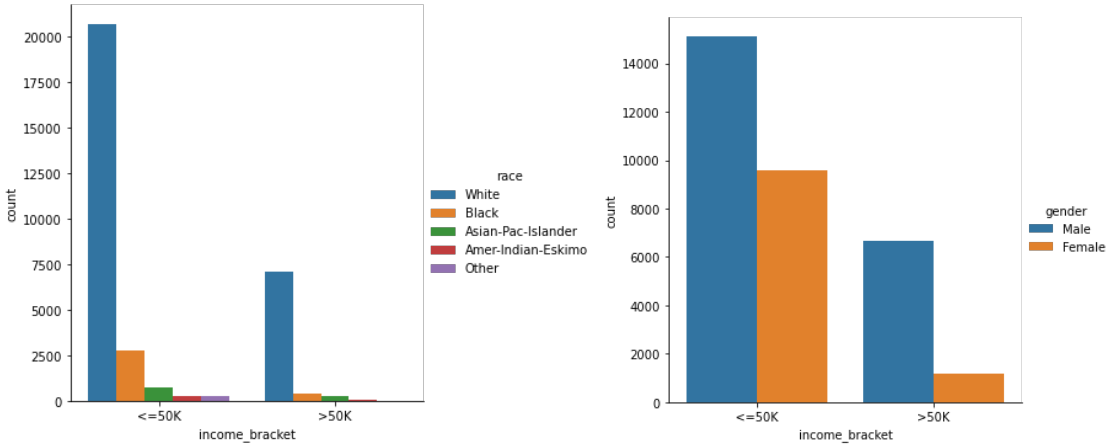
School of Computing and Information Systems
The University of Melbourne
COMP90049 Introduction to Machine Learning (Semester 1, 2023)

Week 12: Sample Solutions

1. We have dataset containing demographic and income data from United States in 1994. We want to judge the fairness of a classifier we have trained on it. The data set consists of about 48,000 individuals, where each instance X is characterized through a range of 14 demographic attributes (gender, origin, education, race, occupation, etc.). The target variable Y is the income of the person (>50K or <=50K). To give you a snapshot of the training data, we have plotted the distribution of income across different genders and races.



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

a) Historical Bias

The data set presumably reflects the demographic reality in (parts of) the USA in 1994, however, society changes! In the 1990s presumably fewer women were among the working population; social disparity between different demographic groups was different than today; and education of society in general went up. A classifier trained on the data would simply reflect the predictive patterns from the past – a ML model cannot differentiate between "useful knowledge" and "spurious (or unstable) correlations". It would use all statistical patterns in the data set, and base its predictions on it, and potentially propagate them into the future.

b) Demographic disparity

Machine learning models learn to generalize on the basis of individual observations (labelled training instances). Clearly, the more frequent a model observes a certain pattern (e.g., feature-label combination) the more it will impact the generalisation the model derives. As such: generalizations are to a large extent based on the majority classes in the data set. In the context of the data set above, consider the "ethnicity (race)" feature. Presumably, "white" is the majority class, and the classifier will make generalisations between education and income largely based on this majority class – however this relation may look very different for different ethnicities.

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.

Both applications must be viewed with caution because the function of X (features) → Y (labels) changes between the training scenario and test scenario.

A classifier that's very good at predicting income levels, can probably not be used "off-the-shelf" to predict the credit score of an applicant! Credit worthiness depends on other features as well, and the statistical relation between the given features and the new label may be different. When we deploy a ML algorithm, we must make sure that the task we are solving is the same that the model was trained on.

The relation between demographic variables and income are presumably quite different across countries, or continents! There is little reason to expect that the model trained on the adult data fares well on a population of Australians. Again: we need to make sure that the data distribution our model observes when it is deployed is (as) similar (as possible) to the data distribution it was trained on.

(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?

The lectures covered three approaches to improving fairness of classifiers: (1) changing the data; (2) changing the ML model/loss; (3) post-processing of the model predictions. Our boss prohibited option (2), so options (1) and (3) are left.

For option (1), pre-processing the data, we could re-sample the data set such that both groups are represented similarly in the data set, for example by down-sampling instances of the majority class. In this way, our classifier would base its generalisations on both groups, rather than focussing on the majority group disproportionately.

Similarly, we could assign each instance a weight and penalize model errors for instances with higher weight more than instances with lower weight. We would compare the true probability of observing a label with a protected group against the expected probability if the two were independent.

Finally, we could leave the data untouched, and instead post-process the classifier output. Rather than applying a decision threshold of 0.5 ($\hat{y}$=1 if score >0.5 and $\hat{y}$=0 if score <=0.5) we could devise a group-specific threshold for female and male applicants. This threshold could for example be set such that for both genders we expect a comparable number of false negative predictions under our already trained classifier (which corresponds to equal opportunity).

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($\hat{y}$=1|A=f)	P($\hat{y}$=1|A=m)	P($\hat{y}$=1|Y=1, A=f)	P($\hat{y}$=1|Y=1, A=m)	P(Y=1|$\hat{y}$=1, A=f)	P(Y=1|$\hat{y}$=1, A=m)	P(Y=1|$\hat{y}$=1)	P($\hat{y}$=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.

- Positive rate is the fraction of (true or false) positives predicted for each group
- True Positive rate is the fraction of positives among all positives in the data (recall)
- Positive predictive value is the fraction of true positives among all positive predictions (precision)

All of these can be computed for individual groups (hence the conditioning) or for the overall population (last two columns)

$P(\hat{y}=1\|A=f)$	$P(\hat{y}=1\|A=m)$	$P(\hat{y}=1\|Y=1, A=f)$	$P(\hat{y}=1\|Y=1, A=m)$	$P(Y=1\|\hat{y}=1, A=f)$	$P(Y=1\|\hat{y}=1, A=m)$	$P(\hat{y}=1\| Y=1)$	$P(Y=1\|\hat{y}=1)$
0.81	0.75	0.80	0.86	0.73	0.74	0.85	0.74
Female Positive rate	Male Positive rate	Female True positive rate; Recall	Male True positive rate; Recall	Female True positive predictive value; precision	Male True positive predictive value; precision	Recall; True Positive Rate (overall)	Precision; Predictive Parity Value (overall)
$PR_f = \frac{P}{N_f}$	$PR_m = \frac{P}{N_m}$	$TPR_f = \frac{TP_f}{TP_f + FN_f}$	$TPR_m = \frac{TP_m}{TP_m + FN_m}$	$PPV_f = \frac{TP_f}{TP_f + FP_f}$	$PPV_m = \frac{TP_m}{TP_m + FP_m}$	$TPR = \frac{TP}{TP + FN}$	$PPV = \frac{TP}{TP + FP}$

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

a) Group Fairness (Demographic parity)

Demographic parity requires that the positive rate of all groups is identical (i.e., the first two columns in our case). There is a substantial gap: the fraction of positive predictions for females (0.81) is larger than the fraction of positive predictions for males (0.75). Our classifier does not achieve group fairness.

b) Equal opportunity

Equal opportunity requires that the true positive rates are identical across groups: Intuitively, if we know that an individual **is** credit worthy, our classifier should predict the individual as credit worthy with the same probability – irrespective of the gender of the applicant.

We compare columns 3 and 4 in the table above and find that our classifier does not achieve equal opportunity: the TPR is higher for males (0.86) than for females (0.80).

This means that our classifier is more likely to correctly grant credit to a man (who indeed is credit worthy) than to a woman (who indeed is credit worthy).

Or, equivalently: our classifier is more likely to falsely deny a credit to a woman (who indeed is credit worthy) than to a man (who indeed is credit worthy). This is because False Negative Rate is also different across the group. We calculate FNR as

FNR = 1-TPR

c) Predictive parity

Predictive parity requires that positive predictive values are identical across groups.

Intuitively, if we have predicted a positive rating for an applicant, it should be equally likely that the applicant is indeed credit worthy – irrespective of the gender.

We compare columns 5 and 6 in the table above and find that both values are similar (if not identical). Although strictly speaking the classifier doesn't exhibit perfect predictive parity, for any realistic application reducing differences to below a certain threshold is typically sufficient.

For almost all scenarios, PPV values of 0.73 and 0.74 for male and female would be considered fair. Our classifier does achieve 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?

We compute the average across all groups g, as following where ϕ_g denotes a group-specific score (TPR or PPV) and ϕ denotes the overall value.

$$GAP_{avg} = \frac{1}{G}\sum_{g=1}^{G}|\phi_g - \phi|$$

For Max GAP we select the single maximum value pertaining to any group.

$$GAP_{max} = max_{g\in G}|\phi_g - \phi|$$

For our classifier in question 2, we can calculate GAP_{avg} and GAP_{max} for TPR and PPV separately.

TPV:

$$GAP_{avg} = \frac{1}{2}(|0.8 - 0.85| + |0.86 - 0.85|) = \frac{1}{2}(0.5 + 0.01) = 0.03$$

$$GAP_{max} = max(|0.8 - 0.85| + |0.86 - 0.85|) = max(0.5 + 0.01) = 0.5$$

PPV:

$$GAP_{avg} = \frac{1}{2}(|0.73 - 0.74| + |0.74 - 0.74|) = \frac{1}{2}(0.01 + 0.00) = 0.005$$

$$GAP_{max} = max(|0.73 - 0.74| + |0.74 - 0.74|) = max(0.01 + 0.00) = 0.01$$

Unsurprisingly, both gap values are smaller for PPV in compare with TPV.

When to use avg-GAP and when max-GAP? Think about a scenario where we have 5 different sensitive groups (e.g., different ethnicities), and it is very important that no single ethnicity is treated unfairly. The avg-GAP could look reasonable if some ethnicities achieve far larger TPR than the overall measure and others achieve much lower TPR. The max-GAP on the other hand would capture every individual outlier. So: in situations where outliers are really unacceptable, max-GAP should be used.

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?

Recall from earlier lectures, that precision and recall (TPR) are class-specific measures and need to be combined across groups in some way. We talked about macro averaging and micro averaging.

Macro average computes the metric individually for each group, and then averages. It treats each group identical, irrespective of its size.

Micro average aggregates count (TP, FN, etc.) first, before the metric is computed. As such, it takes into account class imbalances.

For the numbers above we can conclude that micro average was used (as TPR is not the average of the individual scores).