Homework Out: February 28
Due Date: March 16, midnight
Reminder:
To submit your homework, please go to gradescope and submit your LaTex, PDF, and any code you use for the assignment. Please name your files "hw2-USERNAME-writeup.tex,pdf" "hw2-USERNAME-code.appropriate file type".
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. **(Uniform Noise Won't Help You.).**

   One of the suggested "solutions" for a dataset with far more positive labels for group g_1 than group g_2 suggested by "Data preprocessing techniques for classification without discrimination", *massaging*, suggests flipping some fraction of the labels of the datapoints in g_1 from negative to positive.

   Suppose the current dataset is *linearly separable* (namely, there is some linear classifier with accuracy 1 in training).

   (a) Describe a set of points in $\mathbb{R}^2$, broken into groups g_1, g_2:
   - with equally sized groups g_1, g_2
   - with 50% positive and negative labels for group g_1
   - 40% positive labels for g_2
   - which is linearly separable

   Then, find a subset of g_2's negative labels (large enough give the groups have equal number of positive labels if flipped) that, if swapped, would not change the set of accuracy-maximizing linear separators in training.

   (b) Do the same, but rather than picking a carefully designed set of points in g_2, argue that a *uniformly random* subset of negatively labeled points in g_2 (of the appropriate size) has low probability of changing the set of accuracy-maximizing linear classifiers. Give an estimate of this probability.

2. **(Interventions, Adapted from Gelman and Hill (2006), Chapter 9).**

   An observational study to evaluate the effectiveness of supplementing a reading program with a television show was conducted in several schools in grade 4. Some classroom teachers chose to supplement their reading program with the television show and some teachers chose not to supplement their reading program. Some teachers chose to supplement if they felt that it would help their class improve their reading scores. The study collected data on a large number of student and teacher covariates measured before the teachers chose to supplement or not supplement their reading program. The outcome measure of interest is student reading scores.

   (a) Describe how this study could have been conducted as a randomized experiment.

   (b) Is it plausible to assume that supplementing the reading program is exchangeable in this observational study?