Fairness-aware Classifier with Prejudice Remover Regularizer

Toshihiro Kamishima*, Shotaro Akaho*, Hideki Asoh*, and Jun Sakuma†

*National Institute of Advanced Industrial Science and Technology (AIST), Japan

†University of Tsukuba, Japan; and Japan Science and Technology Agency

ECMLPKDD 2012 (22nd ECML and 15th PKDD Conference)

@ Bristol, United Kingdom, Sep. 24-28, 2012

Present by: Xuehan Chen

Date: 03/07/2021

Outline

Applications

- fairness-aware data mining applications

Difficulty in Fairness-aware Data Mining

- Calders-Verwer's discrimination score, red-lining effect

Fairness-aware Classification

- fairness-aware classification, three types of prejudices

Methods

- prejudice remover regularizer, Calders-Verwer's 2-naïve-Bayes

Experiments

- experimental results on Calders&Verwer's data and synthetic data

Related Work

- privacy-preserving data mining, detection of unfair decisions,
- explainability, fairness-aware data publication

My opinion

Overview

Fairness-aware Data Mining

fairness, discrimination, neutrality, or independence

- fairness-aware classification, regression, or clustering
- detection of unfair events from databases
- fairness-aware data publication

Examples of Fairness-aware Data Mining Applications

Elimination of Discrimination

- ex. credit scoring, insurance rating, employment application
- socially sensitive information: gender, religion, race

Information Neutral Recommender System

- Social media make biased recommendations ex. conservative or progressive people

Outline

Applications

- fairness-aware data mining applications

Difficulty in Fairness-aware Data Mining

- Calders-Verwer's discrimination score, red-lining effect

Fairness-aware Classification

- fairness-aware classification, three types of prejudices

Methods

- prejudice remover regularizer, Calders-Verwer's 2-naïve-Bayes

Experiments

- experimental results on Calders&Verwer's data and synthetic data

Related Work

- privacy-preserving data mining, detection of unfair decisions,
- explainability, fairness-aware data publication

Conclusion

Difficulty in Fairness-aware Data Mining

US Census Data : predict whether their income is high or low

	Male	Female
High-Income	3,256	590
Low-income	7,604	4,831

fewer

Females are minority in the high-income class

- # of High-Male data is 5.5 times # of High-Female data
- While 30% of Male data are High income, only 11% of Females are

Red-Lining Effect

Calders-Verwer discrimination score (CV score)

Pr[Y+ | S-] -Pr[Y+ | S+]

Pr[High-income | Male] - Pr[High-income | Female]

- US Census Data samples

  The baseline CV score is 0.19

- Incomes are predicted by a naïve-Bayes classifier trained from data containing all sensitive and non-sensitive features

  The CV score increases to 0.34, indicating unfair treatments

- Even if a feature, gender, is excluded in the training of a classifier

  CV score improved to 0.28, but still being unfairer than its baseline

Red-Lining Effect: Ignoring sensitive features is ineffective against the exclusion of their indirect influence

Outline

Applications

- fairness-aware data mining applications

Difficulty in Fairness-aware Data Mining

- Calders-Verwer's discrimination score, red-lining effect

Fairness-aware Classification

- fairness-aware classification, three types of prejudices

Methods

- prejudice remover regularizer, Calders-Verwer's 2-naïve-Bayes

Experiments

- experimental results on Calders&Verwer's data and synthetic data

Related Work

- privacy-preserving data mining, detection of unfair decisions,
- explainability, fairness-aware data publication

Conclusion

Variables

Y

objective variable
- a binary feature variable {0, 1}
- a result of serious decision

ex., whether or not to allow credit

S

Sensitive variable
a binary feature variable {0, 1}
socially sensitive information
ex., gender or religion

X

non-sensitive variables
- a numerical feature vector
- features other than a sensitive featur
- non-sensitive, but may correlate with

Three types of prejudices

Prejudice : the statistical dependences of an objective variable or non-sensitive features on a sensitive feature

Direct Prejudice

- a clearly unfair state that a prediction model directly depends on S
- -> $Y \not\perp\!\!\!\perp S \mid X$

Indirect Prejudice

- statistical dependence of Y on S even lack of direct S
- Red-lining effect
- $Y \not\perp\!\!\!\perp S \mid X$
- but not $Y \perp\!\!\!\perp S$

Latent Prejudice

- statistical dependence of X on S
- completely excluding sensitive information
- $Y \perp\!\!\!\perp S \mid X$, $Y \perp\!\!\!\perp S$
- $X \not\perp\!\!\!\perp S$
- ->$X \perp\!\!\!\perp Y \mid S$

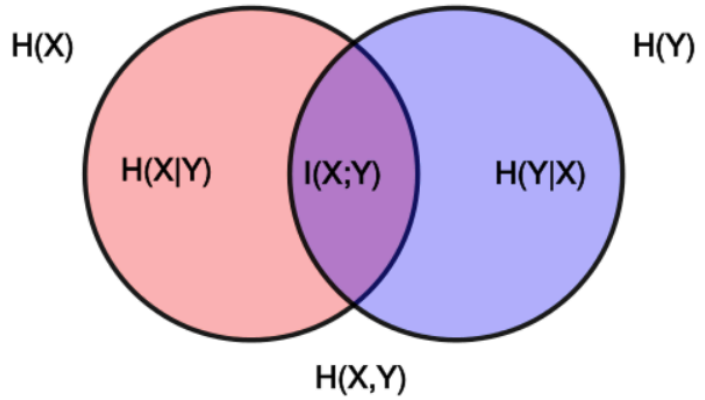
Prejudice Index (Indirect Prejudice)

Prejudice Index (PI):

$$\mathrm{PI} = \sum_{(y,s)\in\mathcal{D}} \hat{\Pr}[y,s] \ln \frac{\hat{\Pr}[y,s]}{\hat{\Pr}[y]\hat{\Pr}[s]}.$$

Normalized Prejudice Index (NPI):

$$\mathrm{NPI} = \mathrm{PI}/(\sqrt{\mathrm{H}(Y)\mathrm{H}(S)}),$$



Information Entropy

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log \mathrm{P}(x_i)$$

Mutual Information

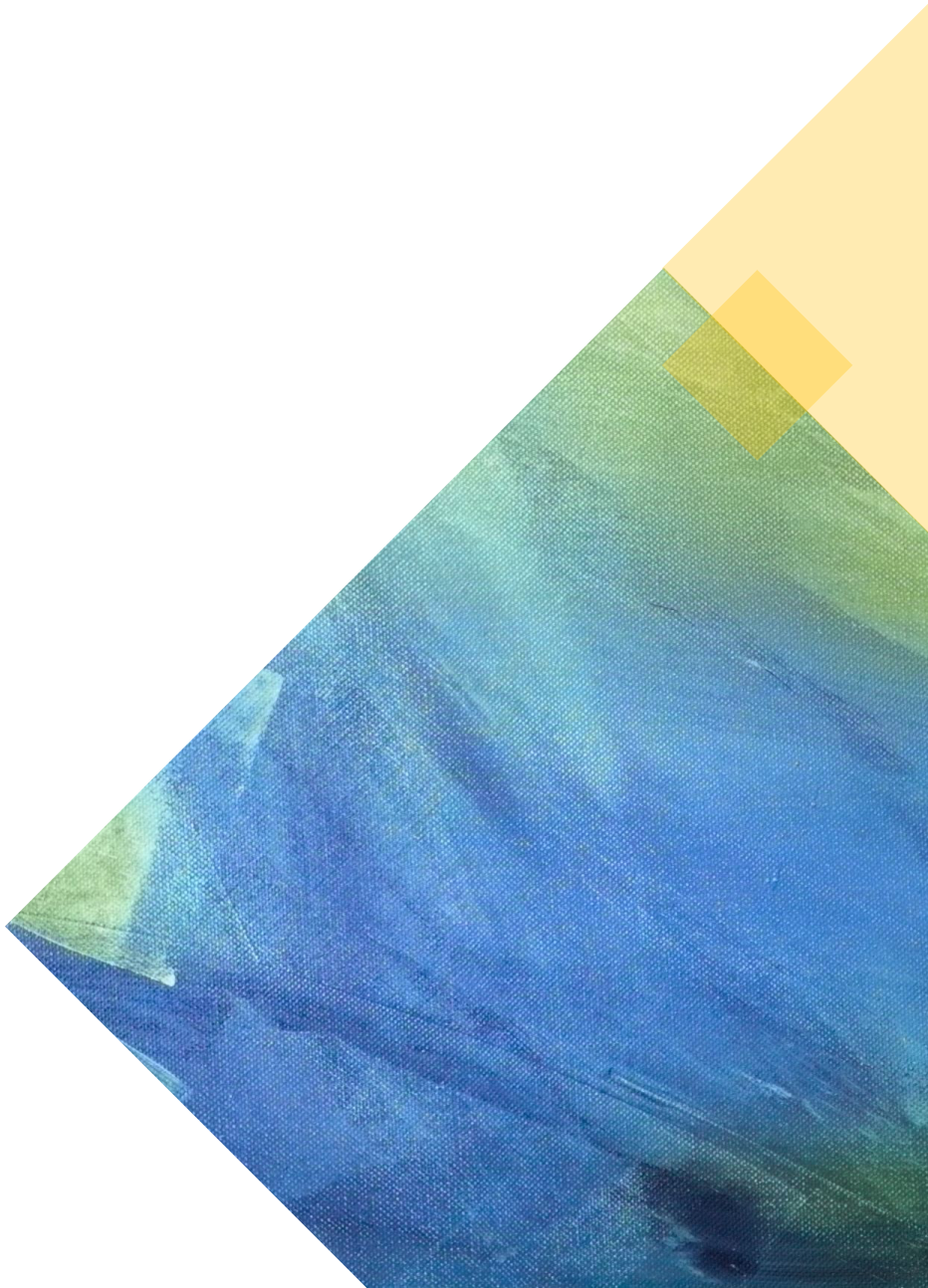
$$\begin{aligned} \mathrm{I}(X;Y) &\equiv \mathrm{H}(X) - \mathrm{H}(X \mid Y) \\ &\equiv \mathrm{H}(Y) - \mathrm{H}(Y \mid X) \\ &\equiv \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \\ &\equiv \mathrm{H}(X,Y) - \mathrm{H}(X \mid Y) - \mathrm{H}(Y \mid X) \end{aligned}$$

Underestimation

A classifer has not yet converged

Negative legacy

Unfair labeling in the training data



Outline

Applications

- fairness-aware data mining applications

Difficulty in Fairness-aware Data Mining

- Calders-Verwer's discrimination score, red-lining effect

Fairness-aware Classification

- fairness-aware classification, three types of prejudices

Methods

- prejudice remover regularizer, Calders-Verwer's 2-naïve-Bayes

Experiments

- experimental results on Calders&Verwer's data and synthetic data

Related Work

- privacy-preserving data mining, detection of unfair decisions,
- explainability, fairness-aware data publication

Conclusion

Logistic Regression
with Prejudice Remover Regularizer

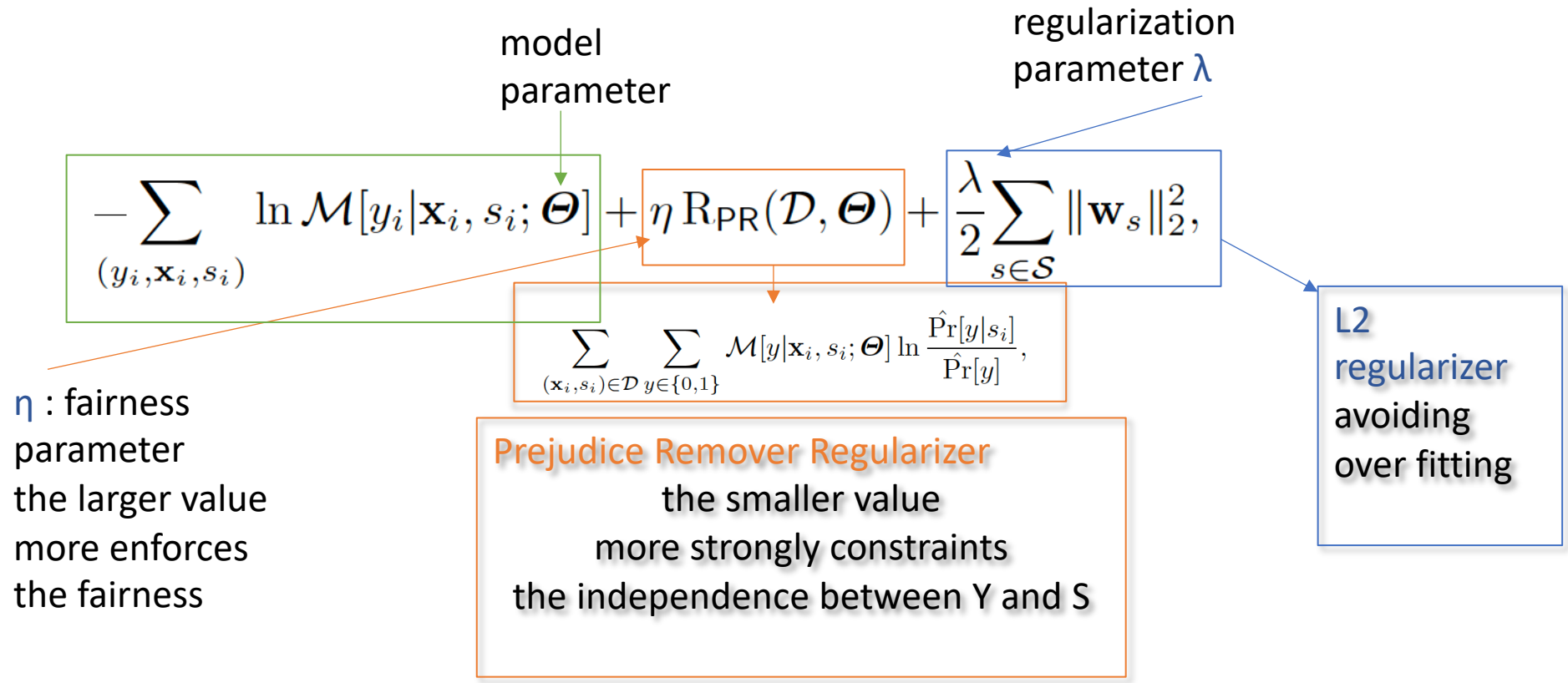
Add ability to adjust distribution of Y depending on the value of S

- Logistic Regression
- Application: applied to any algorithms with probabilistic discriminative

models and are simple to implement.

No Indirect Prejudice Conditon

Add a constraint of a no-indirect-prejudice condition

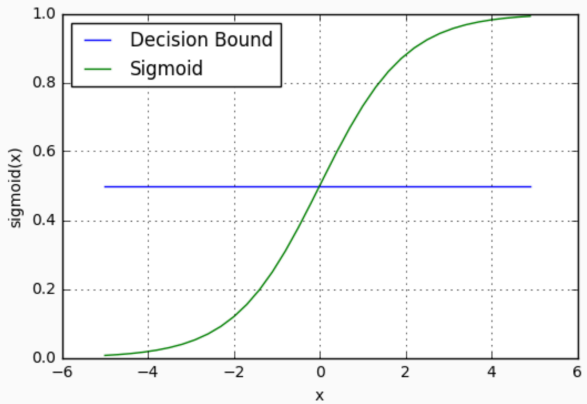


Log-likelihood

$$-\mathcal{L}(\mathcal{D};\boldsymbol{\Theta}) + \eta \mathrm{R}(\mathcal{D},\boldsymbol{\Theta}) + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_2^2,$$

$$S(z) = \frac{1}{1+e^{-z}}$$

$$p \geq 0.5, class = 1$$
$$p < 0.5, class = 0$$



$when\ y = 1, \quad y\sigma(X^T w_s) = \sigma(X^T w_s)$

$when\ y = 0, \quad (1-y)\sigma(X^T w_s) = (1-\sigma(X^T w_s))$

$$\mathcal{M}[y|\mathbf{x}, s; \boldsymbol{\Theta}] = y\sigma(\mathbf{x}^\top \mathbf{w}_s) + (1-y)(1-\sigma(\mathbf{x}^\top \mathbf{w}_s)),$$

$$\mathcal{L}(\mathcal{D};\boldsymbol{\Theta}) = \sum_{(y_i,\mathbf{x}_i,s_i)\in\mathcal{D}} \ln \mathcal{M}[y_i|\mathbf{x}_i, s_i; \boldsymbol{\Theta}].$$

A logarithm of a likelihood ratio is equal to the difference of the log-likelihoods:

$$\log \frac{L(A)}{L(B)} = \log L(A) - \log L(B) = \ell(A) - \ell(B).$$

Prejudice Remover Regularizer

$$-\mathcal{L}(\mathcal{D};\boldsymbol{\Theta}) + \boxed{\eta \mathrm{R}(\mathcal{D},\boldsymbol{\Theta})} + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_2^2,$$

no-indirect-prejudice condition = independence between Y and S

Prejudice Remover Regularizer

mutual information between Y (objective variable) and S (sensitive feature)

$$\mathrm{PI} = \sum_{Y,S} \hat{\Pr}[Y,S] \ln \frac{\hat{\Pr}[Y,S]}{\hat{\Pr}[S]\hat{\Pr}[Y]} = \sum_{X,S} \tilde{\Pr}[X,S] \sum_{Y} \mathcal{M}[Y|X,S;\boldsymbol{\Theta}] \ln \frac{\hat{\Pr}[Y,S]}{\hat{\Pr}[S]\hat{\Pr}[Y]}.$$

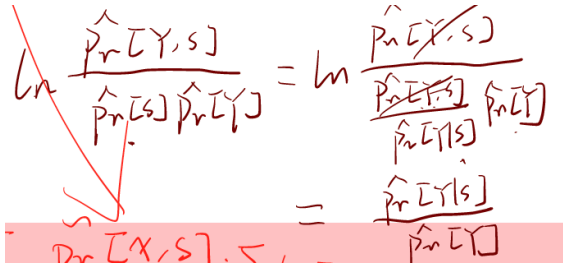
$$\sum_{Y,S} \widehat{\Pr[Y,S]} = \sum_{Y,S}\sum_{X} P[Y,S|X]\,P[X]$$

$$= \sum_{Y,S}\sum_{X} P[Y|X]\,P[S|X]P[X]$$

$$= \sum_{Y,S}\sum_{X} P[Y|X,S]\,P[X,S]$$

$\Pr[Y,X,S] = \mathcal{M}[Y|X]\Pr^*[S|X]\Pr^*[X].$

Y ⊥ S | X



$$\sum_{(\mathbf{x}_i,s_i)\in\mathcal{D}} \sum_{y\in\{0,1\}} \mathcal{M}[y|\mathbf{x}_i,s_i;\boldsymbol{\Theta}] \ln \frac{\hat{\Pr}[y|s_i]}{\hat{\Pr}[y]},$$

Prejudice Remover Regularizer

$$-\mathcal{L}(\mathcal{D};\boldsymbol{\Theta}) + \eta \mathrm{R}(\mathcal{D},\boldsymbol{\Theta}) + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_2^2,$$

$$\sum_{(\mathbf{x}_i,s_i)\in\mathcal{D}} \sum_{y\in\{0,1\}} \mathcal{M}[y|\mathbf{x}_i,s_i;\boldsymbol{\Theta}] \ln \frac{\hat{\Pr}[y|s_i]}{\hat{\Pr}[y]},$$

expectation over X and S
is replaced with
the summation over samples

This distribution can be derived by marginalizing over X

$$\hat{\Pr}[y|s] = \int_{\mathrm{dom}(X)} \Pr^*[X|s]\mathcal{M}[y|X,s;\boldsymbol{\Theta}]dX,$$

$$\hat{\Pr}[y|s] \approx \frac{\sum_{(\mathbf{x}_i,s_i)\in\mathcal{D} \text{ s.t. } s_i=s} \mathcal{M}[y|\mathbf{x}_i,s;\boldsymbol{\Theta}]}{|\{(\mathbf{x}_i,s_i)\in\mathcal{D} \text{ s.t. } s_i=s\}|}.$$

But this is computationally heavy...

approximate by the sample mean over X for each pair of y and s

$$\hat{\Pr}[y] \approx \frac{\sum_{(\mathbf{x}_i,s_i)\in\mathcal{D}} \mathcal{M}[y|\mathbf{x}_i,s_i;\boldsymbol{\Theta}]}{|\mathcal{D}|}.$$

Limitation: This technique is applicable, only if both Y and S are discrete

L2 regularizer

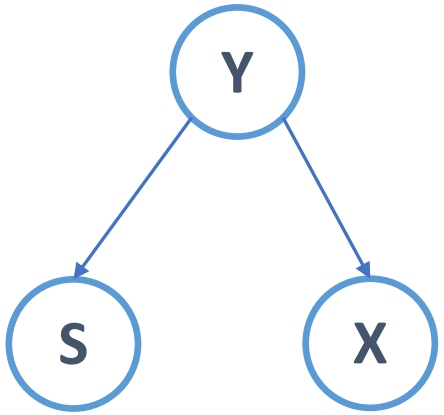
$$-\mathcal{L}(\mathcal{D};\boldsymbol{\Theta}) + \eta \mathrm{R}(\mathcal{D},\boldsymbol{\Theta}) + \boxed{\frac{\lambda}{2}\|\boldsymbol{\Theta}\|_2^2,}$$

***L_2* regularization**: sum of the squares of all the feature weights:

$$L_2 \text{ regularization term} = ||\boldsymbol{w}||_2^2 = w_1^2 + w_2^2 + \ldots + w_n^2$$

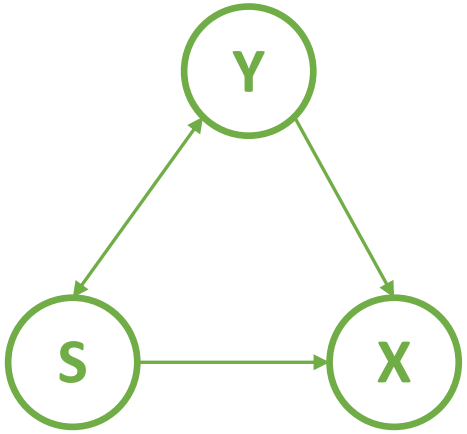
Calders-Verwer's 2 Naïve Bayes

Naïve Bayes



S and X are conditionally independent given Y

Calders-Verwer Two Naïve Bayes (CV2NB)



non-sensitive features X are mutually conditionally independent given Y and S

Calders-Verwer's 2 Naïve Bayes

It is as if two naïve Bayes classifiers are learned depending on each value of the sensitive feature

$$\Pr[Y, \mathbf{X}, S] = \mathcal{M}[Y, S] \prod_i \mathcal{M}[X_i | Y, S].$$

while CVscore > 0
    if # of data classified as "1" < # of "1" samples in original data then
        increase M[Y=+, S=-], decrease M[Y=-, S=-]
    else
        increase M[Y=-, S=+], decrease M[Y=+, S=+]
    reclassify samples using updated model M[Y, S]

Update the joint distribution so that its CV score decrease

Outline

Applications

- fairness-aware data mining applications

Difficulty in Fairness-aware Data Mining

- Calders-Verwer's discrimination score, red-lining effect

Fairness-aware Classification

- fairness-aware classification, three types of prejudices

Methods

- prejudice remover regularizer, Calders-Verwer's 2-naïve-Bayes

Experiments

- experimental results on Calders&Verwer's data and synthetic data

Related Work

- privacy-preserving data mining, detection of unfair decisions,
- explainability, fairness-aware data publication

Conclusion

Experiments

Calders & Verwer's Test Data

- Adult / Census Income @ UCI Repository
- Y : a class representing whether subject's income is High or Low
- S : a sensitive feature representing whether subject's gender
- X : non-sensitive features, all features are discretized, and 1-of-K representation is used for logistic regression
- # of samples : 16281

Methods

- LRns : logistic regression without sensitive features
- NBns : naïve Bayes without sensitive features
- PR : our logistic regression with prejudice remover regularizer
- CV2NB : Calders-Verwer's two-naïve-Bayes

Other Conditions

L2 regularization parameter $\lambda = 1$
five-fold cross validation

Evaluation Measure

Accuracy

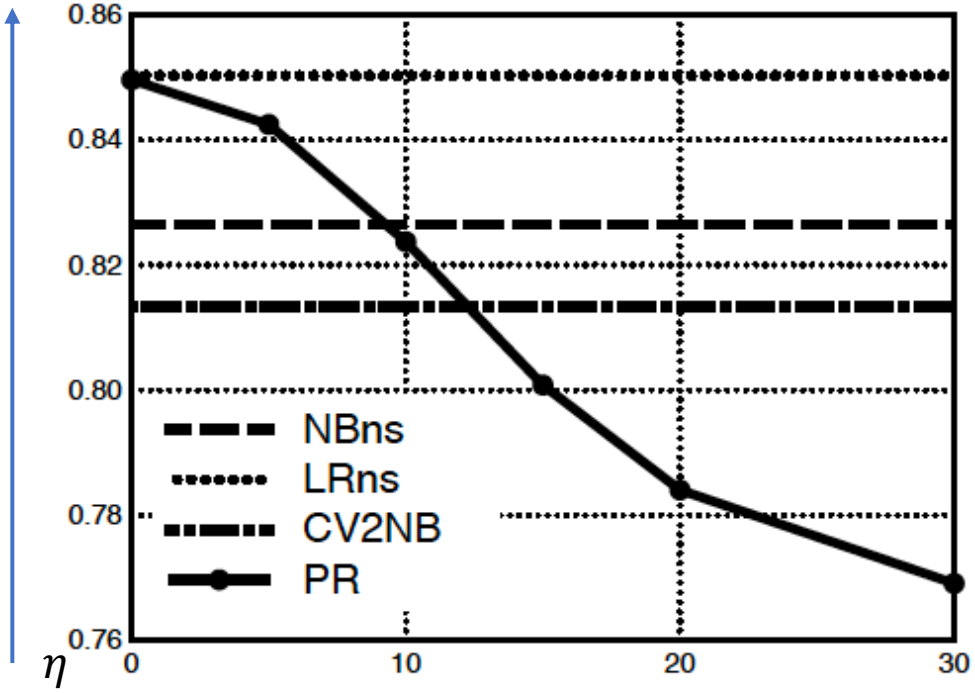
- How correct are predicted classes?
- the ratio of correctly classified sample

NMI (normalized mutual information)

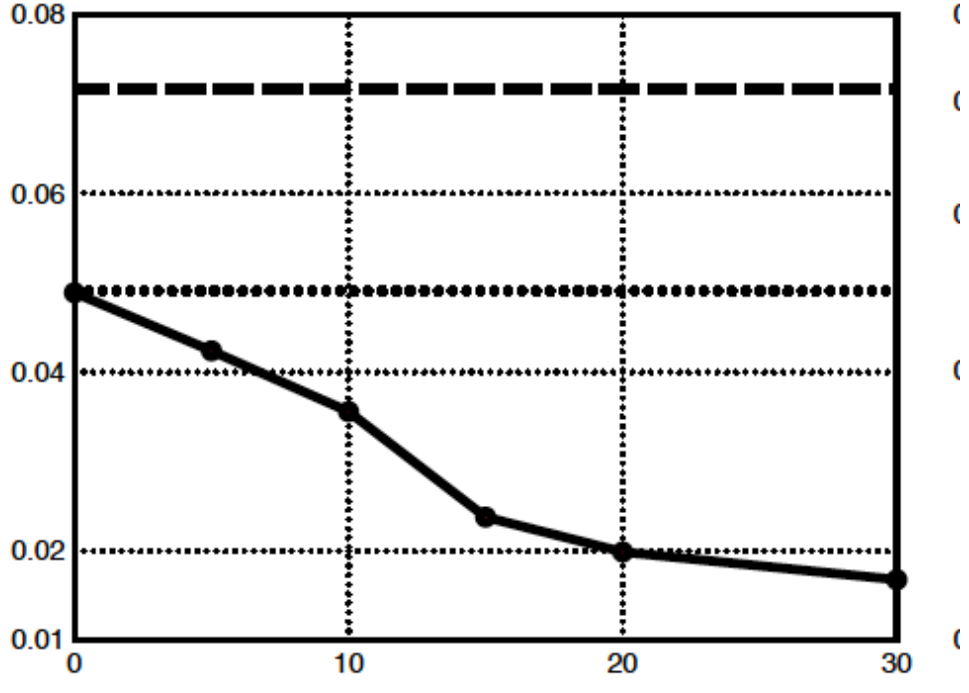
- How fair are predicted classes?
- mutual information between a predicted class and a sensitive feature, and it is normalized into the range [0, 1]

$$\mathrm{NMI} = \frac{\mathrm{I}(Y;S)}{\sqrt{\mathrm{H}(Y)\mathrm{H}(S)}}$$

Higher Accuracy



(a) Acc



(b) NPI

Fairer

fairness parameter η:
$larger\ \eta\ means\ enhance\ fairness\ more$

η ↑ accuracy ↓ fairness(NPI) ↑

PR (Prejudice Remover) could make fairer decisions than pure LRns (logistic regression) and NBns (naïve Bayes)

PR could make more accurate prediction than NBns or CV2NB

CV2NB achieved near-zero NMI, but PR could NOT achieve it

Synthetic Data

Why did our prejudice remover fail to make a fairer prediction than that made by CV2NB?

$\epsilon_i \leftarrow$ sample $\sim N(0,1)$

$S_i \in \{0,1\}$

$x_{ai} = \epsilon_i$

$x_{bi} = 1 + \epsilon_i \mid S_i = 1$

$x_{bi} = -1 + \epsilon_i \mid S_i \neq 1$

$\sigma(x_{ai}^T W s_i)$

$\sigma(x_{bi}^T W s_i)$

$y_i = 0 \mid x_{ai} + x_{bi} < 0$

$y_i = 1 \mid \text{---} > 0$

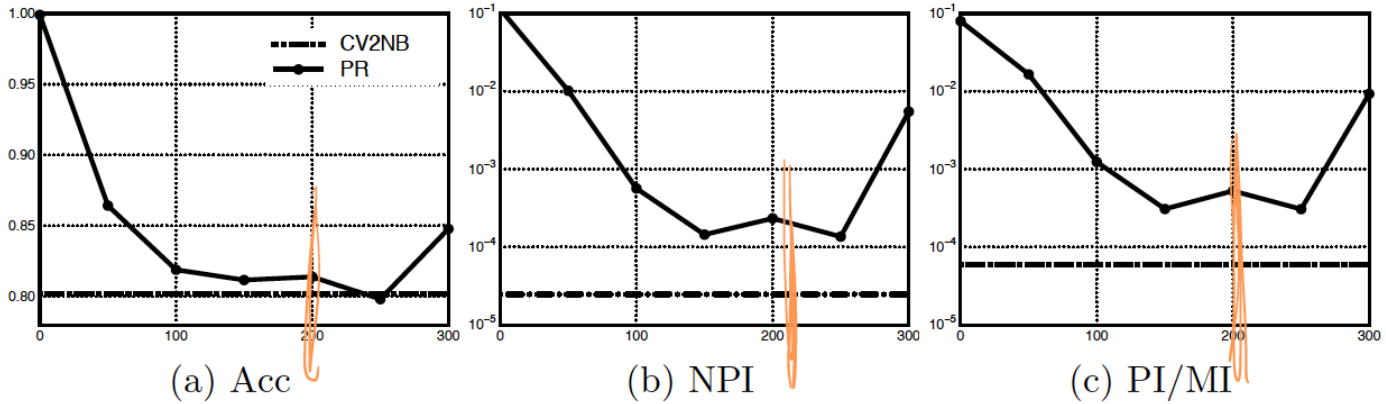


Fig. 2. The change in performance for our synthetic data according to η

Synthetic Data

Table 2. The learned weight vectors $\mathbf{w}_0$ and $\mathbf{w}_1$ in equation (8)

	$\mathbf{w}_0$	$\mathbf{w}_1$
$\eta = 0$	$[11.3,\ 11.3\ ,-0.0257]$	$[11.3,\ 11.4\ ,0.0595]$
$\eta = 150$	$[55.3,-53.0,\ -53.6\ ]$	$[56.1,-54.1,\ 53.6\ ]$
	x_{ai} x_{bi}	

When η=0, PR regularizer doesn't affect weights
When η=150, w(xbi) < w(xai).

Outline

Applications

- fairness-aware data mining applications

Difficulty in Fairness-aware Data Mining

- Calders-Verwer's discrimination score, red-lining effect

Fairness-aware Classification

- fairness-aware classification, three types of prejudices

Methods

- prejudice remover regularizer, Calders-Verwer's 2-naïve-Bayes

Experiments

- experimental results on Calders&Verwer's data and synthetic data

Related Work

- privacy-preserving data mining, detection of unfair decisions,
- explainability, fairness-aware data publication

Conclusion

Indirect prejudice
the dependency between a objective Y and a sensitive feature S

From the information theoretic perspective
mutual information between Y and S is non-zero

From the viewpoint of privacy-preservation
leakage of sensitive information when an objective variable is known

- Kamiran : fairness in Decision trees
- Finding Unfair Association Rules
- Luong: Situation Testing

Outline

Applications

- fairness-aware data mining applications

Difficulty in Fairness-aware Data Mining

- Calders-Verwer's discrimination score, red-lining effect

Fairness-aware Classification

- fairness-aware classification, three types of prejudices

Methods

- prejudice remover regularizer, Calders-Verwer's 2-naïve-Bayes

Experiments

- experimental results on Calders&Verwer's data and synthetic data

Related Work

- privacy-preserving data mining, detection of unfair decisions,
- explainability, fairness-aware data publication

Conclusion

Opinions

- Definition of fairness, sensitive variable
- lack of convexity: local minima
- binary sensitive variable and target variable
- Socially responsible mining