Fairness in Learning via Regularization: A Project

November 29, 2017

1 General

The project will focus on testing and further validating different methods of regularization aimed for learning fair classifiers. To that extent, it will incorporate acquiring and pre-processing relevant datasets, implementing the suggested algorithms as presented in the relevant paper (”Learning Fair Classifiers: A Regularization-Inspired Approach”), in addition to extending the suggested methods to other types of regularizers, in particular convex ones. Finally, producing an efficient and meaningful visualization of the observed results which will demonstrate the ability of the suggested methods to prevent discrimination in real life learning tasks without major loss of performance.

2 Datasets

The first section will consist of acquiring, pre-processing, and handling the following datasets:

1. The Adult dataset - http://archive.ics.uci.edu/ml/datasets/Adult

   Data from the UC Irvine Repository, contains 1994 Census data. Goal: Predict whether the income of an individual in the dataset is more than 50K per year or not. Protected attribute: Gender.

2. The COMPAS dataset - https://github.com/propublica/compas-analysis

   Data from Broward County, Florida, 2014, originally compiled by ProPublica. Goal: Predict whether a convicted individual would commit a violent crime in the following two years or not. Protected attribute: Race.

3. The Default dataset - https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

   Data from Taiwanese credit card users. Goal: Predict whether an individual will default on payments. Protected attribute: Gender.

4. The UCLA Law School dataset - http://www2.law.ucla.edu/sander/Systemic/Data.htm

   Records of law students who went on to take the bar exam. Goal: Predict whether a student will pass the exam based on features such as LSAT score and undergraduate GPA. Protected attribute: Gender.

5. Synthetic data - In addition to the real life datasets above, we will also test our methods on synthetically generated data, to simulate basic scenarios.

3 Implementation

1. The project will be implemented in Python, utilizing the CVXPY package. http://www.cvxpy.org/en/latest

2. We will use train and use a logistic regressor over the specified datasets.

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & -ll(\theta; S) \\ & + C_1 R_{FPR}(\theta; S^{neg}) \\ & + C_2 R_{FNR}(\theta; S^{pos}) \\ & + \frac{1}{2} C_3 \left|\left|\theta\right|\right|_2^2 \end{aligned}$$

3. Regularization will be done using:

   (a) Absolute value of the difference between the average score (according to the logistic function) among protected groups of the same label (as proposed and implemented in the paper). **(Non-convex)**.

$$R_{FPR}(\theta; S^{neg}) = \left| \frac{\sum\limits_{i \in N_A^{neg}} h_\theta(x_i)}{\left|N_A^{neg}\right|} - \frac{\sum\limits_{i \in N_B^{neg}} h_\theta(x_i)}{\left|N_B^{neg}\right|} \right|$$

$$R_{FNR}(\theta; S^{pos}) = \left| - \frac{\sum\limits_{i \in N_A^{pos}} h_\theta(x_i)}{\left|N_A^{pos}\right|} + \frac{\sum\limits_{i \in N_B^{pos}} h_\theta(x_i)}{\left|N_B^{pos}\right|} \right|$$

   (b) Absolute value of the difference between the average score (according to the linear function) among protected groups of the same label. **(Convex, Non-differentiable at 0)**.

$$R_{FPR}(\theta; S^{neg}) = \left| \frac{\sum\limits_{i \in N_A^{neg}} \theta^T x_i}{\left|N_A^{neg}\right|} - \frac{\sum\limits_{i \in N_B^{neg}} \theta^T x_i}{\left|N_B^{neg}\right|} \right| = \left| \theta^T \underbrace{\left(\frac{\sum\limits_{i \in N_A^{neg}} x_i}{\left|N_A^{neg}\right|} - \frac{\sum\limits_{i \in N_B^{neg}} x_i}{\left|N_B^{neg}\right|} \right)}_{\overline{x}_{neg}} \right| = \left|\theta^T \overline{x}_{neg}\right|$$

$$R_{FNR}(\theta; S^{pos}) = \left| - \frac{\sum\limits_{i \in N_A^{pos}} \theta^T x_i}{\left|N_A^{pos}\right|} + \frac{\sum\limits_{i \in N_B^{pos}} \theta^T x_i}{\left|N_B^{pos}\right|} \right| = \left| \theta^T \underbrace{\left(- \frac{\sum\limits_{i \in N_A^{pos}} x_i}{\left|N_A^{pos}\right|} + \frac{\sum\limits_{i \in N_B^{pos}} x_i}{\left|N_B^{pos}\right|} \right)}_{\overline{x}_{pos}} \right| = \left|\theta^T \overline{x}_{pos}\right|$$

(c) The squared difference between the average score (according to the linear function) among protected groups of the same label. **(Convex, Differentiable)**.

$$R_{FPR}(\theta; S^{neg}) = \left(\theta^T \overline{x}_{neg}\right)^2$$

$$R_{FNR}(\theta; S^{pos}) = \left(\theta^T \overline{x}_{pos}\right)^2$$

4 Experiments

Observation We need to observe that as the true problem of learning a fair classifier is hard even under the relaxed setting of a convex loss function, we consider proxies for the equalized odds constraint. Some of these proxies are convex, which makes them easy to optimize. However, it is important to stress that the quality of the found solutions for the proxy problems with regards to the original problem is a function of the ability of the proxy constraints to emulate the true constraints - which is, of course, data-dependent. For example, considering the proxy suggested in (b) would make sense under the assumption that a solution which provides us with equal distances from the decision boundary among the different protected groups of the same label, would also be an equalized-odds classifier, which is of course not the general case.

Cross validation In order to solve the proxy regularized problem (in the convex cases), we will need to perform cross validation in order to detect the optimal value of C_3 the weight put on the 'usual' L_2 regularization given the weight for the FP/FN regularizers.

Results While we expect to view a clear pareto-optimal trade-off front between accuracy and (proxy) fairness in the convex cases, when we turn to the real problem, we would expect that this will not be the case. Still, we hope to shed light on the inherent trade-offs and prices that are dictated by the model and the distribution, and focus on showing what is possible.

Trade-offs In order to better understand the fairness-accuracy trade-offs under the equalized odds notion, we would like to understand the following, for each of the aforementioned regularization methods:

1. We would like to test three major cases:

   (a) FP regularization only. ($C_1 > 0$ has a varying value, $C_2 = 0$, C_3 to be validated using cross validation).

   (b) FN regularization only. ($C_2 > 0$ has a varying value, $C_1 = 0$, C_3 to be validated using cross validation).

   (c) Both FP and FN regularization, tuned to the same significance ($C_1 = C_2$ have a varying value, C_3 to be validated using cross validation).

2. We would like to produce, for each of these 3 settings, and for each dataset, a visualization demonstrating the highest accuracy possible for a given amount of discrimination (FP/FN differences).

5 Visualization

The main challenge in this section is to produce a clear, easy to understand visualization of the results, demonstrating what is possible for each dataset under each of the notions, and at what price.