Homework 2

Note: The homework is due on Tuesday, February 26th by 1:30 pm (to be submitted via Gradescope). If you choose to use some of your 5 available late days, please remember to submit by 1:30 pm on that day.

Please write concise and clear solutions typeset in LaTeX. You are allowed to discuss ideas for solving homework problems in groups of up to 3 people but *you must write your solutions independently*. Also, you must write on your homework the names of the students with whom you discussed.

In this assignment, you will prove an "impossibility result" for the compatibility of multiple definitions of fairness. Here, we consider a population P of people who are each represented by a feature vector x, a target attribute $y \in \{+1, -1\}$, and a group label $g \in 1, 2$. We also assume that some hypothesis h is being used to generate labels in $\{+1, -1\}$ for every person in P.

The table below describes the types of decisions which can be made by h for the population. Each value in the table represents some numerical count, corresponding to the number of people in P with that label and prediction result ($|P| = A + B + C + D$).

True Labels vs Predictions for a Population		
	True label = (+)	True label = (-)
Prediction = (+)	A (true positives)	B (false positives)
Prediction =(-)	C (false negatives)	D (true negatives)

The population can be broken down by group into P_1 and P_2, where $|P| = |P_1| + |P_2|$. Similarly, we can divide each quantity in our table by group label. We can express the true positives in group 1 as A_1 and in group 2 as A_1, where $A = A_1 + A_2$. Similarly, we define $B = B_1 + B_2$, $C = C_1 + C_2$, and $D = D_1 + D_2$.

Across the problems in this homework, you will show that it is impossible to satisfy a set of three fairness notions for groups 1 and 2 simultaneously, *regardless of the hypothesis used*, unless extreme conditions are met.

Three Notions of Fairness

Three notions of fairness we might want to consider are:

1. Equality of false positive rates, where the *false positive rate* for the population can be expressed as:
$$FPR = \frac{B}{B + D}$$
2. Equality of false negative rates, where the *false negative rate* for the population can be expressed as:
$$FNR = \frac{C}{A + C}$$
3. Equality of positive predictive value, where the *positive predictive value* for the population can be expressed as:
$$PPV = \frac{A}{A + B}$$

We have discussed the first two definitions in class; you can think of positive predictive value as being a measure of the accuracy of the *positive predictions* made by a hypothesis. For us to say that the quantities are satisfied, it must be the case that the corresponding values are equalized after substituting in the quantities from the table for both groups. For example, equality of false positive rates would imply that

$$FPR_1 = \frac{B_1}{B_1 + D_1} = \frac{B_2}{B_2 + D_2} = FPR_2$$

You will show that it is impossible to satisfy all three fairness notions at the same time, unless one of the following conditions holds:

- The hypothesis makes no false predictions
- The base rates of the population are identical

Recall that the base rate of the population does not depend on a hypothesis, and is simply the proportion of the population whose true labels are positive. Using the values in the table above, this would be:

$$BR = \frac{A + C}{|P|}$$

Problem 1. (10 points)

Show that $(1 - FNR) = \frac{A}{A+C}$.

Problem 2. (10 points)

Show that $(1 - PPV) = \frac{B}{A+B}$.

Problem 3. (15 points)

Show that $\frac{BR}{1-BR} = \frac{A+C}{B+D}$.

Problem 4. (15 points)

Using the results from Problems 1, 2, and 3, show that:

$$FPR = \frac{BR}{1 - BR} \cdot \frac{1 - PPV}{PPV} \cdot (1 - FNR)$$

.

Problem 5. (25 points)

The statement from Problem 4 holds for the entire population as well as for each group individually. Suppose that all three fairness notions are satisfied by our hypothesis, i.e. $FPR_1 = FPR_2$, $FNR_1 = FNR_2$, and $PPV_1 = PPV_2$. Further, assume that all of these values, as well as the base rates, are neither 0 nor 1. Show that this implies that the base rates of the groups must be equal.

Problem 6. (25 points)

Show that if our hypothesis makes no mistakes (i.e. $B_1 = B_2 = 0$ and $C_1 = C_2 = 0$), all three fairness notions will be satisfied, regardless of the base rates for each group.