Mitigating Unwanted Biases with Adversarial Learning

Brian Hu Zhang, Blake Lemoine, Margaret Mitchell

EECS 598 Presentation
04/06/2021
Sahil Farishta

Motivation

- Machine learning models can be extremely susceptible to biases in data
  - Example: Face recognition software often struggles for minorities
  - Predictions should be independent of protected variables
    - Race, gender, zip code, etc.
- We need the model to be fair despite biases in training data
  - Demographic parity
    - Given protected variable Z, the predictors output is independent of the value of Z
  - Equality of odds
    - For all possible values of the true label, the prediction probability is the same for all Z
  - Equality of opportunity
    - For a specific value of the true label, the prediction probability is the same for all Z
  - One or more of the above 3 fairness definitions can be used for a certain task

Background

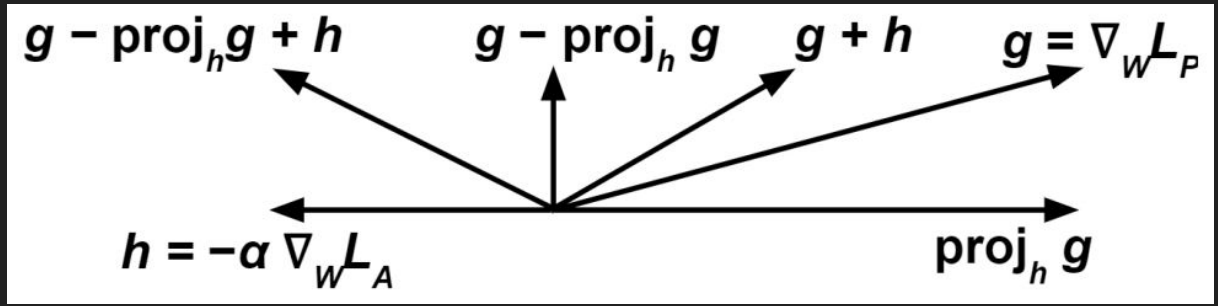
- Previous work has been done in attempting to remove bias in a model
  - Debiasing word embeddings by Bolukbasi et al. [2]
    - Relies on human curated set of gender specific words to remove bias
  - Simple models by Lum and Johndrow [3]
    - Demonstrated that removing protected variable doesn't help since other attributes can be correlated with the protected variable
  - Adversarial training
    - Goodfellow et al. [4] used GANs to compete and attempt to create real life-like images
    - Beutel et al. [5] achieved equality of opportunity in discrete output cases

Threat Model

- Need to train a model f which meets one of the definitions of equality
  - Given input data X, f outputs predicted label $\hat{Y}$
  - Predictor has Loss $L_P(\hat{Y},Y)$ and weights W
  - Z is the protected variable
    - Z may be a variable present in the training data or it may be inferred from other variables
- Adversary is in the form of a discriminator in the GAN setup
  - Has a loss term $L_A(\hat{z}, z)$ and weights U and takes in input from the output layer of the predictor
    - Outputs $\hat{z}$ which is predicted value for protected variable
  - For demographic parity
    - Adversary gets predicted variable $\hat{Y}$ only and succeeds if they predict Z
  - For equality of odds
    - Adversary gets predicted variable $\hat{Y}$ and true label Y and succeeds if they predict Z
  - For equality of opportunity on class y
    - Adversary only is trained on predictions that have a true label Y=y

Adversarial Debiasing

- The adversary has a simple loss update function
  - Updated weights U to minimize $L_A(\hat{z}, z)$
- Predictor takes into account information visible to adversary during update
  - Update weights W using rule $(\nabla_w L_P) - \text{proj}_{\nabla_w L_A}(L_P) - \alpha(\nabla_w L_A)$
  - First term minimizes predictor loss
  - Second term prevents predictor from moving in a direction that decreases adversarial loss
  - Third term attempts to increase adversarial loss



Effect of each term on adversarial loss

Adversarial Debiasing Technique Advantages

- Technique presented provides several advantages over previous work
  - Usable on any of the three fairness definitions in discrete or continuous cases
  - Compatible with any gradient based predictor model
  - Model converges to a definition that matches the desired fairness definition
    - Predictor should perform well as well since update tries to minimize general loss as well

Theoretical Guarantees - Information from $\hat{Y}$

- Need to assume the adversary's loss is convex in U and concave in W
  - U is weights of adversary and W is weights of predictor
  - Not necessarily the case but many machine learning proofs require similar assumptions
- Begin with W_0 where predictor always outputs some $\hat{Y}$ for all input X
  - Effectively ignores input
  - U_0 will minimize L_A in this case
- Allow models to converge to W* and U*
- Based on update function, the predictor will never move in a direction that allows adversary to reduce loss
  - Thus $L_A(W^*,U^*)=L_A(W^*,U_0)$
  - Effectively, the adversary will have the same loss as if all output from the predictor was identical
    - No information gained from $\hat{Y}$

Theoretical Guarantees - Demographic Parity

- Let the training set be triples of the form (X, Y, Z)
  - X is training data, Y is true labels, Z is discrete value of protected variable
  - Adversary has only the prediction $\hat{Y}$
- From before, the trained adversary gains no information about Z from $\hat{Y}$
  - $H(Z)=H(Z|\hat{Y})$ then where H is the entropy function
  - $H(Z|\hat{Y}) = E[-\log P(A(\hat{y})=z|\hat{Y}=\hat{y})]$ where E is the expected value and A is the adversary model
  - This is the cross entropy loss of the adversary
  - So $L_A=H(Z|\hat{Y})=H(Z)$
    - H(Z) is the uncertainty of the value of Z, so the adversary with access to $\hat{Y}$ is as uncertain about Z as if it has no information about anything

Theoretical Guarantees - Equality of Odds and Opportunity

- Let the training set be triples of the form (X, Y, Z)
  - X is training data, Y is true labels, Z is discrete value of protected variable
  - Adversary has only the prediction $\hat{Y}$ and the true label Y
- From before, the trained adversary gains no information about Z from $\hat{Y}$
  - Adversary cannot have loss less than H(Z|Y) where H is the entropy function
    - It is given access to $\hat{Y}$ and Y only and gains no information from $\hat{Y}$
  - $H(Z|\hat{Y},Y) = E[-\log P(A(\hat{y})=z|\hat{Y}=\hat{y},Y=y)]$ where E is the expected value and A is the adversary model
  - This is the cross entropy loss of the adversary
    - But adversary cannot have loss $H(Z|\hat{Y},Y) < H(Z|Y)$
  - So $L_A=H(Z|\hat{Y},Y)=H(Z|Y)$
    - Conditional independence between Z and $\hat{Y}$ given Y which is the desired property

Experiments - Toy Scenario

- Given training sample (x,y,z) where x is input, y is label, z is protected variable
  - Let x consist of u and z (two variables)
    - v ~ N(z,1) and u,y ~ N(v,1)
  - Regression model outputs $y=\sigma(0.7u+0.7z)$
    - Correlated with protected value which is biased
    - Cannot solve by removing z from training data and u and z are correlated
  - Demographic parity model
    - $y=\sigma(0.6u-0.6z+0.6)$
    - u-z is drawn from N(0,2) which isn't dependent on z so the model is independent of z

Experiments - Word Embeddings

- Perform analogy task
  - Man:Woman :: He:?
    - Fill in the question mark with a meaningful word that completes the analogy
- Use data from Google analogy data set
  - Words are provided as embeddings
  - x_1, x_2, x_3 to be the first three words (provided) and y to be fourth word (predicted)
  - Keep track of g which is the gender direction
    - Proj_g y is the protected variable - gender direction of the label y
- Model is represented as $y=v-ww^Tv$
  - $v=x_2+x_3-x_1$ - used commonly for analogy tasks
  - w is the learned vector to help debiasing
    - Expected to be approximately 0 without using the proposed debiasing structure

Gender Direction - Word Embeddings

- Gender direction allows us how to see how gendered a word is
  - Calculated by projection onto gender space
  - Negative and positive are both bad
  - 0 means not-gendered (not-biased)

word	gender_score	word	gender_score
his	-0.660903	husband	0.950914
he	-0.584860	policewoman	0.816518
unimpressive	-0.544877	women	0.758530
Anaheim	-0.503144	mom	0.732802
Libya	-0.486664	princess	0.719736
playing	-0.472259	she	0.706431
play	-0.459254	sisters	0.699692
sharpest	-0.455905	stepson	0.686814
Detroit	-0.454509	her	0.683995
calmly	-0.448576	queen	0.682354

Results - Word Embeddings

- Before gender debiasing, they found w to be very perpendicular to g
  - $w^T \cdot g = 0.08$, $||w|| = 0.82$
- With gender debiasing w points towards g
  - $w^T \cdot g = 0.55$, $||w|| = 0.96$
  - Still capable gendered analogies
    - Man:Woman :: He:She which makes sense
  - Biased analogies go away

biased		debiased	
neighbor	similarity	neighbor	similarity
nurse	1.0121	nurse	0.7056
nanny	0.9035	obstetrician	0.6861
fiancée	0.8700	pediatrician	0.6447
maid	0.8674	dentist	0.6367
fiancé	0.8617	surgeon	0.6303
mother	0.8612	physician	0.6254
fiance	0.8611	cardiologist	0.6088
dentist	0.8569	pharmacist	0.6081
woman	0.8564	hospital	0.5969

Results for He:She :: Doctor:?

Experiments - UCI Adult Dataset

- Use dataset which contains information about adults to determine if income>50k
  - Enforcing equality of odds as fairness measure
  - Protected variable is the gender
  - Cat type features are categorical (discrete)
  - Cont type features are continuous
- Predictor model uses logistic regression
  - σ is sigmoid function
  - $\hat{y}=\sigma(wx+b)$
  - b is a learnable scalar and w is a learnable vector
- Adversary model is similar
  - $s=\sigma((1+|c|)\sigma^{-1}(\hat{y}))$
  - $\hat{z}=u[s,sy,s(1-y)]+b$
  - c is a learnable scalar and u is a learnable vector

Feature	Type	Description
age	Cont	Age of the individual
capital_gain	Cont	Capital gains recorded
capital_loss	Cont	Capital losses recorded
education_num	Cont	Highest education level (numerical form)
fnlwgt	Cont	# of people census takers believe that observation represents
hours_per_week	Cont	Hours worked per week
education	Cat	Highest level of education achieved
income	Cat	Whether individual makes > $50K annually
marital_status	Cat	Marital status
native_country	Cat	Country of origin
occupation	Cat	Occupation
race	Cat	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
relationship	Cat	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
sex	Cat	Female, Male
workclass	Cat	Employer type

UCI Adult Dataset Description

Results - UCI Adult Dataset

- Authors found that the model was able to enforce demographic parity easily
  - Harder to enforce equality of odds
  - Reminder: Update function for predictor is $(\nabla_w L_P)$ - $\text{proj}_{\nabla w} L_A(L_P)$ - $\alpha(\nabla_w L_A)$
    - Authors found that setting α=sqrt(t) where t is the step counter helped
      - Overtime try and hurt the adversary more
- False positive and false negative rates are approximately equal
  - FNR female=0.0647, male=0.0701
  - FPR female=0.4458, male=0.4349
  - Overall accuracy was about the same
    - 86% before and 84.5% after
  - Difference is not statistically significant
    - Equality of odds achieved

Without Debiasing			With Debiasing		
Female	Pred 0	Pred 1	*Female*	Pred 0	Pred 1
True 0	4711	120	True 0	4518	313
True 1	265	325	True 1	263	327
Male	Pred 0	Pred 1	*Male*	Pred 0	Pred 1
True 0	6907	697	True 0	7071	533
True 1	1194	2062	True 1	1416	1840

Confusion Matrix for UCI Adult Dataset

Benchmark Comparison - UCI Adult Dataset

- Results shown by authors beat previous baseline
  - Difference in FPR and FNR in previous work was much higher
    - FPR difference was 0.147 vs 0.0054
    - FNR difference was .0698 vs 0.0109
      - FNR rates were much higher however

		Female		Male	
		Without	With	Without	With
Beutel et al. (2017)	FPR	0.1875	0.0308	0.1200	0.1778
	FNR	0.0651	0.0822	0.1828	0.1520
Current work	FPR	0.0248	0.0647	0.0917	0.0701
	FNR	0.4492	0.4458	0.3667	0.4349

False Positive and Negative Rates vs Benchmark

Reproduced Results

- Reran on analogy set
  - Different results than paper
    - Dataset has changed
    - Similar trends
- He:She::Doctor:?
  - We see that different types of doctors dominated the results, especially in the unbiased dataset
  - Thinks like nurse, midwife, and pharmacist fall significantly

Biased		Unbiased	
Neighbor	**Similarity**	**Neighbor**	**Similarity**
Doctor	0.865327	Doctor	0.810452
Nurse	0.692169	Doctors	0.633919
Gynecologist	0.690843	Gynecologist	0.618335
Nurse Practitioner	0.645798	Physician	0.604603
Doctors	0.643204	Nurse	0.601004
Pediatrician	0.641547	Pediatrician	0.596106
Physician	0.630148	Nurse Practitioner	0.568029
Midwife	0.616777	Dermatologist	0.563209
Pharmacist	0.609216	Midwife	0.551024
Dentist	0.604598	Pharmacist	0.548549

Future Work

- Making sure debiased results are still useful and applicable in tasks
  - Analogies worked well and the accuracy on the UCI Adult set was high
  - Need to see if this holds in more complex task
- Training is touchy with immense dependence on correct hyperparameters
  - UCI Adult set needed specific hyperparameter α to converge
  - Need to improve and guarantee convergence
- Can we use complex adversaries for tasks like image classifications on faces
  - Even combining multiple powerful adversaries
- Do more complex predictors need more complex adversaries to ensure the data is entirely debiased
  - So far simple adversaries work well but particularly in discrete cases

Conclusion

- Method presented is powerful and can generate usable unbiased models
  - Better than the previous state of the art method
  - Requires assumptions on the setup for the predictor and adversary but these are similar to other assumptions needed for various machine learning proofs
  - Models still perform well even after debiasing
- Future work will involve exploring other use cases to test method
  - Find ways to improve convergence guarantees
  - Explore continuous cases further

References

[1] Mitigating Unwanted Biases with Adversarial Learning. Brian Hu Zhang, Blake Lemoine, Margaret Mitchell

[2] Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. Bolukbasi et. al.

[3] A statistical framework for fair predictive algorithms. Kristian Lum, James Johndrow.

[4] Generative Adversarial Nets. Goodfellow et. al.

[5] Data decisions and theoretical implications when adversarially learning fair representations. Alex Beutel, Jilin Chenn, Zhe Zhao, Ed Chi.