公平ロジスティック回帰での
確定的決定則の影響

神嶋 敏弘¹，赤穂 昭太郎¹，麻生 英樹¹，佐久間 淳²
¹産業技術総合研究所
²筑波大学/理化学研究所 革新知能統合研究センター
2018年度人工知能学会全国大会（第32回）@ 鹿児島市，2018-6-6
http://www.kamishima.net
Overview

Fair Logistic Regression
Logistic Regression whose decision is designed to ignore a specified information

Ex. In the decision of employment, the decision is not influenced by socially sensitive information, such as a gender or a race

Trade-off: accuracy vs fairness
The efficiency of the trade-off is POOR in our logistic regression

Ignorance of the influence of a decision rule and model bias

The trade-off is drastically improved
Applications
- Suspicious Placement Keyword-Matching Advertisement

Fairness in Machine Learning
- Notations and Independence between a target variable and a sensitive feature

Fairness-aware Classifier
- Our fairness-aware Classifier, logistic regression with prejudice remover

Model-based Independence & Actual Independence
- Two types of independence and experimental results

Smoothing Relaxation
- The objective is approximated by a smooth function

Conclusion
Applications
Online advertisements of sites providing arrest record information

Advertisements indicating arrest records were more frequently displayed for names that are more popular among individuals of African descent than those of European descent.

African descent’s name

Arrested?

negative ad-text

European descent’s name

Located:

neutral ad-text
Suspicious Placement Keyword-Matching Advertisement

Selection of ad-texts was unintentional

Response from advertiser:
- Advertise texts are selected based on the last name, and no other information in exploited
- The selection scheme is adjusted so as to maximizing the click-through rate based on the feedback records from users by displaying randomly chosen ad-texts

No sensitive information, e.g., race, is exploited in a selection model, but suspiciously discriminative ad-texts are generated

An annotation bias is caused due to the unfair feedbacks from users reflecting the users’ prejudice
Fairness in Machine Learning
Notations of Variables

**Y** target variable / object variable

An objective of decision making, or what to predict
- \( Y \): true / population, \( \hat{Y} \): predicted, \( Y^* \): fairized
  - ex., loan approval, university admission, what to recommend

**S** sensitive feature

To ignore the influence to the sensitive feature from a target
- Specified by a user or an analyst depending on his/her purpose
- It may depend on a target or other features
- It can be multivariate
  - ex., socially sensitive information (gender, race), items’ brand

**X** non-sensitive feature vector

All features other than a sensitive feature
**Annotation Bias**: Target values or feature values in a training data are biased due to annotator’s cognitive bias or inappropriate observation schemes.

Annotations are not reliable, and never accessible to a correct dataset.

**Assumptions about the conditions that values or distributions of target variables and sensitive features should satisfy**

**Examples of assumptive conditions**:
- \( Y \perp S \mid X=x \): \( Y \) and \( S \) are context-sensitive independent given \( X=x \)
- \( Y \perp S \mid X \): \( Y \) and \( S \) are conditionally independent given \( X \)
- \( Y \perp S \): \( Y \) and \( S \) are (unconditionally) independent
Independence between $Y$ and $S$

Removing annotation bias: $\hat{Y} \perp S$

<table>
<thead>
<tr>
<th>$S = 0$</th>
<th>$S = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{Y} = 0$</td>
<td>$\hat{Y} = 1$</td>
</tr>
<tr>
<td>$\hat{Y} = 1$</td>
<td>$\hat{Y} = 0$</td>
</tr>
</tbody>
</table>

Ratios between positives and negatives in prediction are matched

[Calders+ 10, Dwork+ 12]
Red-Lining Effect: Simple elimination of a sensitive feature from training dataset fails to remove the influence of sensitive information to a target.

- \( \Pr[ Y \mid X, S ] \): A model trained from a dataset with both sensitive and non-sensitive features.
- \( \Pr[ Y \mid X ] \): A model that does not depend on \( S \) by eliminating a sensitive feature from a training dataset.

\[
\Pr[ Y, X, S ] = \Pr[ Y \mid X, S ] \Pr[ S \mid X ] P[ X ] \Rightarrow \Pr[ Y \mid X ] \Pr[ S \mid X ] P[ X ]
\]

This is a condition \( Y \perp S \mid X \) (not \( Y \perp S \))

\( S \) still influences \( Y \) through \( X \).
Fairness-aware Classifier
fairness-aware classification

find a fair model that approximates a true distribution instead of a fair true distribution under the fairness constraints

We want to approximate fair true distribution, but samples from this distribution cannot be obtained, because samples from real world are potentially unfair
Prejudice Remover: a regularizer to impose a constraint of independence between a target and a sensitive feature, $Y \independent S$

The objective function is composed of classification loss and fairness constraint terms.

$$ - \sum_{D} \ln \Pr[Y \mid X, S; \Theta] + \frac{\lambda}{2} \|\Theta\|_2^2 + \eta I(Y; S) $$

- A class distribution, $\Pr[Y \mid X, S; \Theta]$, is modeled by a set of logistic regression models, each of which corresponds to $s \in \text{Dom}(S)$
  $$ \Pr[Y=1 \mid x, s] = \text{sig}(w^{(s)}x) $$

- As a prejudice remover regularizer, we adopt a mutual information between a target and a sensitive feature, $I(Y; S)$

[Prejudice Remover Regularizer] [Kamishima+ 12]
Model-based Independence & Actual Independence
Even if $Y$ and $S$ are independent, actual class labels may not satisfy a fairness constraint.

**Deterministic decision rule**: Class labels are generated not probabilistically, but deterministically by a decision rule.

**Difference**: $Pr[Y, S] - Pr[Y] Pr[S]$

- **Always Independent**: Labels probabilistically generated according to $Pr[Y] Pr[S] Pr[X | Y, S]$
- **Not Independent in general**: Bayes optimal Labels are generated by a deterministic decision rule: $y^* \leftarrow \arg \max_y Pr[y|x, s]$

**Model bias**: Models doesn’t contain true distribution to learn in general.
Model-Based & Actual Independence

**Model-based Independence**: Class labels are assumed to be generated probabilistically

\[ \hat{Y} \perp S, \text{ where } (\hat{Y}, S) \sim \Pr[\hat{Y}, S] \]

**Actual Independence**: Class labels are assumed to be deterministically generated by applying a decision rule

\[ \tilde{Y} \perp S, \text{ where } (\tilde{Y}, S) \sim \Pr[\tilde{Y}, S] = \frac{1}{n} \sum_{x \in D_S} \Pr[\tilde{Y}|x, S] \]

\[
\begin{align*}
\Pr[\tilde{y}|x, s] &= 1, & \tilde{y} &= \arg \max_y \Pr[\tilde{y}|x, s] \\
\Pr[\tilde{y}|x, s] &= 0, & \text{otherwise}
\end{align*}
\]

Satisfy actual independence instead of model-based independence

Fairness in class labels can be drastically improved
Experimental Results

Accuracy (Acc) versus Fairness (NMI) for different values of the fairness parameter $\eta$.

- More accurate: $\eta$ increases, $\text{Acc}$ decreases, $\text{fairness}$ increases.
- Fairer: $\eta$ increases, $\text{fairness}$ decreases, $\text{Acc}$ increases.

- Accuracy and fairness have a trade-off relation.
- By satisfying actual independence, instead of model-based independence, the trade-off was drastically improved.
Smoothing Relaxation
The objective satisfying actual independence is hard to optimize

\[
\begin{cases}
\Pr[\hat{y}|x, s] = 1, \quad \hat{y} = \arg \max_y \Pr[\hat{y}|x, s] \\
\Pr[\hat{y}|x, s] = 0, \quad \text{otherwise}
\end{cases}
\]

The objective contains discrete function; and it is indifferentiable

Replace a step function with a smooth sigmoid function \( \text{sig}(x) \)

equivalent to satisfying model-based independence; and meaningless

Use more steepest sigmoid function \( \text{sig}(\phi x) \)
Results: Smoothing Relaxation

- The smoothing relaxation can perform better than the best method
- The performance was very sensitive to the parameter $\phi$

Plotting accuracy vs fairness
- Initialized by standard LR and ROC-AI
- **ROC-AI**: the best performed method by tuning an intercept or LR
Conclusions

- We examined the reason why the trade-offs between accuracy and fairness is poor in a fairness-aware logistic regression classifier.
- We advocate the notions of model-based independence and actual independence.
- We empirically show the more fair classifiers can be obtained by satisfying actual independence, instead of model-based independence.
- To improve the computational efficiency, we develop a modified objective function, called by a smoothing relaxation.

More Information: http://www.kaishima.net/fadm/

Acknowledgements: This work is supported by MEXT/JSPS KAKENHI Grant Number JP24500194, JP15K00327, and JP16H02864