Fairness in Classification

March 2020

1 Introduction

Algorithmic Decision making might lack fairness in its results. For e.g. Pretrial Risk Assessment , Mortgage Apporvals, NYPD Stopquestion-and-frisk program, content recommendations, etc. We need systems not to discriminate based on senstive attributes (race, gender, etc.)

There are various notion of fairness.

- Disparate Treatment : A decision making process suffers from Disparate Treatement if the decisions are based on subjects' senstive attributes. (REF)
- Disparate Impact : A decision making process suffers from Disparate Impact if the outcomes disproportionately hurt (or benifit) people with certain senstive attributes. (REF) Even if we remove senstive attributes from our decision making process, still there might be correlation between sensitive attributes and class labels, which will cause disparate impact. E.g. If in a dataset, gender is correlated whether you get the job or not, then % males getting job and % females getting job will be different.
- Disparate Mistreatment : A decision making process to be suffering from disparate mistreatment with respect to a given sensitive attribute (e.g., race) if the misclassification rates differ for groups of people having different values of that sensitive attribute (REF)

2 Paper

- Title : Fairness Constraints: Mechanisms for Fair Classification
- Goal : To design a fair classifier covering two scenarios: 1) Maximizing accuracy with given fairness constraints 2) Maximizing fairness given accuracy constraints (business necessicity). Also, to generalize over any convex classifiers, dataset having multiple sensitive attributes, and each sensitive attributes might have multiple values.
- Setting of the paper : We will not consider senstive attributes while training our classifers, hence we won't have Disparate Treatment. We consider the scenario where we know that the training data already has bias against certain attributes, in that case balancing the results over those attributes (apply p% rule) will migitate Disparate Impact. But directly incorporating p % rule in convex-margin based classsifier will result into a non convex optimization problem.
- p% rule : If the ratio between the percentage of users with a particular sensitive attribute value having $d_\theta(x) \geq 0$ and the percentage of users without that value having $d_\theta(x) \geq 0$ is no less than (p:100) i.e. 80 % rule means the ratio is atleast 80:100
- Paper Approach :
  - Formulating measure of decision boundary (un)fairness as decision boundary covariance. i.e. measuring the unfairness by finding covariance between user's sensitive attributes z and signed distance from user's feature vectors (x) to decision boundary $d_\theta(x)$

    $$Cov(z, d_\theta(x)) \; = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i)$$

    If a decision boundary satisfies the 100 % rule then the covariance will be approximately zero for large dataset

- Maximizing Accuracy Under Fairness Constraints : Design a classifier that maximizes accuracy to fairness constraints (i.e. a specific p % rule)

$$
\begin{aligned}
&minimize\ \ L(\theta) \\
&subject\ to\ \frac{1}{N}\sum_{i=1}^{N}(z-\bar{z})d_\theta(x_i) \leq c \\
&\qquad\qquad \frac{1}{N}\sum_{i=1}^{N}(z-\bar{z})d_\theta(x_i) \geq -c \\
&\qquad where\ \ \text{x is the feature vector without sensitive attributes} \\
&\qquad\qquad\quad \text{z is the sensitve attributes of feature vector} \\
&\qquad\qquad\quad \text{c is covariance threshold, as c goes towards 0, covaranice will go towards 0}
\end{aligned}
$$

- Maximizing Fairness Under Accuracy Constraints : Design a classifiers that maximizes fairness to accuracy constraints i.e. without any fairness constraints, we find the loss of our classifier, and that will be our optimal loss. so we minimize the covaranice i.e. unfairness, subject to accuracy

$$
\begin{aligned}
&minimize\ \ |\frac{1}{N}\sum_{i=1}^{N}(z_i-\bar{z})| \\
&subject\ to\ L_i(\theta) \leq (1+\gamma_i)L_i(\theta^*)\ \ \forall i \in 1,2,...,N \\
&\qquad where\ \ L_i(\theta^*)\ \text{is individual optimal loss of ith data sample found} \\
&\qquad\qquad\quad \gamma_i \geq 0\ \text{is allowed additional loss} \\
&\qquad\qquad\quad \gamma_i = 0\ \text{means that loss should be less than or equal to optimal loss}
\end{aligned}
$$

- Experimental Results:
  - DataSets : Synthetic data by adding attribute that is correlated to class labels, and Real data - Adult income dataset and Bank Marketing dataset. In Synthetic data, there is single sensitive attribute which has binary values, and it is binary Classification In Adult income data, classification of data is based on whether an individual has income above 50K USD or not. It contains two sensitive attributes - gender (having binary values) and race (having multiple values).

    In Bank Marketing dataset, classification is based on whether an individual has subscribed or not. It contains one sensitive attribute - age, which here takes binary values to indicate whether age is between 25 to 60 years.
  - Classifiers : All the above data set are trained over Logistic regression and SVM classifiers
  - Comparision with existing results (Observations):