. Given a score $S = S(x)$, we consider the notion of **calibration**, where for all values s, we need

$$P(Y = 1|S = s, A = a) = P(Y = 1|S = s, A = b).$$

To define fairness notions for classification, we consider the false positive rate (FPR), the true positive rate (TPR), and the positive predictive value (PPV). Additionally, we define the group-wise rates for our fairness analysis:

$$\begin{aligned}\mathrm{FPR}_a(\hat{Y}) &= P(\hat{Y} = 1|Y = 0, A = a)\\ \mathrm{FNR}_a(\hat{Y}) &= P(\hat{Y} = 0|Y = 1, A = a)\\ \mathrm{PPV}_a(\hat{Y}) &= P(Y = 1|\hat{Y} = 1, A = a).\end{aligned}$$

With these rates, we can define **Predictive Parity** and **Equal Odds**. We say that the classifier $\hat{Y}$ respects **Equal Odds** if

$$\mathrm{FNR}_a(\hat{Y}) = \mathrm{FNR}_b(\hat{Y}) \text{ and } \mathrm{FPR}_a(\hat{Y}) = \mathrm{FPR}_b(\hat{Y}).$$

Similarly, we say that the classifier $\hat{Y}$ respects **Predictive Parity** if

$$\mathrm{PPV}_a(\hat{Y}) = \mathrm{PPV}_b(\hat{Y}).$$

Lastly, we define the **prevalence** $p = P(Y = 1)$, and the group-wise prevalence $p_a = P(Y = 1|A = a)$.

(a) Download the notebook and the dataset. Explore how fair the decisions are with respect to the above fairness notions (for evert possible threshold s_0). Can all of the above fairness notions be fulfilled for the same classifier?

(b) *Theoretical Part.* With the empirical observations made in the notebook, we now would like to know: Can a binary classifier fulfill both Predictive Parity and Equalized Odds?

For this, first prove the following general equation:

$$\mathrm{FPR} = \frac{p}{1-p}\frac{1-\mathrm{PPV}}{\mathrm{PPV}}\left(1-\mathrm{FNR}\right).$$

Hint: The equation is not related to groups. Use the confusion matrix to express PPV, FPR, FNR, and p.

(c) (i) If there is more than one group, the above formula holds for each group. Now, assume that Predictive Parity holds. Using the above formula, what is the necessary condition such that Equalized Odds holds as well?

(ii) *Back to the notebook.* Check this condition in the notebook for the COMPAS dataset.

(d) Again in the notebook, discard the decile scores, and learn your own classifier from the data. Let us check how fair this classifier is, and explore an easy (and trivial) way to be more fair with respect to Equalized Odds.

Exercise 2 (Sub-optimality of post-hoc correction, 2+3+3+1+1 points)
Given binary labels $Y \in \{0,1\}$, binary predictions $\hat{Y} \in \{0,1\}$, and a binary group label $A \in \{0,1\}$, we call the classifier $\hat{Y}$ **fair** if it fulfills **Equal Odds** as defined above. (Using the group-conditional false negative rates FNR_a and false positive rates FPR_a). In post-hoc correction, we post-process given predictions $\hat{Y}$ and the group label A into a randomized predictor $\tilde{Y} = f(\hat{Y}, A)$. It is randomized because we are learning the probabilities of $\tilde{Y}$ to predict 1 for given $\hat{Y}$ and A:

$$p_{ya} := P(\tilde{Y} = 1 | \hat{Y} = y, A = a).$$

In the following optimization problem, we are learning the p_{ya}, which determine the random function $\tilde{Y} = f(\hat{Y}, A)$. With a loss function ℓ, we write

$$\begin{aligned} \tilde{Y} = \underset{p_{00}, p_{01}, p_{10}, p_{11}}{\operatorname{argmin}} \quad & E[\ell(f(\hat{Y}, A), Y)] \\ & \mathrm{FNR}_0(f) = \mathrm{FNR}_1(f) \\ & \mathrm{FPR}_0(f) = \mathrm{FPR}_1(f). \end{aligned}$$

We want to compare the optimality of the post-hoc correction $\tilde{Y}$ with the optimal fair predictor Y^* in an hypothesis class $\mathcal{H}$, which had access to the features X:

$$\begin{aligned} Y^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \quad & E[\ell(h(X, A), Y)] \\ & \mathrm{FNR}_0(h) = \mathrm{FNR}_1(h) \\ & \mathrm{FPR}_0(h) = \mathrm{FPR}_1(h). \end{aligned}$$

In the best case, the post-hoc correction $\tilde{Y}$ would yield a classifier that has a loss not far from Y^*. We are going to show that this is not true when the loss function is the 0-1 loss by considering the following counter example. Let $X, A, Y \in \{0,1\}$ and $\varepsilon \in (0, 1/4)$. The distribution $\mathcal{D}_\varepsilon$ has the following properties

- $P(Y = 1) = 0.5$
- $P(A = y | Y = y) = 1 - \varepsilon$
- $P(X = y | Y = y) = 1 - 2\varepsilon$
- $X \perp A | Y$.

In other words, the label is evenly distributed, and A and X are very predictive of Y. However, using A is slightly better. For this distribution, we prove that

- the optimal fair predictor Y^* has loss at most 2ε, and
- for the unrestricted Bayes optimal predictor $\hat{Y}$, the post-hoc correction of $\hat{Y}$ returns a predictor $\tilde{Y}$ with loss at least 0.5.

Hence, the post-hoc correction cannot be better than random guessing, and is therefore suboptimal. Prove this statement with the following steps:

(a) Show that $Y^* = X$ is fair and has a loss of 2ε. This predictor is an upper bound on the loss of the optimal fair predictor with 0-1 loss.

(b) The unconstrained Bayes optimal predictor is the hypothesis

$$\hat{Y} = \underset{y \in \{0,1\}}{\operatorname{argmax}} \quad P(Y = y | X = x, A = a).$$

Show that $\hat{Y} = A$. Is $\hat{Y}$ fair?

(c) Consider now the post-hoc correction $\tilde{Y}$ of $\hat{Y}$. We use the following linear program to find the randomized function f:

$$\begin{aligned}\tilde{Y} = \underset{p_{00},p_{01},p_{10},p_{11}}{\operatorname{argmin}} \quad & E[\ell(f(\hat{Y}, A), Y)] \\ & \mathrm{FNR}_0(f) = \mathrm{FNR}_1(f) \\ & \mathrm{FPR}_0(f) = \mathrm{FPR}_1(f) \\ & \begin{bmatrix}\mathrm{FPR}_a(f) \\ \mathrm{TPR}_a(f)\end{bmatrix} \in \mathrm{ConvHull}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\mathrm{FPR}_a(\hat{Y}) \\ \mathrm{TPR}_a(\hat{Y})\end{bmatrix}, \begin{bmatrix}\mathrm{FPR}_a(1-\hat{Y}) \\ \mathrm{TPR}_a(1-\hat{Y})\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}\right) \text{ for all } a \in \{0,1\}.\end{aligned}$$

The first and second constraint requires equal true positive and false negative rates (Equal Odds). The third constraint requires that the true positive and the false positive rate be in the convex hull of the constant 0 predictor, the constant 1 predictor, the predictor $\hat{Y}$ and its opposite. This is where $\tilde{Y}$ is derived from $\hat{Y}$.

Use the predictor $\hat{Y}$ that we found, and rewrite the third constraint. What does the third constraint require here? Hint: No need to know the exact definition of the convex hull. Just look at the 4 points a bit closer.

(d) Now, putting all constraints together, what is the optimal (randomized) post-hoc classifier $\tilde{Y}$? What is its loss?

(e) Summarize the implications of these findings. Why is the post-hoc correction suboptimal? In particular, discuss the fact, that the post-hoc classifier is randomized. In general, think about the **pros and cons of randomized classifiers** (which might not be as bad as in this exercise).