]Rishabh Samra
ee17s300

Assignment 2

Q1.(i) FairML

Statistical Classification Criteria

What makes classifier good for classification is a score of accuracy P(y = $\hat{y}$) Score function converts classification by solving as regression problem which helps to summarize data in single valued real score which is a scalar value depending on features of data.Example is of Body Mass Index.
A natural score function is expectation of target variable y conditional on x or R = E[y|X=x] which gives best guess for target variables given observation we have

Sensitive Charecteristics

In many classification tasks feature X encode sensitive charecteristics of an individual. Many neutral features give higher order predictions of sensitive charecteristics.The choice of sensitive attributes helps us decide which group to highlight and what conclusion we draw from investigation.

No Fairness through Unawareness

Removing sensitive attributes ensures impartiality of a classifier.Several features that are slightly predictive of sensitive attribute used to build high accuracy classifiers.In large feature spaces sensitive attributes are generally redundant given the other features. If a classifier trained on the original data uses the sensitive attribute and we remove the attribute, the classifier will then find a redundant encoding in terms of the other features. This results in an essentially equivalent classifier

Non-Discrimination Criteria

Most of the proposed fairness criteria are properties of the joint distribution of the sensitive attribute A, the target variable Y, and the classifier or score R.

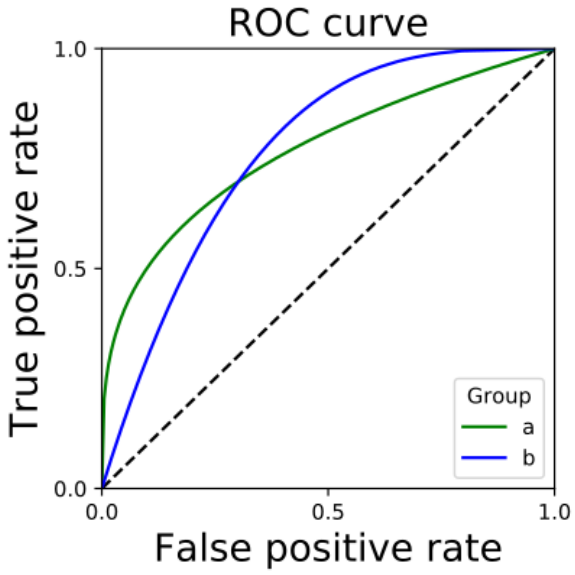
INDEPENDENCE

The random variables(A,R) satisfies independence if A is independent of R (R $\perp$A).
Hence P(R=1|A=a) = P(R=1|A=b) for all groups a and b.If R=1 is acceptance than it implies rate of cceptance to be same for all groups.

Limitations:
Decisions based on classifier can have undesirable properties.Imagine a company in group A hires diligent selected applications while other group B hires carelessly selected applicants at same rate p but company in group A will have more profit.

SEPARATION

Random variables (R,A,Y) satisfy separation if R $\perp$ A|Y.
So P(R=1|Y=1,A=a) = P(R=1|Y=1,A=b) and P(R=1|Y=0,A=a) = P(R=1|Y=0,A=b) Sensitive charecteristics may be correlated with target variable. The separation criterion allows correlation between the score and the sensitive attribute to the extent that it is justified by target variable.
P(R = 1|Y = 1) is called the true positive rate of the classifier,the rate at which the classifier correctly recognizes positive instances. The false positive rate P(R = 1 | Y = 0) highlights the rate at which the classifier mistakenly assigns positive outcomes to negative instances. What separation therefore requires is that all groups experience the same false negative rate and the same false positive rate.
Achieving Separation We can achieve separation by post-processing a given score function without the need for retraining. A binary classifier that ensures separation must achieve same true positive and false positive rate in all groups. Figure highlights ROC Curve of two groups.



Since two groups have different curves indicates that not all trade-off between true and false positive rate are achievable.Tradeoff that are achievable in both groups are those that lie under both curves.Points that are not on curve but under the curve requires randomization.

SUFFICIENCY

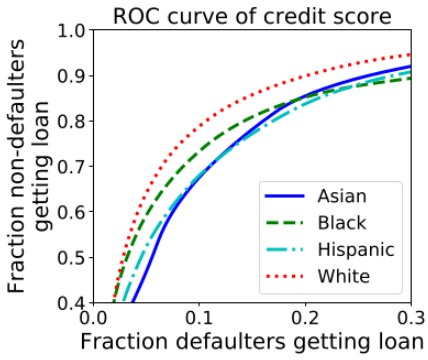
We say Random Variables (R,A,Y) satisfy separation if R ⊥ A| Y. So we have P (Y = 1 | R = r, A = a) = P (Y = 1 | R = r, A = b) Score already subsumes sensitive charecteristic for purpose of predicting target.
Calibration we say that a score R is calibrated if for all score values r in the support of R, we have P (Y = 1 | R = r) = r.This condition means that the set of all instances assigned a score value r has an r fraction of positive instances among them.The fact is calibration in group implies sufficiency.

How to Satisfy Fairness Criteria

1.**Pre-Processing**: Adjust feature space to be uncorrelated with sensitive attribute.It ensures independence after pre-processing.
2.**At Training Time**: Work the constraint into optimization process that construct classifier from training data.Achieving independence at training time can lead to highest utility since we optimize classifier with this criterion in mind.Disadvantage is we give up a fair bit of generality as this approach applies to specific model class or optimization problem.
3.**Post Processing**: It refers to process of taking a trained classifier and adjusting it depending on sensitive attribute such that it is uncorrelated with sensitive attribute.Derived classifier is a randomized function of given score, R and sensitive attribute, A.

Performance Variables and ROC curves

ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Figure shows ROC Curve of credit score of different groups. A particular trade-off



between true positive rate and false positive rate achieved at a threshold t in one group could require a different threshold t' in the other group.

Comparision of Different Criteria

- Maximum profit: Pick possibly group-dependent score thresholds in a way that maximizes profit.
- Single threshold: Pick a single uniform score threshold for all groups in a way that maximizes profit
- Separation: Achieve an equal true/false positive rate in all groups.Subject to this constraint, maximize profit.
- Independence: Achieve an equal acceptance rate in all groups. Subject to this constraint, maximize profit.

Q.1.(ii)Causal Bayesian Viewpoint on Fairness

Consider a dataset $\Delta = (a^n, x^n, y^n)$ corresponding to N individuals. where a^n is sensitive attribute using observations x^n to predict $\hat{y}^n$ of outcome y^n. Unfairness in Δ is the presence of an unfair causal path from A to X or Y.

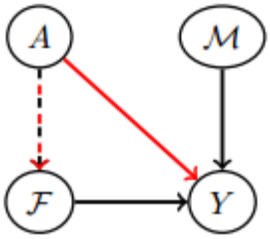


Figure 1. CBN for COMPAS dataset

In the figure 1, F denotes the number of prior arrest of the person which encodes the racial information of the person(A), Y encodes the information about the person if he will re-offend after being granted bail. Unfair path is denoted by link A-Y in which a person is judged based on his race. Another possible unfair path can be A-F-Y as it's not correct to predict if he will commit crime only based on prior arrest as may be the person has changed his nature.

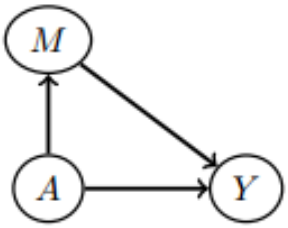
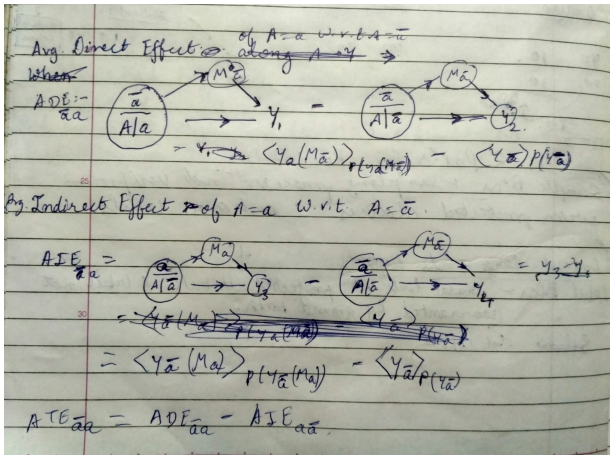


Figure 2. Average Direct and Indirect Effect

Let $Y_a(M_{\bar{a}})$ be the random variable with distribution equal to the conditonal distribution of Y given A restricted to path with A = a along A-Y and A = $\bar{a}$ along A-M-Y.

Average Direct Effect of A=a w.r.t A = $\bar{a}$ is
$ADE_{\bar{a}a} =< Y_a(M_{\bar{a}}) >_{p(Y_a(M_{\bar{a}}))} - < Y_{\bar{a}} >_{p(Y_{\bar{a}})}$
Average Indirect Effect of A=a w.r.t A = $\bar{a}$ is
$AIE_{\bar{a}a} =< Y_{\bar{a}}(M_a) >_{p(Y_{\bar{a}}(M_a))} - < Y_{\bar{a}} >$
Average Total Effect,
$(ATE_{\bar{a}a}) =< Y_a >_{p(Y_a)} - < Y_{\bar{a}} >_{p(Y_{\bar{a}})}$



(Q.2)

To estimate the effect along a specific causal path, we can perform intervention at A=a along group of interest and A = $\bar{a}$ along other paths.

In figure 1, if we consider path A-F hence path A-F-Y to be unfair, unfairness in overall population can be quantified using average total effect calculated as above, that is $ATE_{\bar{a}a}) =< Y_a >_{p(Y_a)} - < Y_{\bar{a}} >_{p(Y_{\bar{a}})}$ If the path A-F-Y is considered to be fair, unfairness can be quantified with the path specific effect along the direct path A-Y given by:
$PSE_{\bar{a}a} =< Y_a(F_{\bar{a}}) >_{p(Y_a(F_{\bar{a}}))} - < Y_{\bar{a}} >_{p(Y_{\bar{a}})}$

Q.1.(iii) Selective Labels Problem

In many domain, data is selectively labelled and observed outcomes are consequence of existing choice of human decision makers.

Problem formulation

Consider a judge granting bail to a criminal. Here Data is selectively labelled because if someone is denied bail we did not have any data for him.If criminal is granted bail and if he returns without committing crime, label is positive else its negative.

REASON FOR EXISTENCE OF BIAS:

- Presence of selective labels as judgement of decision-makers determines which instances are labelled in the data.
- There exists unobservables which are available to the decision-makers when making judgement but are not recorded in the data and hence can't be used by predictive models.
- Results may vary under same conditions for Multiple Judges because of varying acceptance rate.

Let x_i denote the feature values of subject i which are recorded in the data and is associated with unobservable z_i.Human decision maker (j_i) who makes a yes($t_i = 1$) or no has access to both x_i and z_i.Let $y_i \in$ (0,1,NA) denotes the resulting outcome.Problem occurs because observation of outcome y_i is constrained based on decision made by judge (j_i)

INPUT DATA

A dataset D = (x_i, j_i, t_i, y_i) consisting of N observations, each of which corresponds to a subject (individual) from an observational study

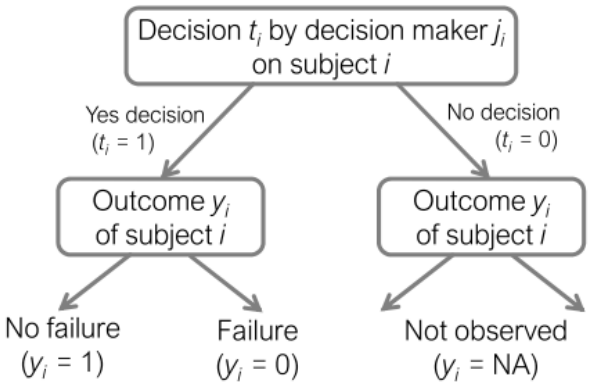


Figure 3. Selective Labels Problem

BLACK BOX PREDICTIVE MODEL

Black box predictive model B assigns risk scores to each observations in D. Scores $\in$ [0,1] indicates how confident the model is in assigning observations to t=0 (denying bail). Given observational data D and predictive model,B goal is to evaluate performance of B and benchmark it against human decisions in D, given the selective labelling of the data and the presence of unobservables.

Acceptance and Rejection Rate:

Acceptance Rate: Ratio of number of subject assigned to yes to total number of subjects.
Failure Rate: Ratio of number of crimes due to decision to total number of subjects.
For instance judge gives decision for 100 people of them 70

were released and 20 commits crime.Failure Rate = 0.2 and Acceptance Rate = 0.7

Contraction Technique:

Comparing black box predictive model to any human decision maker by forcing acceptance rate of model to be same as that of judge and measuring failure rate.If model exhibits lower failure rate then judge so model is judge better.Also imputation and other counterfactual inference techniques can't be used due to presence of unobservables.
Consider bail setting where each judge decides on bail petitions of 100 defendants.

- Compare performance of black box model with judge j' who releases 70% of defendants who appear before him.
- To compare model with j' we run black box model on set of defendants judged by most lenient judge q who releases 90 % defendants(lesser unobserved variables).
- constraining the black box model to detain the same 10 defendants who were detained by q thus avoiding the missing labels.
- In addition it allows black box model to detain another 20 defendants deemed as high risk by model.
- Then compute failure rate on remaining 70 defendants who are released by the model.

The idea behind the contraction technique is to simulate the black box model on the sample of observations judged by the decision-maker q with the highest acceptance rate by contracting the set of observations assigned to yes decision by q while leveraging the risk scores (or probabilities) assigned to these observations by the model.
Algorithm is as follows:

Experimental Evaluation

Accuracy of the estimates of model failure rates obtained using contraction technique was analyzed by simulating the selective labels problem using synthetic data.They also compare the effectiveness of the contraction technique to various state-of-the-art baselines commonly used for counterfactual inference and imputation.

Figure shows the effect of selective labels on estimation of predictive model failure rate (error bars denote standard errors)Green curve is the true failure rate of the predictive model and the machine evaluation using contraction (blue curve) follows it very closely. However, various imputation techniques heavily underestimate the failure rate. Based on the estimates of imputation, one would conclude that

Input: Observational data $\mathcal{D}$, Probability scores $\mathcal{S}$, Acceptance rate r
Procedure:
Let q be the decision-maker with highest acceptance rate in $\mathcal{D}$

$\mathcal{D}_q = \{(x, j, t, y) \in \mathcal{D} | j = q\}$
▷ $\mathcal{D}_q$ is the set of all observations judged by q

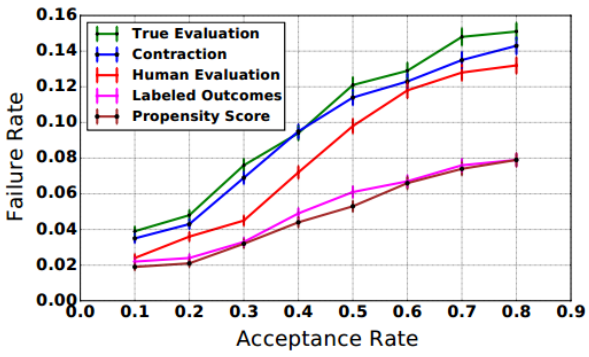
$\mathcal{R}_q = \{(x, j, t, y) \in \mathcal{D}_q | t = 1\}$
▷ $\mathcal{R}_q$ is the set of observations in $\mathcal{D}_q$ with observed outcome labels

Sort observations in $\mathcal{R}_q$ in descending order of confidence scores $\mathcal{S}$ and assign to $\mathcal{R}_q^{sort}$
▷ Observations deemed as high risk by $\mathcal{B}$ are at the top of this list

Remove the top $[(1.0-r)|\mathcal{D}_q|]-[|\mathcal{D}_q|-|\mathcal{R}_q|]$ observations of $\mathcal{R}_q^{sort}$ and call this list $\mathcal{R}_\mathcal{B}$
▷ $\mathcal{R}_\mathcal{B}$ is the list of observations assigned to $t = 1$ by $\mathcal{B}$

Compute $u = \sum_{l=1}^{|\mathcal{R}_\mathcal{B}|} \frac{\mathbb{1}(y_l=0)}{|\mathcal{D}_q|}$
Return u

Figure 4. Contraction technique for estimating failure rate at acceptance rate r



the predictive model outperforms human judges (red curve), while in fact its true performance is worse than that of the human judges.

Construction of Bounds:

There exists some additional data where there is no decision(t=0) and which is not common to both decision maker q and Predictive model, B. The error in calculation of True Failure Rate is due to this data and width of the bound measured from true value is : $\frac{(1-a)|D_q-R_q|}{|D_q|}$. Here a is fraction of observations in the set $D_q - R_q$ that both B and q agrees on assigning to a no decision (t=0). When B favors assigning the additional subjects to t=1 , in worst case all might result in undesirable outcomes. The upper bound is (u + $\frac{(1-a)|D_q-R_q|}{|D_q|}$) and vice versa for lower bound (u - $\frac{(1-a)|D_q-R_q|}{|D_q|}$).

Q.3

We have data from different judges which have different acceptance rate so under the same condition, the decision for the criminal will be different for each judge causing bias. We can take weighted sum of failure rate from each judge to calculate true failure rate. Weights can be obtain by fitting a regression model taking as parameters all the features.
Also unlabelled data can be assigned label according to the score obtained by fitting logistic regression thus avoiding problem due to selective labels.

(Q.4) Fairness in children across various cultures

A key component of the human sense of fairness is inequity aversion which refers to the willingness to sacrifice material payoffs for the sake of greater equality. Disadvantageous Inequity(DI) aversion occurs to avoid receive less than a peer and vice-versa for advantageous inequity(AI).
An experiment was devised to study how fairness develops in children across 7 countries and measure DI and AI among them. In the inequity game, two children sat across from each other at an apparatus, randomly assigned to either disadvantageous (DI condition) or advantageous (AI condition) allocations. One child, the actor, had a choice between accepting the allocation or rejecting it. The recipient played no part in the decision.
It was observed that Disadvantageous Inequity Aversion was common and emerged in middle childhood while Advantageous Inequity Aversion was dominant among children of US, Canada and Uganda and appears after DI.
So culture and society strongly influences inequity aversion and shapes acquisition of fairness behavior during childhood.
Reference: The ontogeny of fairness in seven societies