Learning fair representations (LFR) vs Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment (DM and DM-sen) ¶

Today we are going to explore the methods of unfairness problem in Machine Learning. Since Propublica organization pointed out that the COMPAS (Correctional Offender Management Profiling for Alternative Sanction) system which is a database containing the criminal history are discrminatory against race and gender.It's analysis shows that Black defandants were often predicted to be at a higher risk of recidivism than they actually were while white defandants were predicted to be less risky. Recidivism is defined as defandants reoffend and get arrested again within two years. And the risk scores evaluated by their systemun results in unfairness between the defandant's recidivism situation.

We call a decision-making process to suffer from disparate mistreatment (DM) concerning a given sensitive attribute (e.g., race) if the misclassification rates differ for groups of people with different values of that sensitive attribute (e.g., blacks and whites). For example, in the case of the NYPD Stopquestion- and-frisk program (SQF), where pedestrians are stopped on suspicion of possessing an illegal weapon, having different weapon discovery rates for different races would constitute a case of disparate mistreatment.

User Attributes			Ground Truth (Has Weapon)	Classifier's Decision to Stop		
Sensitive	Non-sensitive					
Gender	Clothing Bulge	Prox. Crime		C_1	C_2	C_3
Male 1	1	1	✓	1	1	1
Male 2	1	0	✓	1	1	0
Male 3	0	1	✗	1	0	1
Female 1	1	1	✓	1	0	1
Female 2	1	0	✗	1	1	1
Female 3	0	0	✓	0	1	0

	Disp. Treat.	Disp. Imp.	Disp. Mist.
C_1	✗	✓	✓
C_2	✓	✗	✓
C_3	✓	✗	✗

Figure 1: Decisions of three fictitious classifiers (C_1, C_2 and C_3) on whether (1) or not (0) to stop a pedestrian on the suspicion of possessing an illegal weapon. Gender is a sensitive attribute, whereas the other two attributes (suspicious bulge in clothing and proximity to a crime scene) are non-sensitive. Ground truth on whether the person is actually in possession of an illegal weapon is also shown.

In [1]:
```
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
%%capture
%run ../lib/LFR_model.ipynb
%run ../lib/DM_DM_sen_Model.ipynb
```

Data Preprocessing: From the ProPublica notebook, we removed the rows that

1. charge date of a defendants Compas scored crime was not within 30 days from when the person was arrested
2. the recidivist flag - is_recid == -1 if we could not find a compas case at all
3. those with a c_charge_degree of 'O' which means ordinary traffic offenses. It will not result in Jail time are removed

4. since we are only intersted in sample fairness between two races: African-American and Caucasian, we subsets our datasets

Here we introduce Learning Fair Representations techiniques to solve unfairness problem, the learning algorithm for fair classification is achieved by formulating fairness as optimization problem of finding good representation. The main idea in this model is to map each individual, represented as a data point in a given input space, to a probability distribution in a new representation space. General speaking, the goal of our model is to learn a good prototype set with the consideration of accuracy and statiscal parity.

We also use Learning Classification without Disparate Mistreatment techniques to solve unfairness problems. These methods avoid disparate treatment and disparate mistreatment simultaneously. In particular, DM is avoided by using fairness constraints, while disparate treatment is avoided by ensuring that sensitive attribute information is not used while making decisions, i.e., by keeping user feature vectors (x) and the sensitive features (z) disjoint.

Reference: Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, Cynthia Dwork, Learning Fair Representations, http://proceedings.mlr.press/v28/zemel13.html (http://proceedings.mlr.press/v28/zemel13.html)

Before we built the LFR model, we first transform each variable to a learnable indicator value. And according to the research paper, defandants with African-American race are regarded as non-sensitive group , Caucasian defandants are regarded as protected group.

As defined in the Learning Fair Representation paper, the Loss function **L = A_z * L_z+ A_x * L_x + A_y * L_y**, where **Ax, Ay, Az** are hyper-parameters governing the trade-off between the system's desire data, And relative **Lz, Lx, Ly** are defined as follows

The second term constrains the mapping to Z to be a good description of X. We quantify the amount of information lost in the new representation using a simple squared-error measure:

$$L_x = \sum_{n=1}^{N} (\mathbf{x}_n - \hat{\mathbf{x}}_n)^2 \tag{8}$$

where $\hat{\mathbf{x}}_n$ are the reconstructions of $\mathbf{x}_n$ from Z:

$$\hat{\mathbf{x}}_n = \sum_{k=1}^{K} M_{n,k} \mathbf{v}_k \tag{9}$$

The second term constrains the mapping to Z to be a good description of X. We quantify the amount of information lost in the new representation using a simple squared-error measure:

$$L_x = \sum_{n=1}^{N} (\mathbf{x}_n - \hat{\mathbf{x}}_n)^2 \tag{8}$$

where $\hat{\mathbf{x}}_n$ are the reconstructions of $\mathbf{x}_n$ from Z:

$$\hat{\mathbf{x}}_n = \sum_{k=1}^{K} M_{n,k} \mathbf{v}_k \tag{9}$$

The final term requires that the prediction of y is as accurate as possible:

$$L_y = \sum_{n=1}^{N} -y_n \log \hat{y}_n - (1 - y_n) \log(1 - \hat{y}_n) \tag{10}$$

Here $\hat{y}_n$ is the prediction for y_n, based on marginalizing over each prototype's prediction for Y, weighted by their respective probabilities $P(Z = k|\mathbf{x}_n)$:

$$\hat{y}_n = \sum_{k=1}^{K} M_{n,k} w_k \tag{11}$$

Therefore we defined the following function to calculate the relative value. And here we use scipy.optimizeo package to minimize our Loss function.

We split the protected group and unprotected group first and concatenate them together. Traning sets and testing sets are split proportionally as 6:1, you can see how does each defandants variables are being rescaled and manipulated

In addition, we implement a method to avoid disparate mistreatment only (DM-sen). The user feature vectors (x) and the sensitive features (z) are not disjoint, that is, z is used as a learnable feature. Therefore, the sensitive attribute information is used for decision-making, resulting in disparate treatment.

$$
\begin{aligned}
\min \quad & L(\theta) \\
\text{s.t.} \quad & \frac{-N_1}{N} \sum_{(x,y)\in\mathcal{D}_0} g_\theta(y,x) + \frac{N_0}{N} \sum_{(x,y)\in\mathcal{D}_1} g_\theta(y,x) \leq c \\
& \frac{-N_1}{N} \sum_{(x,y)\in\mathcal{D}_0} g_\theta(y,x) + \frac{N_0}{N} \sum_{(x,y)\in\mathcal{D}_1} g_\theta(y,x) \geq -c
\end{aligned} \tag{11}
$$

where $\mathcal{D}_0$ and $\mathcal{D}_1$ are the subsets of the training dataset $\mathcal{D}$ taking values $z = 0$ and $z = 1$, respectively. $N_0 = |D_0|$ and $N_1 = |D_1|$.

The results of the LFR model versus a logistic regression are:

In [3]:
```
return_lfr_accuracy()
```

```
the overall test accuracy for LFR is: 56.120000000000005%
the test accuracy for LFR for sensitive: 60.099999999999994%
the test accuracy for LFR for nonsensitive: 53.459999999999994%
the test accuracy for logistic regression is: 67.09%
the test accuracy for logistic regression for sensitive is: 64.56%
the test accuracy for logistic regression for nonsensitive is: 64.5700000
0000001%
```

The results of the DM model are:

In [4]:
```
return_dm_accuracy()
```

```
== Unconstrained (original) classifier ==


Accuracy: 0.642
||  s  || FPR. || FNR. ||
||  0  || 0.26 || 0.44 ||
||  1  || 0.28 || 0.49 ||



== Constraints on FPR ==


Accuracy: 0.642
||  s  || FPR. || FNR. ||
||  0  || 0.26 || 0.44 ||
||  1  || 0.28 || 0.49 ||
```

Explanation and Interpretation of Results

As noted above the accuracy for a basic logistic model is about 67% overall and 64.6% for

sensitive and non-sensitive groups, the accuracy for the LFR model is 47.54% and the accuracy of the DM/DM-sen model was 64.9%. A decrease in accuracy when implementing fair classification is to be expected, particularly in this case where our initial data is an imbalanced dataset. This brings up a couple of problems. There could be sampling bias, which is a situation in which the labeled training data was from a non-representative sample of the population. There might also be label bias which is a situation in which the labeled training data contains samples that were mis-labeled in a way that correlates with being part of a sensitive group.The data can be purposefully unfair and imperfect, making things more fair does not necessarily make things more accurate. There are simply other factors to take into account.

The disparate mistreatment model is intended to bridge the gap created by this imbalance and under-representation of minority groups, and inaccuracy of classification of data that is not necessarily linearly seperable. In terms of accuracy it was about on-par with the logistic regression models of the sensitive and nonsensitive groups, if not a little higher but lover than the logistic regression of the data overall. This makes sense, the DM model should have higher accuracy than the LFR model since it accounts for disparate mistreatment but lower accuracy than the baseline model due to possible bias.

In [5]:
```
1 !jupyter nbconvert --to html main.ipynb
```

```
[NbConvertApp] Converting notebook main.ipynb to html
[NbConvertApp] Writing 587156 bytes to main.html
```

In [ ]:
```
1
```

In [ ]:
```
1
```