Question 1 (5 points)

For a given classifier, let C denote the classifier output, Y denote the correct output, D denote the sensitive characteristic, X denote the input features, and $s(\cdot)$ denote the score function.

Calibration (3 points)

If the score function is used for the classifier's output:

$$p(Y = 1|s(X) = S, D = d) = p(Y = 1|s(X) = S, D = d') \quad \forall S \in \text{range}(s) \tag{1}$$

where $S \in \mathcal{R}$ is any value the score function takes; d and d' are two values the feature D takes.

If the classifier's score output is assumed to be binary, the following results in:

$$p(Y = 1|c = 1, D = d) = p(Y = 1|c = 1, D = d') \quad \text{and} \tag{2}$$
$$p(Y = 1|c = 0, D = d) = p(Y = 1|c = 0, D = d') \tag{3}$$

False Positive Parity (2 points)

$$p(C = 1|Y = 0, D = d) = p(C = 1|Y = 0, D = d') \tag{4}$$

Marking Scheme

For both calibration and FPR:

- Full marks for correct formulas.
- Partial marks for minor typos.
- No mark for incorrect formulas.

Question 2 (5 points)

Example Solution

Basic Idea

The base rates for the different demographics must be equal.

Dataset

Two populations A and B. People with no prior arrests have a 10% chance of arrest, and people with prior arrests have a 50% chance of re-arrest. In both A and B, 50% of people have prior arrests, and 50% don't.

Classifier

Predict 0 if no prior arrests, 1 if there are prior arrests.

Calibration

By the definition of the population and our classifier rule:

$$p(Y = 1|C = 1, D = d) = p(Y = 1|C = 1, D = d') = 0.5 \tag{5}$$
$$p(Y = 1|C = 0, D = d) = p(Y = 1|C = 0, D = d') = 0.1 \tag{6}$$

False Positive Rate

$$p(C=1|Y=0,D=d) = \frac{p(C=1,Y=0|D=d)}{p(Y=0|D=d)} \tag{7}$$

$$= \frac{p(Y=0|C=1,D=d)p(C=1|D=d)}{p(Y=0|D=d)} \tag{8}$$

$$= \frac{0.5 \times 0.5}{0.5 \times 0.1 + 0.5 \times 0.5} \tag{9}$$

The false positive rate is the same regardless of demographics.

Making scheme

- 1 point for setting base rates to be the same.
- 1 point for specifying a sensible classifier.
- 2 points for proving that false positive rate parity and calibration are satisfied, explicitly using the fact that the base rates are the same.
- 1 point for an explanation that makes sense.

Question 3

Suppose we observe sex and red car and aggressiveness (aggr) predicts both car colour and accidents. We have something like:

$p(redcar = 1) = \sigma(sex + aggr)$

$p(accident = 1) = \sigma(aggr)$

If we predict accidents using only red car, the classifier won't satisfy calibration: we will over-predict acc for $sex = 1$, and under-predict acc for $sex = 0$. If we control sex, the classifier can be calibrated since we can correctly predict aggr if we have both sex and red car.

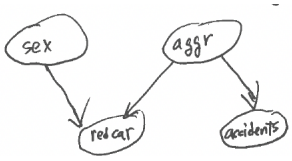

Figure 1: model

Marking scheme

- 4 pts for the idea of omitted variable bias
- 3 pts for correctly saying which fairness criterion is not satisfied and why
- 3 pts for explaining why adding the variable can make the problem go away

Question 4

Consider the following probability model where we imagine we measure the temperature at different altitudes during different seasons.

$p(low = 1) = 0.5$

$p(low = 0) = 0.5$

$p(summer = 1) = 0.25$

$p(summer = 0) = 0.75$

$p(hot = 1) = \sigma(summer + low \times 0.1 - 1)$

Explaining away refers to variables that cause a variable Z becoming anti-correlated when Z is observed. $p(summer = 1|hot = 1, low = 1) < p(summer = 1|hot = 1)$

Making scheme

- 3 pts: plausible probability distribution
- 7 pts: good explanation including an expression of probabilities involving the example

Question 5

Cost function:

$$J = \sum_{i,j} f(X_{ij})(u_i.v_j + b_i + b'_j - \log(X_{ij}))^2 \tag{10}$$

Figure 2: f function

0.1 Intuition 1

Want the similarity of words i and j to correspond to the number of co-occurrences of the words so:
Make $u_i \cdot v_j$ approximate $\log(X_{ij})$
Do not want frequent words to have an out-sized influence on the embeddings, so limit the influence of high co-occurrence counts using f.

0.2 Intuition 2

Want the embeddings to reflect the probability of word i conditional on seeing word j so make $P(Word_i|Word_j) \approx \log(X_{ij}/X_i) \approx u_i \cdot v_j$. Absorb (X_i) into the biases, to get $(u_i.v_j + b_i + b'_j - \log(X_{ij}))^2$
Intuition for f is same as above.

Marking scheme

- 4 pts for formula
- 6 pts for intuition (2 pts for f + 4 pts for either intuition 1 or 2

Question 6

The training set distribution is the probability distribution implied by the training set. That is, if we generated examples from the training set distribution, we'd get examples that look like examples from the training set.
The generator distribution is the distribution implied by the generator. That is, samples from the distribution can be obtained using
$Z \sim \mathcal{N}(0, I)$
$S = G_\theta(Z)$.

Marking scheme

- 2 pts for training distribution
- 3 pts for generator distribution (at most 1 pt if $z \sim N(0, I)$, $sample = G_\theta(z)$ is not in the explanation

Question 7

We first derive the optimal discriminator given a generator G. The optimal discriminator optimizes the cost function:

$$\begin{aligned} C(G) = \max_D V(G, D) &= \max_D E_{X \sim P_{\text{data}}} \log D_G(x) + E_{Z \sim P_z} \log(1 - D_G(G(z)) \\ &= \max_D E_{X \sim P_{\text{data}}} \log D_G(x) + E_{X \sim P_G} \log(1 - D_G(x)) \\ &= \max_D \int_x \left[P_{\text{data}}(x) \log D_G(x) + P_G(x) \log(1 - D_G(x)) \right] dx \end{aligned}$$

We will now maximize the integrand with respect to $D_G(x)$:

$$\begin{aligned} f(y) &= a \log b + b \log(1 - y) \Rightarrow \\ f^{'}(y) &= \frac{a}{y} - \frac{b}{1 - y} \Rightarrow \\ y &= \frac{a}{a + b} \end{aligned}$$

So, $D_G^*(x) = \frac{P_{\text{data}}(x)}{P_{\text{data}}(x) + P_G(x)}$ is the optimal discriminator
So, when optimizing the generator, we are optimizing:

$$\begin{aligned} C(G) &= E_{X \sim P_{\text{data}}} \log \frac{P_{\text{data}}(x)}{P_{\text{data}}(x) + P_G(x)} + E_{X \sim P_G} \log \frac{P_G(x)}{P_{\text{data}}(x) + P_G(x)} \\ &= E_{X \sim P_{\text{data}}} \log \frac{P_{\text{data}}(x)}{\frac{P_{\text{data}}(x) + P_G(x)}{2}} - \log 2 + E_{X \sim P_G} \log \frac{P_G(x)}{\frac{P_{\text{data}}(x) + P_G(x)}{2}} - \log 2 \\ &= KL(P_{\text{data}} || \frac{P_{\text{data}} + P_G}{2}) + KL(P_G || \frac{P_{\text{data}} + P_G}{2}) - \log 4 \\ &= 2 \; JS(P_{\text{data}} || P_G) - \log 4 \end{aligned}$$

So minimizing $C(G)$ is the same as minimizing $JS(P_{\text{data}} || P_G)$

Marking scheme

For full points, there should be a clear explanation of what is going on.

Question 8

Node2vec learns the embedding z_u for each node u.

We start out with wanting to maximize the probability of u and v co-occurring according to the model if v occurs on a random walk starting with u. We express this probability as:

$$\frac{\exp(z_u \cdot z_v)}{\sum_{n \in V} \exp(z_u \cdot z_u)}$$

We are trying to make it so the embeddings are close (understood as making $z_u \cdot z_v$ high)

Instead of maximizing $\sum_u \sum_{v \in N(u))} \log \frac{\exp(Z_u Z_v)}{\sum_{u \in V} (\exp(Z_u Z_m)}$ we try to maximize the numerators and minimize an estimate for the denominators. An objective function for a minibatch can look like

$$L_u = \sigma(z_u \cdot z_v) - \sum_{i=1}^{K} \sigma(z_u \cdot z_{n_i}) \tag{11}$$

we sample the n_i according to the degree of nodes so that more connected nodes have more influence.

We then maximize the L_u's.

Marking scheme

- 3 pts for the correct function to optimize
- 2 pts for saying we optimize probability of co-occurrence and giving the formula for the probability
- 3 pts for connecting maximizing $z_u \cdot z_v$, the embeddings being similar, and u and v co-occurring
- 2 pts for explaining that the negative sampling formula is an approximation

Question 9

Permutation invariant: $f(A) = f(B)$ if graph A is an rearrangement of graph B
Permutation equivariant: If i and j are permuted between A and B, $f(A)_i, f(B)_i = f(B)_j, f(A)_j$

Marking scheme

- 1 point each for "invariant" and "equivariant"

Question 10

If we stack every embedding, computing the embeddings is *Permutation-equivariant* because the embeddings depend on what the neighbours are, and not on their order.

Marking scheme

- Only accept answers with an explanation
- If the explanation makes sense, also accept invariant+equivariant