# DS 102 Homework 2

If you are handwriting your solution please make sure your answers are legible, as you may lose points otherwise.
Data science is a collaborative activity. While you may talk with others about the homework, please write up your solutions individually. If you do discuss the homework with others, please include their names on your submission.

**Due on Gradescope by 9:29am, Tuesday 25th February, 2020**

1. (35 points) **Non-discrimination Criteria**

   In ProPublica's 2016 investigation, they claim that COMPAS exhibited racial bias against black individuals. Specifically, the investigation revealed a racial disparity in the *error rates* of the tool. Among the defendants who ultimately did not recidivate, the Black defendants were labeled "high risk" at a higher rate than the White defendants.

   Northpointe, the company that sells COMPAS, published a report in response, arguing that their risk scores are equally accurate and predictive for white and black defendants. In order to evaluate whether both of their claims are true, we analyze the allegations and response in the framework of our non-discrimination criteria: *equal positive rates*, *equal error rates* and *calibration by groups.*

   We will view this as a binary problem. We let the classifier $\hat{Y}$ be 1 if a defendant is "high risk" and 0 if they are "low risk" according to their COMPAS score. Let $Y$ be the true outcome, 1 if an individual recidivated and 0 otherwise. Finally, let $A$ be the race of the defendant.

   (a) (8 points) **Translate the following statements from ProPublica as inequalities between conditional distributions of the classifier, the true outcome, and sensitive attribute:**

      (1) Black defendants who did not recidivate within two years were strictly more likely to be misclassified than their white counterparts.

      (2) White defendants who re-offended within the next two years were mistakenly labeled "low risk" strictly more often.

   (b) (4 points) **Do ProPublica's statements in 1 (a) violate any of the three fairness criteria discussed in class? If so, which one?**

   (c) (4 points) **Translate the following statements from Northpointe into relationships between conditional probabilities:** (use "$\approx$" to represent "comparable")

      ***Among the defendants who received the same COMPAS score, a comparable percentage of the black defendants re offend in comparison to the white defendants.***

   (d) (4 points) **Which of the fairness criteria does Northpointe's claim in the previous part satisfy?**

(e) (8 points) Define $p_a$ as the proportion of group $a$ that recidivate, i.e. $p_a = P(Y = 1|A = a)$, $\text{TPR}_a$ as the true positive rate within group $a$, $\text{FPR}_a$ as the false positive rate within group $a$. Define $\text{PPV}_a$ as the positive preditive value for group $a$, and $\text{NPV}_a$ as the negative predictive value group $a$ as follows:

$$\text{PPV}_a = P(Y = 1|\hat{Y} = 1, A = a)$$

**Prove the following relationship:**

$$\text{PPV}_a = \frac{\text{TPR}_a \cdot p_a}{\text{TPR}_a \cdot p_a + \text{FPR}_a \cdot (1 - p_a)}$$

for all $a \in \{\text{black}, \text{white}\}$.

*Hint: First, try to express* $\text{PPV}_a$ *in terms of* $P(\hat{Y} = 1, Y = 1|A = a)$ and $P(\hat{Y} = 1|A = a)$. *Use Bayes rule.*

(f) (7 points) Suppose that recidivism probability is not independent of race, i.e. $P(Y = 1|A = \text{black}) \neq P(Y = 1|A = \text{white})$, and the two groups have nonzero true and false positive rates. **Show that if *equalizing error rates* holds, then the two groups *cannot* have the same positive predictive values.**
*Hint:* what does *equalized error rates* imply? You might want to use the result from the previous part.

2. (25 points) **Fairness threshold**

In this problem, we consider an automated resume screening tool which is used by a company to sort candidates based on whether or not they should be invited for an on-site interview after an initial phone screen.

Let the random variable $X$ denote the features of a candidate's application and $Y$ denote whether a candidate is invited for an on site interview, where $Y = 1$ indicates that an individual was invited.

Below is a table which shows the predictions and outcomes for applicants split by membership in a minority religious group, with $A = 1$ indicating that an individual is a member of this group and $A = 0$ indicating that they are not. We assign $\widehat{Y} = 1$ for those who will be considered more closely by recruiters and $\widehat{Y} = 0$ for those who will not.

| | $A = 1$ | | | $A = 0$ | | |
|---|---|---|---|---|---|---|
| | $\widehat{Y} = 0$ | $\widehat{Y} = 1$ | | $\widehat{Y} = 0$ | $\widehat{Y} = 1$ | |
| $Y = 0$ | 360 | 40 | 400 | 2700 | 300 | 3000 |
| $Y = 1$ | 40 | 60 | 100 | 600 | 900 | 1500 |
| | 400 | 100 | | 3300 | 1200 | |

(a) (10 points) With membership in the religious group as the sensitive attribute, **does this classifier satisfy *equalizing positive rates* criterion exactly? Does it satisfy *equalizing error rates* criterion exactly?** Justify your answer.

(b) (5 points) Suppose we use **a threshold rule** $\widehat{Y} = \mathbb{1}\{R(X) > t\}$ which performs binary classification based on some score R(X) to assign $\widehat{Y} = 1$ for those who will be considered more closely by recruiters and $\widehat{Y} = 0$ for those who will not.

For the criteria that the classifier doesn't satisfy, **propose group dependent thresholds**, i.e. using threshold $t_a$ in group $a$, that result in a classifier that does satisfy the criteria. You do not need to specify exact quantities, rather comparisons with the current threshold. You should not propose a trivial threshold that results in 0% or 100% acceptance rates.

(c) (10 points) Suppose that the score is estimated using **some historical data**. **Compare and contrast** the value of the intervention you suggested in the previous part for the following two circumstances:

(1) You learn that the historical data comes from a hiring manager who is a member of the religious group and has been heard telling members that they have an "in" regardless of their qualifications.

(2) You learn that there is a well-regarded religious university nearby that sends the resumes of highly qualified students to the company. Historically, these candidates have highly relevant skill sets and make up a majority of applications from the religious group.

*There are many possible conclusions that could be drawn, so any thoughtful and well explained response will receive credit.*

3. (15 points) **Visualizing calibration by group**. Consider a binary classification problem where $X$ denotes the input variable, $Y \in \{0, 1\}$ denotes the output variable, and we have a score function $R(X)$ with which we construct a decision rule.

Recall from section that we call a score function $R(X)$ that takes on values in $[0, 1]$ *calibrated* if, for all $r \in [0, 1]$,

$$\mathbb{P}(Y = 1 \mid R(X) = r) = r.$$

Furthermore, for an attribute $A$ that we care about, we call $R(X)$ *calibrated by group* if, for all $r \in [0, 1]$ and values $a$ that $A$ can take on,

$$\mathbb{P}(Y = 1 \mid R(X) = r, A = a) = r.$$

In practice, we do not know these probabilities. However, we can estimate them given a dataset. In this problem you'll implement and interpret a common plot used to visualize how calibrated (or not) a score function is, known as a *calibration plot.*

(a) (5 points) We'll create calibration plots for a dataset used to predict whether a patient has heart disease. Download the data file `heart.csv` from the course website. The file consists of three columns:

- Column 1 contains the scores (between 0 and 1) of the $n$ patients, output by a logistic regression model trained to predict whether a patient has heart disease given several medical attributes.
- Column 2 contains the true binary label $Y$, indicating whether or not the patient has heart disease.
- Column 3 contains a binary label $A$ indicating whether the patient is male (1) or female (0).

Write a function `calibration_plot()`, which takes as input

- an array of probabilities `r` of length $m$, such as `np.arange(0, 1.1, 0.1)`
- an array of scores of length $n$
- an array of true binary labels of length $n$

and outputs an array `p` of length $m - 1$, where the $i$-th element contains

$$\hat{\mathbb{P}}(Y = 1 \mid R(X) \in [\texttt{r[i]}, \texttt{r[i + 1]}))$$

The notation $\hat{\mathbb{P}}$ indicates that this is an estimate, where we compute the fraction of instances with $R(X) \in [\texttt{r[i]}, \texttt{r[i + 1]})$ that have the true label $Y = 1$.

(b) (5 points) Run your implementation of `calibration_plot()` on the downloaded data, to produce an output array `p` of length $m-1$. Plot `r` against `p`, with `r` on the $x$-axis and `p` on the $y$-axis. This is called a calibration plot.

A perfectly calibrated score function has a calibration plot that lies on the $y = x$ diagonal line. The more the calibration plot deviates from this line, the less calibrated the score function is. To help you judge this, on top of the calibration plot, plot the $y = x$ diagonal line. (For example, if using `matplotlib.pyplot` in python, `plt.plot([0, 1], [0, 1], "--k")`.)

1. Is this score function close to calibrated?
2. Pick any point (`r[i]`,`p[i]`) on your calbration plot. State the values (`r[i]`,`p[i]`), and interpret what they mean.

(c) (5 points) Note that for any value $a$ that an attribute $A$ can take on, you can run `calibration_plot()` on only the scores and labels corresponding to data points where $A = a$. This allows you to compute calibration plots specific to a group. For example, let $A$ denote gender. Compute and plot a calibration plot for the male patients, where the $i$-th element of the output is

$$\hat{\mathbb{P}}(Y = 1 \mid R(X) \in [\texttt{r[i]}, \texttt{r[i + 1]}), A = 1 \text{ (male)})$$

and a calibration plot for the female patients, where the $i$-th element of the output is

$$\hat{\mathbb{P}}(Y = 1 \mid R(X) \in [\texttt{r[i]}, \texttt{r[i + 1]}), A = 0 \text{ (female)}).$$

Compare the two plots with each other, along with the first plot you generated in the previous part. Is this score function close to calibrated by group? Is it more calibrated on one group than another?

4. (25 points) **Linear Regression and the Gauss-Markov Theorem** This problem explores the implications of the Gauss-Markov Theorem for the ordinary least-squares (OLS) estimator. We consider $n$ data points $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$ and their corresponding target values $y^{(1)}, \ldots, y^{(n)} \in \mathbb{R}$. Our model assumes

$$y^{(i)} = \langle \beta^*, x^{(i)} \rangle + \epsilon^{(i)}$$

where the errors $\epsilon^{(i)}$ are independently distributed. Recall that the OLS estimator is given by

$$\hat{\beta} = \left( \frac{1}{n} \sum_{i=1}^{n} x^{(i)} x^{(i)T} \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} x^{(i)} y^{(i)} \right).$$

The Gauss-Markov theorem states that if the $\epsilon^{(i)}$ are independent from each other and $\mathbb{E}[\epsilon^{(i)} | x^{(i)}] = 0$ for all $i$, then

1. $\hat{\beta} = \beta^*$ if $n = \infty$
2. $\mathbb{E}[\hat{\beta}] = \beta^*$ if $n$ is finite. This means the estimator is unbiased.
3. If $\mathrm{Var}[\epsilon^{(i)}]$ is the same for all $i$, $\hat{\beta}$ has the lowest variance among all unbiased estimates of $\beta^*$. A common way to say this is that the OLS estimator is the *best linear unbiased estimator* (BLUE), where "best" means lowest variance.

(a) (20 points) For each of the following two models, state whether the Gauss-Markov theorem allows us to conclude that $\hat{\beta}$ is an unbiased estimate of $\beta^*$, and explain why or why not by specifying which assumptions are satisfied and not satisfied. For all models, the $z^{(i)}$ are independent of each other and of the $x^{(i)}$.

1. $y^{(i)} = \langle \beta^*, x^{(i)} \rangle + \sin(x^{(i)}) \cdot z^{(i)}$, where $z^{(i)} \sim N(0, 1)$.
2. $y^{(i)} = \langle \beta^*, x^{(i)} \rangle + z^{(i)2}$, where $z^{(i)} \sim N(0, 1)$.
3. $y^{(i)} = \cos(\langle \beta^*, x^{(i)} \rangle) + z^{(i)}$, where $z^{(i)} \sim N(0, 1)$.
4. $y^{(i)} = \langle \beta^*, x^{(i)} \rangle + z^{(i)}$, where $z^{(i)} \sim N(0, |x^{(i)}|)$.
5. $y^{(i)} = \langle \beta^*, x^{(i)} + z^{(i)} \mathbb{1} \rangle$, where $z^{(i)}$ takes on the value $-1$ with probability $1/2$ and the value 1 with probability $1/2$. $\mathbb{1}$ is the all-ones vector, so $x^{(i)} + z^{(i)} \mathbb{1}$ is equivalent to adding $z^{(i)}$ to $x^{(i)}$ element-wise.

(b) (5 points) Consider the intercept-only model, where $d = 1$ and $x^{(i)} = 1$:

$$y^{(i)} = \beta^* + \epsilon^{(i)}$$

What is the OLS estimator $\hat{\beta}$ for the intercept-only model?