Mitigating Unwanted Biases with Adversarial Learning

論文情報

- Brian Hu Zhang, Blake Lemoine, Margaret Mitchell
- Stanford University, Google
- AAAI 2018

なんで選んだの？

公平性配慮データマイニングで使えるツール "AIF360" のなかで実装されているものの元論文であったため.

はじめに

- 機械学習は，学習に使われるデータに強く影響を受けてしまう.
- 機械学習の出力に公平性を考慮する方法として，公平性の指標; Demographic Parity, Equality of Odds, Equality of Opportunityを損失関数のなかに組み込んで，不公平を軽減する方法がある.
- 本論文では，adversarial debiasingの文脈で公平性の指標を使っていく.
- 使っていく公平性の指標は次の通りに定義される.

用語について

- X: データセット(センシティブ属性を含まない．)
- Y: 予測ラベル
- Z: センシティブ属性
- W: 予測の際のパラメータ
- U: aのパラメータ

Definition 1. DEMOGRAPHIC PARITY. A predictor $\hat{Y}$ satisfies *demographic parity* if $\hat{Y}$ and Z are independent.

This means that $P(\hat{Y} = \hat{y})$ is equal for all values of the protected variable Z: $P(\hat{Y} = \hat{y}) = P(\hat{Y} = \hat{y}|Z = z)$.

Definition 2. EQUALITY OF ODDS. A predictor $\hat{Y}$ satisfies *equality of odds* if $\hat{Y}$ and Z are conditionally independent given Y.

This means that, for all possible values of the true label Y, $P(\hat{Y} = \hat{y})$ is the same for all values of the protected variable: $P(\hat{Y} = \hat{y}|Y = y) = P(\hat{Y} = \hat{y}|Z = z, Y = y)$

Definition 3. EQUALITY OF OPPORTUNITY. If the output variable Y is discrete, a predictor $\hat{Y}$ satisfies *equality of opportunity* with respect to a class y if $\hat{Y}$ and Z are independent conditioned on $Y = y$.

This means that, for a *particular* value of the true label Y, $P(\hat{Y} = \hat{y})$ is the same for all values of the protected variable: $P(\hat{Y} = \hat{y}|Y = y) = P(\hat{Y} = \hat{y}|Z = z, Y = y)$

Adversarial debiasing

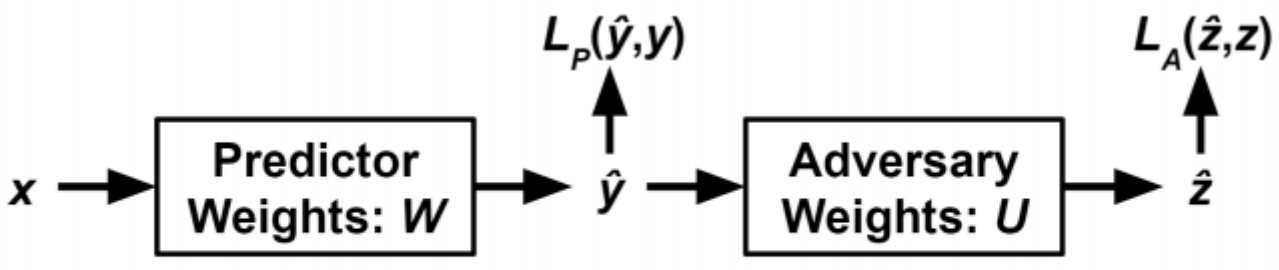

Figure 1: The architecture of the adversarial network.

全体としては，PredictorとAdvesaryの二つで構成要素でなりたっている．Predictでは，与えられたデータXを用いて\hat{Y}を予測する．

Adversaryの部分では，予測器の出力を入力として用いて，センシティブ属性があるかどうかを予測する．（GANでいうdiscriminatorに相当）．パラメータUの更新は損失関数La(Z,\hat{Z})を用いて行なっていく．

使用する公平性の指標によって，Adversaryにはさらに追加する情報がある．

Wを更新するときは，LpとLaの両方を用いて更新していく．

$$\nabla_W L_P - \text{proj}_{\nabla_W L_A} \nabla_W L_P - \alpha \nabla_W L_A \qquad (1)$$

where α is a tuneable hyperparameter that can vary at each time step and we define $\text{proj}_v x = 0$ if $v = 0$.

第一項，第三項はそれぞれpredictorとadversaryの損失関数での勾配．第二項のprojは,下付き文字が0のとき，値が0なると定義．つまり，Laの勾配が0の時は，Lpの勾配だけが残る．それ以外は，Laの勾配だけが残るようになる．(ここの部分が敵対的をさしている？)

理論保証

具体的な話は割愛します．．．

1. LpとLaの凸性と凹性が保証できるよって話.
2. Zかどうかの識別にYを用いると,収束がいいよって話．(評価指標がDPよりもEOのほうが収束がいいよ)
3. Adversaryが強くなりすぎると収束しなくなるよって話(GANでもバランスが取りにくいのと同じ話)

実験

実験では，Word Embedding, UCI Adult Datasetを使っていく．

Word Embedding

analogy taskを取り組むようにしていく．(i.e. man:woman :: he:?)

word embeddingでは，データにあるバイアスに強く影響を受けてしまうことがよく知られている．(Bolukdasi et al.2016)まず"gender direction"について計算する．この計算した値によって今回用いるセンシティブな属性を決定する.

データセットとしては，センシティブ属性（今回は性別）はwikipediaから取得．そのほかはGoogle analogy datasetを使う.

結果は次のようになった.

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

Table 1: Completions for `he : she :: doctor : ?`

上から答えになりそうな順でランキングされている．biasedではmother,womanなど女性に関する答えが含まれている．それに対して，debiasedでは，職業などが多く出ており，性別に関する情報は出てきていない． このとこから，debiaseがきちんと機能していると言える．

Adult dataset

センシティブ属性zとして性別を用いる．debaiasと通常のモデルでそれぞれの属性における金剛行列は次のようになった．

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

Table 3: Confusion matrices on the UCI Adult dataset, with and without equality of odds enforcement.

この表から言えるのは,全体での精度は86%から84.5%にしか下がっていない.

またFPR,FNRは次のようになった.

		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

Table 4: False Positive Rate (FPR) and False Negative Rate (FNR) for income bracket predictions for the two sex subgroups, with and without adversarial debiasing.

この表から言えるのは，FPRでは 0.0647 ≈ 0.0701, FNRでは，0.4458 ≈ 0.4349 で男女で差が小さくなっているので，EOの観点から公平性が達成されていると言える.

まとめ

敵対的学習でデータのバイアスを取り除く手法を提案した．また公平性のツールも後悔してる.

感想

予測したラベルからセンシティブを判別するところ(UやLa(z,\hat{z}))が本当にそれでいいのかよくわからなかった.