Project Part 2: Dataset Review
CSE 446: Machine Learning

University of Washington

Deadline: ~~October 26, 2017~~ October 30, 2017

October 23, 2017: updates in red.

For each part of the project, your team will be evaluated as a whole; you will share a grade. The second task for your team to complete is to write a brief review of three datasets constructed by other teams, which we will assign to you.

For each of the three datasets we assign your team, work together to answer these questions:

1. Does the dataset instantiate a binary classification problem, or a regression problem? If neither, the dataset will be disqualified (please report this), but you should still answer the other questions.
2. Summarize in a short paragraph what the machine learning problem is.
3. Does the team make a convincing case that y is predictable from x? If you believe there is anything they should have taken into consideration, but didn't, explain.
4. Does the team make a convincing case on the matter of ethics? Do you see any ethical reasons why this dataset shouldn't be used by the class?
5. Are you concerned that the dataset might be "too easy," "too hard," "too small," or "too big" (as each of these was explained in the prompt for part 1 of the project)? Explain your concern. We recognize that you don't have a lot of experience with these issues yet. It's reasonable, though, to open up the dataset, try loading the data up into a Python program, even run your code from recent assignments on the data to try to identify problems that might come up if we use the data. Think of this as "debugging the data": if there are reasons not to use this dataset for the project, we'd rather find out now than later. So work with your team and do the best you can.
6. Would you like to use this dataset later in the course? (We will use some datasets for the final project, and some for future assignments; some won't be used.) This is a yes-or-no question. It is fine if you say "yes" to all three datasets you review, and it is fine if you say "no" to all three datasets you review (i.e., we don't want you to think about this as a choice among the datasets we assigned to you).
7. Does the dataset follow the format required from the instructions and as described in its accompanying document? Please see the compliant example that we posted on Canvas.

Please submit a separate pdf for with your answer for each dataset. The name of the pdf file should be `NAME.pdf`. Put these files together into a single gzipped tarball and submit through Canvas. We will route the reviews back to the teams who created the dataset—please be constructive and professional. Do not include any information about your team or its members in your reports; the reviews are intended to be “blind.” Remember that your grade depends on the reviews you *write*, not the reviews others write about your dataset!