School of Computing and Information Systems
The University of Melbourne
COMP90049 Introduction to Machine Learning (Semester 1, 2021)

Workshop: Week 12

1. Using the Adult dataset from assignment2, we want to judge the fairness of a classifier we have trained on it. In this dataset each instance *X* is a person, represented through a range of demographic attributes (gender, origin, education, ...). The target variable *Y* is the income of the person (>50K or <=50K).

   (i). Discuss the following concepts in the context of this data set.

   a) Historical Bias

   b) Demographic disparity

   c) Using the system in the context of (1) a bank which wants to use a model trained on this data for predicting credit ratings; and (2) a government institution in Australia which has access to the features of the Adult for a small population of Australians and wants to predict their income based on it.

   (ii). You are asked to develop an income classifier that is fair with respect to the protected attribute *gender*. Your boss is a big believer in logistic regression classifiers and asks you to apply this particular classifier architecture with no modification. What approach(es) could you take to still test/improve the performance of your classifier?

2. Using the dataset in question 1, assume we selected *gender* as our protected attribute. We trained our classifier and observed the following outcomes. The label y=1 means "income >50K", and y=0 means "income <=50K.

P(ŷ=1|A=f)	P(ŷ=1|A=m)	P(ŷ=1|Y=1, A=f)	P(ŷ=1|Y=1, A=m)	P(Y=1|ŷ=1, A=f)	P(Y=1|ŷ=1, A=m)	P(Y=1|ŷ=1)	P(ŷ=1| Y=1)
0.81	0.75	0.80	0.86	0.73	0.74	0.74	0.85

   (i). Name each of the statistics and provide a formula for its measurement. Be sure you understand the intuition / connection behind the statistical notion and its metric.

   (ii). For each of the following criteria, decide whether the classifier meets this criterion.

   a) Group Fairness (Demographic parity)

   b) Equal opportunity

   c) Predictive parity

3. A common metric for assessing classifier fairness is the GAP in scores achieved across groups. If we choose true positive rate (TPR) as our score of interest, we will check the classifier for "equal opportunity". If we choose positive predictive value as score of interest, we test our classifier for "predictive parity". Verify your observations in question 2 using (a) max-GAP and (b) avg-GAP. When would avg-GAP be preferred, and when max-GAP?

4. For our classifier above, we reported that TPR_f=0.8, TPR_m=0.86 and TPR=0.85 (cf. Columns 3, 4 and 8 in the table). How do you think TPR was computed, and what does it tell us about the data?