CS 4803/8803: Big Data and Society - Misuse, Abuse, and Algorithms
Spring 2019

Final Project

Due: April 23, 2019@4:30pm ET

In this assignment, you can work independently or in teams of (no more than) 3 students. Each student should upload the same project files (but indicate which student(s) they collaborated with in the upload).

Step 1 (In-Class): You may select any dataset from the machine learning repository - https://archive.ics.uci.edu/ml/datasets.html or from Kaggle - https://www.kaggle.com/datasets based on the following characteristics [*Note: on Kaggle – many of the datasets provide links to the original dataset such that you do not have to set up a new Kaggle profile. Kaggle was acquired by Google in 2017*]:

1. Must have at least a sample size of 10,000 observations
2. Must have at least three variables belonging to a legally recognized protected class
3. Must have at least two dependent variables (outcome variables) that depends on inputs from the protected class variables (*Note: Use your subjective opinion based on the discussions we've had in class*)
4. Must belong to a regulated domain in law (*Note: Feel free to explore regulations established by either federal or state laws*)

Answer the following questions in the final project report:

- Which dataset did you select?
- How many observations are in the dataset?
- Which regulated domain in law does the dataset belong to?
- How many variables in the dataset?
- Which variables did you select as your dependent variables?
- How many and which variables in the dataset are associated with a legally recognized protected class? Which protected classes?

Step 2:

1) Identify the protected class variables and discretize non-numerical values associated with the protected class variables into discrete categories/numerical values (e.g. Male = 0; Female = 1; Other = 2).
2) Select your dependent variables and discretize non-numerical values associated with your dependent variables into discrete categories/numerical values
3) Compute the frequency of any discretized values for each protected class variable from Step 2.1
4) Compute the frequency of any discretized values for each of the protected class variables (Step 2.1) as a function of any discretized dependent variables (Step 2.3)
5) Create histogram(s) comparing the frequency values of the protected class variables as a function of the dependent variables (i.e. based on data from Step 2.4)

Note: Only perform Step 2 for variables having non-numerical values

Provide the following in the final project report:

- Table documenting the relationship between non-numerical values and the discrete numerical values associated with the protected class variables (from Step 2.1)

- Table documenting the relationship between non-numerical values and the discrete numerical values associated with your dependent variables (from Step 2.2)
- Table providing the computed frequency values for 1) the discretized values for each protected class variable (from Step 2.3), and 2) the discretized values for each protected class variable as a function of any discretized dependent variables (from Step 2.4)
- Histograms derived from Step 2.5

Step 3: For the next set of questions, you are allowed to modify code found from the AI Fairness 360 Open Source Toolkit to work with your dataset (https://aif360.mybluemix.net/). I recommend using the Credit Scoring example as a code template:
https://nbviewer.jupyter.org/github/IBM/AIF360/blob/master/examples/tutorial_credit_scoring.ipynb

1) Write import statements to import the metrics you will use to check for bias and the bias mitigation algorithms you will use to mitigate bias. You may use any metrics or bias mitigation algorithms found in the toolkit (additional details below).
2) Load your initial dataset, select one of your protected class variables (while dropping the other protected class variables), identify your privileged/unprivileged values, and split your original dataset into training and testing datasets.
3) Select three fairness metrics from the BinaryLabelDatasetMetric class and compute the fairness metrics on the original training dataset (https://aif360.readthedocs.io/en/latest/modules/metrics.html#binary-label-dataset-metric)
4) The toolkit implements several pre-processing mitigation algorithms. Apply two different pre-processing mitigation algorithms to transform the original dataset (https://aif360.readthedocs.io/en/latest/modules/preprocessing.html)
5) Use the three fairness metrics identified in 3.3 and compute the fairness metrics on the transformed training dataset using each of the two pre-processing mitigation approaches.
6) Repeat Steps 1-5 using one of your other protected class variables

Provide the following in the final project report:

- Provide the resulting code as an attachment to the assignment
- Provide a table documenting the two protected class variables selected, the privileged/unprivileged values, the pre-processing mitigation functions selected, and the resulting values from the fairness metrics computed in Step 3.3 and Step 3.5

Step 4: There are two options for Step 4 – Choose one to complete for the final project.

Option A: For the next set of questions, you are allowed to modify code found from the AI Fairness 360 Open Source Toolkit to work with your dataset (https://github.com/IBM/AIF360/tree/master/examples). For example, code for training a classifier based on the Credit Scoring example can be found here:
https://github.com/IBM/AIF360/blob/master/examples/demo_reweighing_preproc.ipynb

1) Train two classifiers using one of your original datasets and one of your transformed datasets; select one of your dependent variables as the output label
2) Compute the classification metrics for both the original data set and the transformed data set: Balanced accuracy, Disparate impact, Equal opportunity difference, Average odds difference and the Theil index.
3) For each classification metric, identify whether the bias mitigation transformation had a positive change, negative change, or no change on that metric. [*Note: Use your subjective opinion*]

Provide the following in the final project report:

- Provide the resulting code as an attachment to the assignment
- Provide a table documenting the protected class variable associated with the dataset, the dependent variable, the resulting classification metrics for the original and transformed dataset, and whether there was positive, negative, or no change on the metrics.

Option B: For the next set of questions, you are to design your own bias mitigation algorithm

- Design your own bias mitigation algorithm (must not already be represented in the aif360.algorithms.preprocessing class) to transform your original dataset [*Note: Provide sufficient comments in your code so that the algorithm/math can be deciphered*]
- Use the three fairness metrics identified in Step 3.3 and compute the fairness metrics on the transformed training dataset using your mitigation approach
- Discuss whether your bias mitigation algorithm provides a better or worse outcome when compared with the bias mitigation algorithms used in Step 3.4

Provide the following in the final project report:

- Provide the resulting code as an attachment to the assignment
- Provide a table documenting the resulting values from the fairness metrics and why your algorithm provides a better or worse outcome