Homework Out: February 4
Due Date: February 18, midnight
Reminder:
To submit your homework, please go to `https://classroom.github.com/a/7rslWmMd`, accept the assignment, and submit your LaTex, PDF, and any code you use for the assignment. Please name your files "hw1-USERNAME-writeup.tex,pdf" "hw1-USERNAME-code.appropriate file type". You will need a github account, and to add, commit, and push your homework.
Please cite all sources you use, and people you work with. The expectation is that you try and solve these problems yourself, rather than looking online explicitly for answers. Submissions due at 23:00 of the due date.
You may use O-notation unless explicitly noted somewhere in the homework.

Problems

1. **(Statistical Guarantees aren't Created Equal.).**

   Consider some $f : \mathcal{X} \to [0, 1]$ and distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathcal{X}$ such that

   $$\mathbb{E}_{x\sim\mathcal{D}_1}[f(x)] = \mathbb{E}_{x\sim\mathcal{D}_2}[f(x)]$$

   Consider two datasets X_1, X_2 drawn independently as $x \sim \mathcal{D}_i, x \in X_i$, with $|X_1| = n_1 > n_2 = |X_2|$.

   (a) As a function of n_1, n_2, give an upper bound on

   $$|\frac{1}{n_1}\sum_{x\in X_1} f(x) - \frac{1}{n_2}\sum_{x\in X_2} f(x)|$$

   which holds with probability $1 - \delta$ over the draw of X_1, X_2. You can use "known" concentration bounds (e.g., Chernoff-Hoeffding).

   (b) Define two distributions D_1, D_2 and a function f such that the expected value of f on a draw from either distribution is equal. Define these such that you can get a large lower bound on the probability below.
   Lower bound the probability with which

   $$\frac{1}{n_1}\sum_{x\in X_1} f(x) \geq \frac{1}{n_2}\sum_{x\in X_2} f(x) + \frac{1}{\sqrt{n_2}}.$$

   Notice this describes the probability that two empirical averages differ by the *larger* of their variances.

2. **(Distortion vs Information Theory.).**

   The paper we read ("On the (Im)possibility of Fairness") for the first class suggested we consider distortion as a measure of WYSIWYG (how much our observed features represent the construct space for different groups). Let $\mathcal{X}, d_{\mathcal{X}}$ represent the construct space and its metric, and $\mathcal{Y}, d_{\mathcal{Y}}$ the observed space and metric, respectively.

   (a) Describe some $\mathcal{X} \subseteq \mathbb{R}^n$ and f comprised of some linear measurements of $\mathcal{X}$, f transforming $\mathcal{X}$ to $\mathcal{Y} \subseteq \mathbb{R}^{n'}$ such that the distortion of f is of order $\max_{x\in\mathcal{X}} ||x||_2$, but the accuracy of the *best* linear classifier is no worse in $\mathcal{Y}$ than in $\mathcal{X}$.
   You can assume $d_x = d_y$ is the ℓ_2 metric.

   (b) Find some $\mathcal{X}$, f, and two groups $\mathcal{X}_1 \cup \mathcal{X}_2 = \mathcal{X}$ such that the distortion of f on $\mathcal{X}_1$ (ignoring points in $\mathcal{X}_2$) and the distortion of $\mathcal{X}_2$ (ignoring points in $\mathcal{X}_1$) are equal, but the best linear classifier for $\mathcal{X}$ has 0 error, the best linear classifier for $f(\mathcal{X}_1)$ has error $\frac{1}{2}$, and the best linear classifier for $f(\mathcal{X}_2)$ has error 0.

3. **(Coding: Training together, training apart.).**

   Find a human-centric dataset[1] which is publicly available with demographic information (age, gender, race, sexual orientation, country of origin, etc).

[1] A dataset such that each entry represents one measurement of one person.

(a) Please include a link to the dataset you used, as well as any documentation that accompanied its release. Clearly describe any "cleaning", binning, bucketing, or discrete-to-continuous feature transformation you did in this process.

(b) For one way of splitting the dataset into different demographic groups, write down the size of the groups, the average value for each (numeric) feature, the variance of each (numeric) feature, the mode for each categorical feature, and the three most frequent values for each categorical feature, each computed on the different demographic subgroups.

If your dataset has more than 20 features, you can report this information only for 20 features.

Do you observe any interesting differences between the different subgroups' statistics?

(c) Now, randomly subsample half of the dataset. That is, include each row in the database in some new dataset with probability $\frac{1}{2}$ independently. Report the same statistics as above.

Do you observe some of the statistics are more "stable" than others (namely, that their values changed little from the previous question)? Do some subgroups have more stable statistics than others?