

Homework Out: October 24

Due Date: Novemer 9, midnight

Reminder:

To submit your homework, please go to https://classroom.github.com/a/_JTnmwoc, accept the assignment, 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”. 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. (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. (Project Outline.)

Please describe what you'd like to do for your course project. Describe the motivation for this project, what concrete tasks you plan to do, and what the final work product of these tasks will be. Is there a particular context you're thinking about (e.g., hiring, advertising) as you formulate this question? If you're planning to do an empirical project, list the dataset(s) you will use or gather.

It isn't necessary that you do precisely what you outline above for the final project, but it should help you start thinking about your work to answer these questions.