PAPER

Prediction with Model-Based Neutrality

Kazuto FUKUCHI[†a)], ***Nonmember***, **Toshihiro KAMISHIMA**[††], ***and*** **Jun SAKUMA**[†], ***Members***

SUMMARY With recent developments in machine learning technology, the predictions by systems incorporating machine learning can now have a significant impact on the lives and activities of individuals. In some cases, predictions made by machine learning can result unexpectedly in unfair treatments to individuals. For example, if the results are highly dependent on personal attributes, such as gender or ethnicity, hiring decisions might be discriminatory. This paper investigates the neutralization of a probabilistic model with respect to another probabilistic model, referred to as a viewpoint. We present a novel definition of neutrality for probabilistic models, η-neutrality, and introduce a systematic method that uses the maximum likelihood estimation to enforce the neutrality of a prediction model. Our method can be applied to various machine learning algorithms, as demonstrated by η-neutral logistic regression and η-neutral linear regression.

key words: *neutrality, fairness, discrimination, logistic regression, linear regression, classification, regression, social responsibility*

1. Introduction

With recent developments in machine learning technology, the resulting predictions can now have a significant impact on the lives and activities of individuals. In some cases, there are safeguards in place so that the predictions do not cause unfair treatment, discrimination, or biased views of individuals [1]. The following two examples describe situations in which predictions made by machine learning can cause unfair treatment.

Example 1 (hiring decision) A company collects personal information from employees and job applicants; this information includes age, gender, race or ethnicity, place of residence, and work experience. The company uses machine learning to predict the work performance of the applicants, using information collected from employees. The hiring decision is then based on this prediction.

Example 2 (personalized advertisement and recommendation) A company that provides web services records user behavior, including usage history and search logs, and uses machine learning to predict user attributes and preferences. The advertisements or recommendations displayed on web pages are thus personalized so that they match the predicted user attributes and preferences.

In the hiring-decision example, if the results are highly dependent on personal attributes, such as gender or ethnicity, hiring decisions might be deemed discriminatory. In the second example, when recommendations are accurately pinpointed to sensitive issues, such as political or religious affiliation, the result may be increasingly biased views. This is known as the problem of the filter bubble [2]. For example, suppose supporters of the Democratic Party wish to read news articles related to politics. If the recommended articles are all related to their party and are absent of criticism, they may develop a biased view of the political situation. In the web-service example, showing advertisements that suit the user's attributes, such as gender or age, would improve the service for some users. Other users, however, may object to advertisements that are apparently based on their race, ethnicity, or gender. Thus, it is difficult to clearly distinguish personalization from discrimination.

We now introduce some terms that will be useful in the following discussion. The input and output of a prediction model are referred to as *input* variables (e.g., race, ethnicity, or web-usage history) and *target* variables (hiring decisions or website recommendations). Factors that might result in discrimination or bias are referred to as *viewpoint* variables (e.g., race, ethnicity, or political affiliation).

The objective of machine learning is to learn prediction functions that predict target variables from given examples. In the example above, if the viewpoint variables (e.g., race or ethnicity) are dependent on the predicted target variables (e.g., hiring decisions), the prediction function causes unfair treatment. In this paper, we introduce a systematic way to remove this dependency from prediction models and neutralize them with respect to a given viewpoint.

1.1 Related Works

Several techniques that take account of fairness or discrimination have recently received attention [5], [7]–[9]. One of the easiest ways to suppress unfair treatment is to remove the values of the viewpoint from the input values before the learning process with the prediction model. If there is no correlation between the input and viewpoint variables, no discrimination or bias will appear after elimination. However, if another input variable is dependent on the viewpoint variable, then even after the viewpoint values are eliminated, the target variable will retain dependency on the viewpoint variable (Table 1, line 1). For example, assume that the

Manuscript received October 21, 2014.
Manuscript revised March 21, 2015.
Manuscript publicized May 15, 2015.

[†]The authors are with the Graduate School of System and Information Engineering, University of Tsukuba, Tsukuba-shi, 305–8577 Japan.

[††]The author is with National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba-shi, 305–8568 Japan.

a) E-mail: kazuto@mdl.cs.tsukuba.ac.jp

DOI: 10.1587/transinf.2014EDP7367

Table 1 Summary of learning algorithms with neutrality guarantee.

method	neutrality guarantee	domain of target	domain of viewpoint	model of viewpoint
elimination of viewpoint variable	no guarantee	any	any	×
CV2NB [3]	CV Score	multiple	multiple	×
PR [4]	mutual information	any	multiple	×
Lipschitz property [5]	statistical parity	multiple	multiple	×
LFR [6]	statistical parity	multiple	multiple	×
η-neutral logistic regression (proposal)	η-neutrality	multiple	multiple	√
η-neutral linear regression (proposal)	η-neutrality	continuous	continuous	√

race or ethnicity attribute is eliminated in Example 1. Even so, hiring decisions may be dependent on race or ethnicity if there is a correlation between individuals' addresses and their race or ethnicity; this is known as the redlining effect [3], [10].

Calders et al. presented the *Calders–Verwer* 2 *Naive Bayes method* (CV2NB), which proactively removes the redlining effect [3]. Let $y \in \{y_+, y_-\}$ be the binary target variable, and let $v \in \{v_+, v_-\}$ be the binary viewpoint variable. Then, the Calders–Verwer (CV) score is defined by $\mathrm{CV}(\mathcal{D}) = \Pr(y_+|v_+) - \Pr(y_+|v_-)$. The CV2NB modifies the näive Bayes classifier in such a way that the CV score becomes zero with respect to the given examples $\mathcal{D}$. The CV2NB guarantees the elimination of discrimination in terms of the CV score. The limitation of the CV2NB is that it cannot be used when the target or viewpoint variables are continuous (Table 1, line 2). Related to the CV2NB, it has been shown [11] that positive CV scores do not necessarily cause discrimination in some situations. There is also a method [12] that uses the kth-nearest neighbor to test for the existence of discrimination. Both these methods are based on the CV2NB, so they share its limitations.

Kamishima et al. introduced the *prejudice remover regularizer* (PR) for fairness-aware classification [4]. The PR regularizer penalizes the objective function if there is a high correlation between the target variable and the viewpoint variable. The penalty, which is called as the *prejudice index*, is evaluated based on the information that is shared by the target variable y and the viewpoint variable v. This penalty function can work with a continuous target variable if it is approximated by a histogram, as demonstrated by Kamishima et al. [13], [14]. Continuous viewpoint variables, however, cannot be treated by the PR method (Table 1, line 3).

Dwork et al. presented a classification method that uses a fairness-aware framework, in which statistical parity is used as the measure of fairness [5]. Intuitively, statistical parity occurs when the demographics of those receiving positive (or negative) classifications are identical to the demographics of the population as a whole. In their fairness-aware framework, the classification is made to be fair by minimizing the empirical risk while satisfying certain constraints that are called the *Lipschitz property*[†]. As is the case with the CV2NB and PR methods, this framework assumes that the viewpoint variables are binary or multiple; continuous viewpoint variables are not considered (Table 1, line 4).

Zemel et al. proposed *learning fair representation* (LFR), aiming to obtain an intermediate representation which encodes the given data while simultaneously removing any information about the viewpoint variable. The fairness of LFR is based on statistical parity, and that leads the limitation for use of continuous viewpoint variables (Table 1, line 5).

1.2 Our Contribution

Modeling viewpoint variables. In this manuscript, we provide a method to neutralize the target prediction model with respect to a probabilistic model of a given viewpoint. Existing methods assume the viewpoint is observed and is explicitly provided in the input, but this is not always the case. For instance, consider the recommendation of articles neutralized with respect to political affiliation, as in Example 2. Political affiliation is not explicitly included in the input, but given as input the logs of keyword searches or subscribed news articles, modern machine learning techniques can easily predict party affiliation. In such a case, our method neutralizes the target prediction model with respect to the probability model of such a "hidden viewpoint".

In order to neutralize a model with respect to a viewpoint, we represent the viewpoint as a probabilistic model and define η-neutrality (Sect. 2), which is a measure of the dependency of the target prediction model on the viewpoint prediction model. With η-neutrality, we can check the neutrality of a target prediction model with respect to any hidden viewpoint, as long as we have a probabilistic model of the viewpoint variable (Table 1, the rightmost column). Furthermore, since η-neutrality is measured with respect to probabilistic models, the neutrality of the prediction model with respect to unseen examples is expected to be effectively guaranteed, and this is demonstrated by experiments (Sect. 5).

Maximum likelihood estimation with η-neutrality. Following the definition of η-neutrality, we introduce a systematic method that removes this dependency from the prediction model obtained by the maximum likelihood estimation (Sect. 2). Our methods can treat target and viewpoint variables that are either discrete (Table 1, line 5) or continuous (Table 1, line 6), as demonstrated by η-neutrality

[†]Lipschitz property differs to commonly known Lipschitz continuity

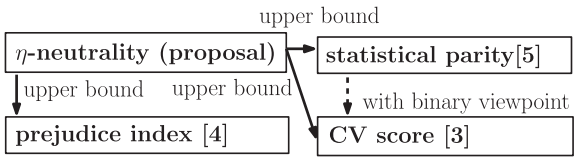


Fig. 1 Relationship between the neutrality measures. Statistical parity with binary viewpoint is equivalent to CV score. η-neutrality upper bounds all of other neutrality measures.

with logistic regression (Sect. 3.1) and linear regression (Sect. 3.2). The effectiveness of our methods is examined by both artificial and real datasets in Sect. 5.

Comparison of neutrality measures.

We clarify the relationship between the existing neutrality measures, η-neutrality, the CV Score [3], statistical parity [5] and the prejudice index [4]. For comprehensive discussion, we introduce a *neutrality factor* (Sect. 4), which represents neutrality of a pair of a target value and viewpoint value. We show that existing neutrality measures are universally represented by aggregation of the neutrality factors.

In Fig. 1, we illustrate the relationship between the neutrality measures. The dashed arrow between statistical parity and the CV score shows that statistical parity with binary viewpoint is equivalent to CV score. Furthermore, η-neutrality is interpreted as an upper-bound of the other neutrality measures represented by the solid arrows. We will prove these relations in Sect. 4.

2. η-Neutrality

We propose a novel definition of neutrality, η-neutrality. We then present a general maximum likelihood estimation method that has a guarantee of neutrality.

2.1 Problem Setting: Maximum Likelihood Estimation

Let $\mathcal{D} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1}^N$ be a set of training examples that are assumed to be i.i.d. samples drawn from a probability distribution $\Pr(X, Y)$. The random variables X and Y are referred to as the input and target, respectively. The realized values of the variables are denoted by the corresponding lowercase letters. Thus, the random variable X can take the value x. In the following discussion, we assume the input random variable X is continuous. We can treat a discrete X by replacing the integral with a sum. For Y, the discussion below is valid for both discrete and continuous variables. Besides, the prediction function of the target variable is represented as a probabilistic model $f(Y|X;\theta) = \Pr(Y|X)$, parametrized by θ. The target prediction model can be obtained by minimization of the negative log-likelihood with respect to the parameter θ:

$$\theta^* = \mathrm{argmin}_{\theta\in\Theta} L(\theta),$$

where

$$L(\theta) = -\sum_{(x_i,y_i)\in\mathcal{D}} \ln f(y_i|x_i;\theta). \tag{1}$$

2.2 Definition of η-Neutrality

In addition to the input random variable X and the target random variable Y, we now introduce the viewpoint random variable V. Let $\mathcal{V}$ be the domain of V. As is the case with Y, the discussion below with respect to V is also valid for both discrete and continuous variables. As we did for the target random variable, we assume that the prediction model of the viewpoint variable is represented as a conditional probability $\Pr(V|X)$. Noting that the values of the target and the viewpoint variables are predicted independently, we assume the joint probability is

$$\Pr(X, Y, V) = \Pr(X)\Pr(Y|X)\Pr(V|X).$$

With this assumption, we consider the dependency of the target random variable Y and the viewpoint random variable V. When V and Y are statistically independent, for any $y \in \mathcal{Y}$ and $v \in \mathcal{V}$, $\Pr(v, y)/\Pr(v)\Pr(y) = 1$. When $\Pr(v, y)/\Pr(v)\Pr(y) > 1$, v and y are more dependent than independent. Hence, our neutrality definition is defined as the ratio of the marginal probabilities, as follows.

Definition 1 (η-neutrality)**.** *Let X and Y be the input and target random variables, respectively. Let V denote the viewpoint random variable. Given $\eta \geq 0$, the probability distribution $\Pr(X, Y, V)$ is η-neutral if*

$$\forall v \in \mathcal{V}, y \in \mathcal{Y}, \quad \frac{\Pr(v, y)}{\Pr(v)\Pr(y)} \leq 1 + \eta. \tag{2}$$

Our neutrality definition simply bounds above the ratio. As a variation of this definition, the ratio can be bounded above and below as

$$\forall v \in \mathcal{V}, y \in \mathcal{Y}, \quad 1 - \eta \leq \frac{\Pr(v, y)}{\Pr(v)\Pr(y)} \leq 1 + \eta. \tag{3}$$

If both the target random variable and the viewpoint random variable are binary, the ratio of our definition is bounded below as Eq. (3). If either or both of the target random variable and the viewpoint random variable take M multiple values, the ratio of our definition is bounded below by $1 - M\eta$ which is different from Eq. (3). We employed Definition 1 for optimization efficiency. The number of constraints derived from Definition 1 can be reduced to half compared to Eq. (3).

Next, given the probabilistic models of $\Pr(Y|X)$ and $\Pr(V|X)$, we derive conditions that the model of the joint probability distribution satisfies η-neutrality. The target and the viewpoint prediction models are described by the probability distributions $f(Y|X;\theta) = \Pr(Y|X)$ and $g(V|X;\phi) = \Pr(V|X)$, respectively, where θ and ϕ are the model parameters. Thus, given the target prediction model $f(Y|X;\theta)$ and the viewpoint prediction model $g(V|X;\phi)$, the probabilistic model of $\Pr(X, Y, V)$ becomes

$$M(X, Y, V;\theta,\phi) = f(Y|X;\theta)g(V|X;\phi)\Pr(X). \tag{4}$$

In what follows, we assume the viewpoint prediction

model is fixed, and so the model parameter ϕ is omitted and g is described by $g(V|X)$. The following theorem shows the condition that the model of Eq. (4) is empirically η-neutral.

Theorem 1. *Suppose the joint probability distribution of input X, target Y, and viewpoint V follows the model* $M(X, Y, V; \theta) = \Pr(X) f(Y|X; \theta) g(V|X)$. *Then M is η-neutral if* $\forall v \in \mathcal{V}, y \in \mathcal{Y}$,

$$\int_x \Pr(x) f(y|x;\theta) \left[g(v|x) - (1+\eta)\bar{g}(v) \right] dx \leq 0, \tag{5}$$

where $\bar{g}(v) = \int_x \Pr(x) g(v|x) dx$.

Proof. By the marginalization of $\Pr(x, y, v)$ with respect to x, (x, y), and (x, v), we have

$$\begin{aligned}
\Pr(y, v) &= \int_x \Pr(x, y, v)\, dx = \int_x \Pr(x) f(y|x;\theta) g(v|x) dx, \\
\Pr(y) &= \int_x \int_v \Pr(x, y, v)\, dv dx = \int_x \Pr(x) f(y|x;\theta) dx, \\
\Pr(v) &= \int_x \int_y \Pr(x, y, v)\, dy dx = \int_x \Pr(x) g(v|x) dx \\
&= \bar{g}(v).
\end{aligned}$$

By substituting the above equations into Eq. (2), we have

$$\begin{aligned}
&\forall v, y, \int_x \Pr(x) f(y|x;\theta) g(v|x) dx \\
&\qquad - (1+\eta)\bar{g}(v) \int_x \Pr(x) f(y|x;\theta) dx \leq 0, \\
&\forall v, y, \int_x \Pr(x) f(y|x;\theta) \left[g(v|x) - (1+\eta)\bar{g}(v) \right] dx \leq 0.
\end{aligned}$$

□

2.3 Approximation of η-Neutrality

When $\Pr(x)$ cannot be obtained, η-neutrality can be empirically evaluated with respect to the frequency distribution $\tilde{\Pr}(x)$ of the examples $\mathcal{D}$. The neutrality condition with respect to this frequency distribution is derived in a similar manner, as follows. Given examples $\mathcal{D}$, we approximate η-neutrality with respect to the frequency distribution

$$\tilde{\Pr}(X = x) = \frac{1}{N} \sum_{i=1}^{N} I(x_i = x),$$

where $I(\cdot)$ denotes the indicator function. From this, we have

$$\tilde{\Pr}(X, Y, V) = \tilde{\Pr}(X) \Pr(Y|X) \Pr(V|X),$$

and an approximation of η-neutrality is defined by this $\tilde{\Pr}(X, Y, V)$.

Definition 2 (Empirical η-neutrality)**.** *Let X and Y be the input and target random variables, respectively. Let V denote the viewpoint random variable. Let* $\tilde{\Pr}(X)$ *be the frequency distribution of X obtained from* $\mathcal{D}$. *Given* $\eta \geq 0$, *if* $\tilde{\Pr}(X, Y, V)$ *is η-neutral,* $\Pr(X, Y, V)$ *is said to be empirically η-neutral with respect to the dataset* $\mathcal{D}$.

The following theorem shows the condition that the model of Eq. (4) is η-neutral with respect to the given examples.

Theorem 2. *Suppose the joint probability distribution of the input X, target Y, and viewpoint V follows the model* $M(X, Y, V; \theta) = \Pr(X) f(Y|X; \theta) g(V|X)$. *Then, given* $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, *M is empirically η-neutral if*

$$\forall y, v, \sum_{i=1}^{N} f(y|x_i;\theta) \left[g(v|x_i) - (1+\eta)\tilde{g}(v) \right] \leq 0,$$

where $\tilde{g}(v) = \frac{1}{N} \sum_{i=1}^{N} g(v|x_i)$.

Proof. Theorem 2 states that $\Pr(X, Y, V)$ is η-neutral if Eq. (5) holds. By substituting $\tilde{\Pr}(X)$ into Eq. (5), the neutrality condition is rewritten as

$$\forall y, v, \frac{1}{N} \sum_{i=1}^{N} f(y|x_i) \left[g(v|x_i) - (1+\eta)\tilde{g}(v) \right] \leq 0.$$

□

For convenience in the following discussion, the neutrality condition is notated as

$$N(y, v) = \sum_{i=1}^{N} f(y|x_i) \left[g(v|x_i) - (1+\eta)\tilde{g}(v) \right] \leq 0. \tag{6}$$

2.4 Maximum Likelihood Estimation with η-Neutrality

Given examples and a viewpoint prediction model, we performed maximum likelihood estimations with the guarantee of η-neutrality. We wanted a target prediction model that would achieve the maximum log-likelihood with respect to the given data. At the same time, we wanted a target prediction function that would make $\Pr(X, Y, V)$ empirically η-neutral with respect to the given data and viewpoint prediction model. This problem is the following constrained optimization problem:

$$\text{minimize } L(\theta) \quad \text{subject to } N(y, v; \theta) \leq 0, \quad \forall y, v.$$

Existing neutrality indexes measure neutrality with certain statistics, such as differences in the conditional probabilities [3] or mutual information [4]. If such measures are used to guarantee neutrality, the neutrality of the model is statistically guaranteed for the set of given examples. In principle, it is desirable to guarantee neutrality with respect to each individual contained in the given examples. However, such prediction functions tend to overfit to the given examples and do not provide neutrality of unseen examples.

Assuming the model of the viewpoint correctly represents the true distribution, a model that satisfies our η-neutrality condition guarantees statistical independency between every combination of target value y and viewpoint value v. Note that η-neutrality can be realized even when the viewpoint values are not contained in the given examples. This is because the evaluation of neutrality is not dependent on the value of the viewpoint but the model of the viewpoint.

2.5 Prediction Model for Viewpoints

In principle, we assume $g(V|X)$ accurately represents the true probabilistic distribution $\Pr(V|X)$, but in reality, this does not always hold. In this subsection, we consider three types of possible viewpoint models.

The first case assumes an extreme example; model $g(V|X)$ is the probabilistic model that outputs random or constant values independent of input x. If we have no knowledge of the viewpoint, we have no choice other than this. Since $g(V|X)$ takes a constant value independent of X, η-neutrality is guaranteed for any $f(Y|X;\theta)$ in this model; however, such neutralization is meaningless.

The second case assumes that model $g(V|X)$ is taken as the empirical distribution of the training examples. Existing methods, including CV2NB, statistical parity, and PR, achieve neutralization with respect to this empirical distribution. This model realizes neutralization with respect to the given training examples, but neutralization with respect to unseen examples is not guaranteed.

The third case considers the situation that is our focus; model $g(V|X)$ is given as a parametrized probabilistic model. In this case, if $g(V|X)$ accurately represent the true distribution without overfitting, the output of the target prediction model is expected to be neutralized with respect not only to the training examples, but also to the unseen examples; this is demonstrated in the following sections by experiments.

The definition of η-neutrality contains all of the above cases, but we specifically consider only the third case, the parametric model.

3. Applications of Maximum Likelihood Estimation with η-Neutrality

In this section, we demonstrate two applications of maximum likelihood estimation with a guarantee of empirical η-neutrality: η-neutral logistic regression and η-neutral linear regression.

3.1 η-Neutral Logistic Regression

We incorporate our neutrality definition into logistic regression. In logistic regression, the domain of the input variable is $\mathcal{X} = \mathbb{R}^d$, and the domain of the target variable is binary, $\mathcal{Y} = \{0, 1\}$. Letting $\boldsymbol{\theta} \in \mathbb{R}^d$ be the model parameter, the target prediction model for logistic regression is

$$f(y|\boldsymbol{x};\boldsymbol{\theta}) = \sigma(\boldsymbol{\theta}^T\boldsymbol{x})^y(1-\sigma(\boldsymbol{\theta}^T\boldsymbol{x}))^{1-y}, \tag{7}$$

where $\sigma(a)$ is the logistic sigmoid function.

Letting Eq. (7) be the target prediction model, the log-likelihood is given by Eq. (1), and then the problem of η-neutral logistic regression is

$$\text{minimize } L(\boldsymbol{\theta}) \quad \text{subject to } N(y, v;\boldsymbol{\theta}) \le 0, \quad \forall v, y.$$

Note that the viewpoint prediction model $g(v|\boldsymbol{x})$ can be any probabilistic model.

We consider the optimization of η-neutral logistic regression. The gradient and Hessian matrix of $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ are, respectively,

$$\nabla L(\boldsymbol{\theta}) = \sum_{i=1}^{N}\left(\sigma(\boldsymbol{\theta}^T\boldsymbol{x}_i) - y_i\right)\boldsymbol{x}_i,$$
$$\nabla^2 L(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sigma(\boldsymbol{\theta}^T\boldsymbol{x}_i)(1-\sigma(\boldsymbol{\theta}^T\boldsymbol{x}_i))\boldsymbol{x}_i\boldsymbol{x}_i^T.$$

Due to the nature of the logistic sigmoid function, the Hessian matrix is positive semidefinite. Hence, the log-likelihood function is convex.

Next, we examine the convexity of the constraints associated with the η-neutrality condition. Since $N(y, v;\boldsymbol{\theta})$ is a linear combination of f, the convexity of f is investigated. The gradient of f with respect to the parameter $\boldsymbol{\theta}$ is

$$\begin{aligned}\nabla f(y, \boldsymbol{x};\boldsymbol{\theta}) &= \nabla \exp(\ln f(y|\boldsymbol{x};\boldsymbol{\theta})) \\ &= \left(y - \sigma(\boldsymbol{\theta}^T\boldsymbol{x})\right) f(y|\boldsymbol{x};\boldsymbol{\theta})\boldsymbol{x}.\end{aligned}$$

The Hessian is similarly obtained as

$$\nabla^2 f(y|\boldsymbol{x};\boldsymbol{\theta}) = \alpha(\boldsymbol{x}, y, \boldsymbol{\theta}) f(y|\boldsymbol{x};\boldsymbol{\theta})\boldsymbol{x}\boldsymbol{x}^T,$$

where $\alpha(\boldsymbol{x}, y, \boldsymbol{\theta}) = 2\sigma(\boldsymbol{\theta}^T\boldsymbol{x})^2 + y^2 - (2y+1)\sigma(\boldsymbol{\theta}^T\boldsymbol{x})$. Since $\alpha(\boldsymbol{x}, y, \boldsymbol{\theta}) \in \mathbb{R}$ can be negative, the Hessian is not positive definite, and f is nonconvex with respect to $\boldsymbol{\theta}$. Thus, unfortunately, the neutrality condition in logistic regression is nonconvex, regardless of the choice of $g(v|\boldsymbol{x})$.

In our experiments with η-neutral logistic regression, we used the nonlinear optimization package, Ipopt, that provides the implementation of the primal-dual interior point method [15]. As the initial point of the primal-dual interior point method to solve the optimization problem of η-neutral logistic regression, we use the optimal point of logistic regression without neutralization. Although the constraint is nonconvex, we show by experiments that η-neutrality can be achieved without sacrificing too much of the accuracy of the prediction in Sect. 5. This nonconvexity arises in part from the nonconvexity of the probability distribution. Further research on convexifying the neutrality constraint is left as an area of future work.

3.2 η-Neutral Linear Regression

We now consider η-neutral linear regression and demonstrate that maximum likelihood estimation with η-neutrality can work with continuous viewpoint variables. In linear regression, the domain of the target variable is $\mathcal{Y} = \mathbb{R}$, and the

input domain is $\mathcal{X} = \mathbb{R}^d$. The target prediction function is given by

$$f(y|\boldsymbol{x};\boldsymbol{w},\beta) = \frac{\beta}{\sqrt{2\pi}} \exp\left[-\frac{\beta(\boldsymbol{w}^T\boldsymbol{x}-y)^2}{2}\right], \tag{8}$$

where $\boldsymbol{w}$ denotes the regression coefficient for the target variable and β denotes the parameter representing the inversed variance of the prediction error of the target variable. The linear regression problem is solved by the minimization of the negative log-likelihood, as given by Eq. (1).

The domain of the viewpoint is $\mathcal{V} = \mathbb{R}$. Similarly, we assume the viewpoint prediction model is

$$g(v|\boldsymbol{x};\boldsymbol{w}_v,\beta_v) = \frac{\beta_v}{\sqrt{2\pi}} \exp\left[-\frac{\beta(\boldsymbol{w}_v^T\boldsymbol{x}-v)^2}{2}\right], \tag{9}$$

where $\boldsymbol{w}_v$ denotes the regression coefficient for the viewpoint variable and β_v denotes the parameter representing the inversed variance of the prediction error of the viewpoint variable.

Predictions of the target random variable Y and the viewpoint random variable V are obtained, respectively, by

$$\hat{y} = \operatorname*{argmax}_y f(y|\boldsymbol{x};\boldsymbol{w},\beta), \quad \hat{v} = \operatorname*{argmax}_v g(v|\boldsymbol{x};\boldsymbol{w}_v,\beta_v).$$

Then, η-neutral linear regression is formulated as an optimization problem with the same constraints as in Eq. (6):

$$\begin{aligned} &\text{minimize } \frac{1}{2}\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} \\ &\text{subject to } \max_{\boldsymbol{x}\in\mathcal{D}}\{N(\boldsymbol{w}^T\boldsymbol{x},\boldsymbol{w}_v^T\boldsymbol{x};\boldsymbol{w},\beta)\} \le 0, \end{aligned}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \ldots, \boldsymbol{x}_N^T)^T$ is the design matrix and $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)^T$ is the vector of target values.

As in the case with η-neutral logistic regression, we investigate the convexity of the neutrality constraint given models f and g by investigating the convexity of f. The gradient and Hessian matrix of f are, respectively,

$$\begin{aligned} &\nabla_w f(\boldsymbol{w}^T\boldsymbol{x}'|\boldsymbol{x};\boldsymbol{w},\beta) \\ =&\nabla_w \exp(\ln f(\boldsymbol{w}^T\boldsymbol{x}'|\boldsymbol{x};\boldsymbol{w},\beta)) \\ =&-\beta(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{x}')f(\boldsymbol{w}^T\boldsymbol{x}'|\boldsymbol{x};\boldsymbol{w},\beta)(\boldsymbol{x}-\boldsymbol{x}'), \\ &\nabla_w^2 f(\boldsymbol{w}^T\boldsymbol{x}'|\boldsymbol{x};\boldsymbol{w},\beta) \\ =&\alpha(\boldsymbol{x},\boldsymbol{x}',\boldsymbol{w},\beta)\beta f(\boldsymbol{w}^T\boldsymbol{x}'|\boldsymbol{x};\boldsymbol{w},\beta)(\boldsymbol{x}-\boldsymbol{x}')(\boldsymbol{x}-\boldsymbol{x}')^T, \end{aligned}$$

where $\alpha(\boldsymbol{x},\boldsymbol{x}',\boldsymbol{w},\beta) = \beta(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{x}')^2 - 1$. Since, depending on $\boldsymbol{w}$, $f(\boldsymbol{w}^T\boldsymbol{x}'|\boldsymbol{x};\boldsymbol{w},\beta) \ge 0$ and $\alpha(\boldsymbol{x},\boldsymbol{x}',\boldsymbol{w},\beta) \in \mathbb{R}$ can take negative values, the Hessian is not positive definite. Hence, unfortunately, f is not convex with respect to $\boldsymbol{w}$. For this non-convex constraint optimization, we again use the primal-dual interior point method of Ipopt in our experiments [15].

4. Comparison of Neutrality Measures

One of the largest difference between η-neutrality and the CV score [3] or statistical parity [5] or the prejudice index [4] is its situation; while the neutralization of η-neutrality is based on the model of the viewpoint variable, that of others is based on the value of the viewpoint variable. In order to discuss the difference of these neutrality measures, we assume the examples $\mathcal{D}$ contains the viewpoint values in the subsequent subsections.

The CV score and the prejudice index are defined as quantities that measures neutrality, whereas η-neutrality and statistical parity are defined as conditions required for the prediction model to be neutral. More precisely, for example, the prediction model is said to be η-neutral only if the η-neutrality of the prediction model is upper-bounded by η for all y and v. We employ the upper bound of η-neutrality and statistical parity as the neutrality measure for η-neutrality and statistical parity, respectively.

For comprehensive discussion of the comparison of the neutrality measures, we define the *neutrality factor*. We introduced that can universally represent all the neutrality measures. The neutrality factor denotes the neutrality with respect to a specific pair of target value y and viewpoint value v:

Definition 3 (Neutrality factor)**.** *Let X and Y be the input and target random variables, respectively. Let V denote the viewpoint random variable. Then, the* neutrality factor *with respect to target $y \in \mathcal{Y}$ and viewpoint $v \in \mathcal{V}$ is defined by*

$$\nu(y,v) = \frac{\tilde{\Pr}(y,v)}{\tilde{\Pr}(y)\,\tilde{\Pr}(v)}.$$

From the definition of η-neutrality, we can say that η-neutrality evaluates the maximum value of the neutrality factors with respect to y and v. In subsequent subsections, we represent the CV score, statistical parity and the prejudice index by using the neutrality factor. Furthermore, we clarify the relationship between these neutrality measures.

4.1 Comparison of η-Neutrality, CV Score and Statistical Parity

In this subsection, we first show that the CV score is a variant measure of statistical parity. Then, we derive the relation between η-neutrality and statistical parity. In what follows, we assume Y is the discrete target variable and V is the binary viewpoint variable.

Let $y \in \{y_+, y_-\}$ and $v \in \{v_+, v_-\}$ be the binary target variable and the binary viewpoint variable, respectively. The CV score [3] with respect to the given example set $\mathcal{D}$ is defined by the difference of the conditional probability:

$$\mathrm{CV}(\mathcal{D}) = \tilde{\Pr}(y_+|v_+) - \tilde{\Pr}(y_+|v_-), \tag{10}$$

where $\tilde{\Pr}(Y|V)$ is empirically evaluated with the given example set $\mathcal{D}$. We can assume $\tilde{\Pr}(y_+|v_+) \ge \tilde{\Pr}(y_+|v_-)$ without loss of generality. If the CV score equals to zero, the classification is empirically neutral with respect to the given example $\mathcal{D}$.

Statistical parity [5] defines the neutrality considering total variation of the two probabilistic distributions of target y, $P(y)$ and $Q(y)$,

$$D_{tv}(P, Q) = \frac{1}{2}\sum_{y\in\mathcal{Y}} |P(y) - Q(y)|. \tag{11}$$

Given $\epsilon \geq 0$ as a neutrality parameter, we say ϵ-statistical parity holds with respect to a given example set $\mathcal{D}$ if

$$D_{tv}(\tilde{\Pr}(Y|v_+), \tilde{\Pr}(Y|v_-)) \leq \epsilon.$$

First, we show that the CV score is a variant measure of statistical parity.

Lemma 1. *Let Y and V be the binary target variable and the binary viewpoint variable, respectively. For any $\epsilon \geq 0$ and example set $\mathcal{D}$, $\mathrm{CV}(\mathcal{D}) \leq \epsilon$ if and only if ϵ-statistical parity with respect to $\mathcal{D}$ holds.*

Proof. By the definition of statistical parity, if ϵ-statistical parity holds

$$D_{tv}(\tilde{\Pr}(Y|v_+), \tilde{\Pr}(Y|v_-)) \leq \epsilon.$$

By the definition of the probability, $\tilde{\Pr}(y_+|v) + \tilde{\Pr}(y_-|v) = 1\ \forall v \in \mathcal{V}$ and we have

$$\begin{aligned}
&\tilde{\Pr}(y_+|v_+) - \tilde{\Pr}(y_+|v_-)\\
=&(1 - \tilde{\Pr}(y_-|v_+)) - (1 - \tilde{\Pr}(y_-|v_-))\\
=&\tilde{\Pr}(y_-|v_-) - \tilde{\Pr}(y_-|v_+).
\end{aligned} \tag{12}$$

By substituting Eq. (12) into Eq. (10), we have

$$\begin{aligned}
&\mathrm{CV}(\mathcal{D})\\
=&\tilde{\Pr}(y_+|v_+) - \tilde{\Pr}(y_+|v_-)\\
=&\frac{1}{2}(\tilde{\Pr}(y_+|v_+) - \tilde{\Pr}(y_+|v_-) + \tilde{\Pr}(y_-|v_-) - \tilde{\Pr}(y_-|v_+))
\end{aligned}$$

We can assume $\tilde{\Pr}(y_+|v_+) - \tilde{\Pr}(y_+|v_-) \geq 0$ and $\tilde{\Pr}(y_-|v_-) - \tilde{\Pr}(y_-|v_+) \geq 0$ without loss of generality. Hence, we have

$$\begin{aligned}
\mathrm{CV}(\mathcal{D}) =&\frac{1}{2}\sum_{y\in\mathcal{Y}} |\tilde{\Pr}(y|v_+) - \tilde{\Pr}(y|v_-)|\\
=&D_{tv}(\tilde{\Pr}(Y|v_+), \tilde{\Pr}(Y|v_-)).
\end{aligned}$$

□

As proved by Lemma 1, the statistical parity with the binary target variable can be interpreted as the CV score.

Next, we provide the relation between η-neutrality and statistical parity. The following theorem shows that η-statistical parity with respect to the given example set $\mathcal{D}$ holds if η-neutrality holds.

Theorem 3. *Let X and Y be the input variable and the discrete target random variable, respectively. Let V denote the binary viewpoint random variable. If the probability $\Pr(X, Y, V)$ is empirically η-neutral, then Y is η-statistical parity with respect to V.*

In order to prove Theorem 3, we use the following lemma that shows another representation of the total variation in statistical parity by using the neutrality factors.

Lemma 2. *Let $D_{tv}(P, Q)$ be total variation between P and Q with respect to Y defined Eq. (11). Then,*

$$D_{tv}(\tilde{\Pr}(Y|v_+), \tilde{\Pr}(Y|v_-)) = E_Y\left[\max_{v\in\{v_+,v_-\}} \nu(y, v)\right] - 1.$$

The proof of Lemma 2 is shown in the Appendix. As proved by Lemma 2, statistical parity is the expectation of the maximum value with respect to v of the neutrality factors. By using Lemma 2, we prove Theorem 3.

Proof of Theorem 3. If η-neutrality holds, $\nu(y, v) \leq 1 + \eta\ \forall y \in \mathcal{Y}, v \in \mathcal{V}$. Then, we have

$$D_{tv}(\tilde{\Pr}(Y|v_+), \tilde{\Pr}(Y|v_-)) \leq \mathrm{E}_Y[1 + \eta] - 1 \leq \eta.$$

□

As proved by the Theorem 3, statistical parity holds if η-neutrality holds. We can immediately show that the CV score is bounded by a certain function of η if η-neutrality holds by using Theorem 3 and Lemma 1.

4.2 Comparison of η-Neutrality and Prejudice Index

In this subsection, we compare our η-neutrality with the prejudice index [4]. The prejudice index is defined as the mutual information of the target random variable Y and the viewpoint random variable V:

$$\mathrm{PI} = I(Y; V) = \mathrm{E}_{Y,V}[\ln \nu(y, v)],$$

where $I(X; Y)$ is the mutual information of the target Y and the viewpoint V.

While the prejudice index is the expectation of the logarithm of the neutrality factors $\nu(y, v)$, the neutrality parameter η of η-neutrality denotes the upper bound of the neutrality factor $\nu(y, v)$. This indicates that prejudice index can be upper bounded with the neutrality parameter η if η-neutrality holds. Following proposition provides this indication.

Proposition 1. *Let X and Y be the input and target random variables, respectively. Let V denote the viewpoint random variable. If the probability $\Pr(X, Y, V)$ is empirically η-neutral with respect to given $\mathcal{D}$ and $\eta \geq 0$, then*

$$I(V; Y) \leq \ln(1 + \eta).$$

Proof. From empirical η-neutrality of the probability $\Pr(X, Y, V)$, we have

$$\forall v \in \mathcal{V}, y \in \mathcal{Y}, \quad \frac{\tilde{\Pr}(v, y)}{\tilde{\Pr}(v)\,\tilde{\Pr}(y)} \leq 1 + \eta.$$

Since natural logarithm is a monotonically increasing function, we have

Table 2 Summary of neutrality measures

Aggregation	$\nu(y, v)$	$\ln \nu(y, v)$
Maximum w.r.t y and v	η-neutrality	equivalent to η-neutrality
Maximum w.r.t v and expectation w.r.t. y	CV-score, statistical parity	-
Expectation w.r.t y and v	-	prejudice index

Table 3 Specification of datasets for classification tasks. #Inst., #Attr., "Viewpoint" and "Target" denote the number of example sets, the number of attributes, the attribute used as the target variable and the attribute used as the viewpoint variable, respectively. $\#y_+$ and $\#v_+$ represent the number of positive target and viewpoint values, respectively. The prediction accuracy of logistic regression for the target variable (Acc (y)) and viewpoint variable (Acc (v)) are also shown.

dataset	#Inst.	#Attr.	Viewpoint	Target	$\#y_+$	$\#v_+$	Acc (y)	Acc (v)
Adult [18]	16281	13	gender	income	3846 (23.6%)	10860 (66.7%)	0.850	0.842
Dutch Census [19]	60420	10	gender	income	31657 (52.4%)	30273 (50.1%)	0.819	0.665
Bank Marketing [18]	45211	17	loan	term deposit	5289 (11.7%)	7244 (16.0%)	0.900	0.839
Credit Approval [18]	690	15	A1	A16	307 (44.5%)	480 (69.6%)	0.875	0.676
German Credit Data [18]	1000	20	foreign worker	credit risk	300 (30.0%)	37 (3.7%)	0.757	0.961

$$\forall v \in \mathcal{V}, y \in \mathcal{Y}, \quad \ln \frac{\tilde{\Pr}(v, y)}{\tilde{\Pr}(v)\tilde{\Pr}(y)} \leq \ln(1 + \eta). \tag{13}$$

Expectation of Eq. (13) with respect to Y and V derives as follows:

$$\mathrm{E}_{Y,V}\left[\ln \frac{\tilde{\Pr}(V, Y)}{\tilde{\Pr}(V)\tilde{\Pr}(Y)}\right] = I(v; y) \leq \ln(1 + \eta).$$

□

As proved by the Proposition 1, the prejudice index is upper bounded by $\ln(1 + \eta)$ if η-neutrality holds.

4.3 Summary of Comparisons

Table 2 shows the summary of the neutrality measures. By definition of η-neutrality, η-neutrality is the maximum value of the neutrality factors. Due to monotonicity of the logarithm function, η-neutrality is equivalent to the maximum value of logarithm of the neutrality factors (Table 2, line 1). Statistical parity can be represented as the expectation of the maximum value with respect to v of the neutrality factors as indicated by Lemma 2. Similarly, as indicated by Lemma 1, the CV score is equivalent to statistical parity with binary target variables (Table 2, left of line 2). The prejudice index is defined as the expectation of logarithm of the neutrality factor (Table 2, left of line 2).

As indicated by Theorem 3, statistical parity and the CV score can be upper bounded by η-neutrality. Moreover, as indicated by Theorem 1, the prejudice index can be upper bounded by η-neutrality.

As shown in Table 2, all of the neutrality measures can be represented with the neutrality factors. The difference of these neutrality measures is only in the way of aggregation.

The prejudice index is defined by the mutual information which represents statistical dependency between the target random variable and the viewpoint random variable. Thus, the neutrality measures are closely connected to the measures of statistical dependency [16], [17].

5. Experiments

5.1 Classification

Settings. In order to examine and compare the classification performance and the neutralization effect of η-neutral logistic regression with other methods, we performed experiments on five real data sets specified in Table 3. In the table, #Inst. and #Attr. denote the number of examples and the number of the attributes, respectively; "Viewpoint" and "Target" denote the attribute used as the target variable and the viewpoint variable, respectively. Table 3 also shows the number of examples with the target variable ($\#y_+$) and the viewpoint variable ($\#v_+$). In addition, the table shows the prediction accuracy of the logistic regression without neutralization with respect to the target variable (Acc(y)) and the viewpoint variable (Acc(v)).

We compared the following methods: logistic regression (LR, no neutrality guarantee), logistic regression that learns without using the values of viewpoint (LRns), the Naive Bayes classifier (NB, no neutrality guarantee), the Naive Bayes classifier that learns without the values of viewpoint (NBns), CV2NB [3], logistic regression that uses the PR [13], and η-neutral logistic regression with viewpoint neutrality (ηLR, proposal). In the PR method, the regularizer parameter λ, which balances the loss minimization and neutralization, was varied as $\lambda \in \{0, 5, 10, 15, 20, 30\}$. The neutrality parameter η, which determines the degree of neutrality, was varied as $\eta \in \{0.00, 0.01, \ldots, 0.40\}$. All dataset attributes were discretized by the same procedure described in [3] and coded by 1-of-K representation for LR, LRns, PR and ηLR.

As neutrality indices of prediction models, normalized prejudice index (NPI) and $\hat{\eta}$ are introduced. NPI is defined as the normalized mutual information of the target random variable Y and the viewpoint random variable V, normalized by the entropy of Y and V [4]:

$$\mathrm{NPI} = \frac{I(X; Y)}{\sqrt{H(Y)H(V)}},$$

Table 4 Summary of the treatment of the viewpoint random variables in two settings.

case	method	learning of $f(y\|x)$	neutrality guarantee	neutrality measure
Case 1	others	$\boldsymbol{x}, v$	v	$f(y\|\boldsymbol{x};\boldsymbol{\theta}), v$
	ours	$\boldsymbol{x}, v$	$g(v\|\boldsymbol{x})$	$f(y\|\boldsymbol{x};\boldsymbol{\theta}), v$
Case 2	others	$\boldsymbol{x}, \hat{v}$	$\hat{v}$	$f(y\|\boldsymbol{x};\boldsymbol{\theta}), v$
	ours	$\boldsymbol{x}, \hat{v}$	$g(v\|\boldsymbol{x})$	$f(y\|\boldsymbol{x};\boldsymbol{\theta}), v$

where $I(X;Y)$ is the mutual information of target Y and viewpoint V, $I(X;Y)/H(Y)$ is the ratio of information of V used for predicting Y, and $I(X;Y)/H(V)$ is the ratio of information that is exposed if a value of Y is known. Thus NPI can be interpreted as the geometrical mean of these two ratios. The range of this NPI is $[0, 1]$.

The neutrality measure $\hat{\eta}$ is defined as

$$\hat{\eta} = \max_{y\in\mathcal{Y},v\in\mathcal{V}} \frac{\tilde{\Pr}(v,y)}{\tilde{\Pr}(v)\,\tilde{\Pr}(y)} - 1,$$

where $\hat{\eta}$ can be interpreted as the degree of the dependency of y and v with which the largest dependency occurs. If Y and V are mutually independent, $\hat{\eta} = 0$. If the neutrality measure with respect to a target prediction model is $\hat{\eta}$, it means the model of Eq. (4) is empirically $\hat{\eta}$-neutral with respect to the given examples.

We compared the three measures: accuracy, normalized prejudice index (NPI), and $\hat{\eta}$ of η-neutrality. These indices were evaluated with five-fold cross validation and the average values of ten different folds are shown in the plots.

The values used for the learning of $f(y|x)$, the guarantee of neutrality, and the measurement of neutrality are summarized in Table 4. For the guarantee of neutrality, we consider the following two cases.

Case 1 assumes that the values of the viewpoint random variable are provided in examples. In this case, our method performs neutralization with respect to the model of the viewpoint learned from the examples, whereas other methods perform neutralization with respect to the actual viewpoint values provided.

Case 2 assumes that the values of the viewpoint are not provided. Instead, the model of the viewpoint variable, $g(v|\boldsymbol{x})$, is provided. In this case, our method again learns the model of the target without using values of the viewpoint and performs neutralization with respect to the given model g. Other methods need the values of the viewpoint, so these are estimated as $\hat{v} = \mathrm{argmax}_v g(v|\boldsymbol{x})$. Other methods then learn the model of the target with $(x, \hat{v})$, and neutralization is performed with respect to $\hat{v}$.

As a measurement of neutrality, all methods used the true viewpoint value v in both cases.

Results. Figure 2 shows the experimental results. In the graphs, the best result is at the left top. Comparing the results of NB and NBns in Adult, Dutch Census and Bank Marketing, we can see that the improvement of neutrality by elimination of the viewpoint variable is limited in both cases. The same applies to LR and LRns.

In both cases, CV2NB achieves better neutrality than NBns in terms of both NPI and $\hat{\eta}$ in Adult and Dutch Census. In addition, the decrease in the accuracy of the prediction is less than 1% in the Adult dataset and 5% in the Dutch Census. On the other hand, neutralization by CV2NB dose not work well in Bank Marketing and German Credit Data; the neutralization level of CV2NB is worse than NBns. As shown in Table 3, the number of positive viewpoint of these datasets is fewer than the negative viewpoint values, in comparison with the other datasets. The degradation of performance of CV2NB in Bank Marketing and German Credit Data can be caused by such imbalanced viewpoint labels.

In both cases, PR successfully balances the NPI or the $\hat{\eta}$ and the accuracy for Adult and Dutch Census datasets, but dominated by ηLR. In order to neutralize the target prediction model, PR adds non-convex NPI term to the objective function. Due to the non-convexity of the objective function, both the accuracy and the neutralization level of the prediction model can be worsen.

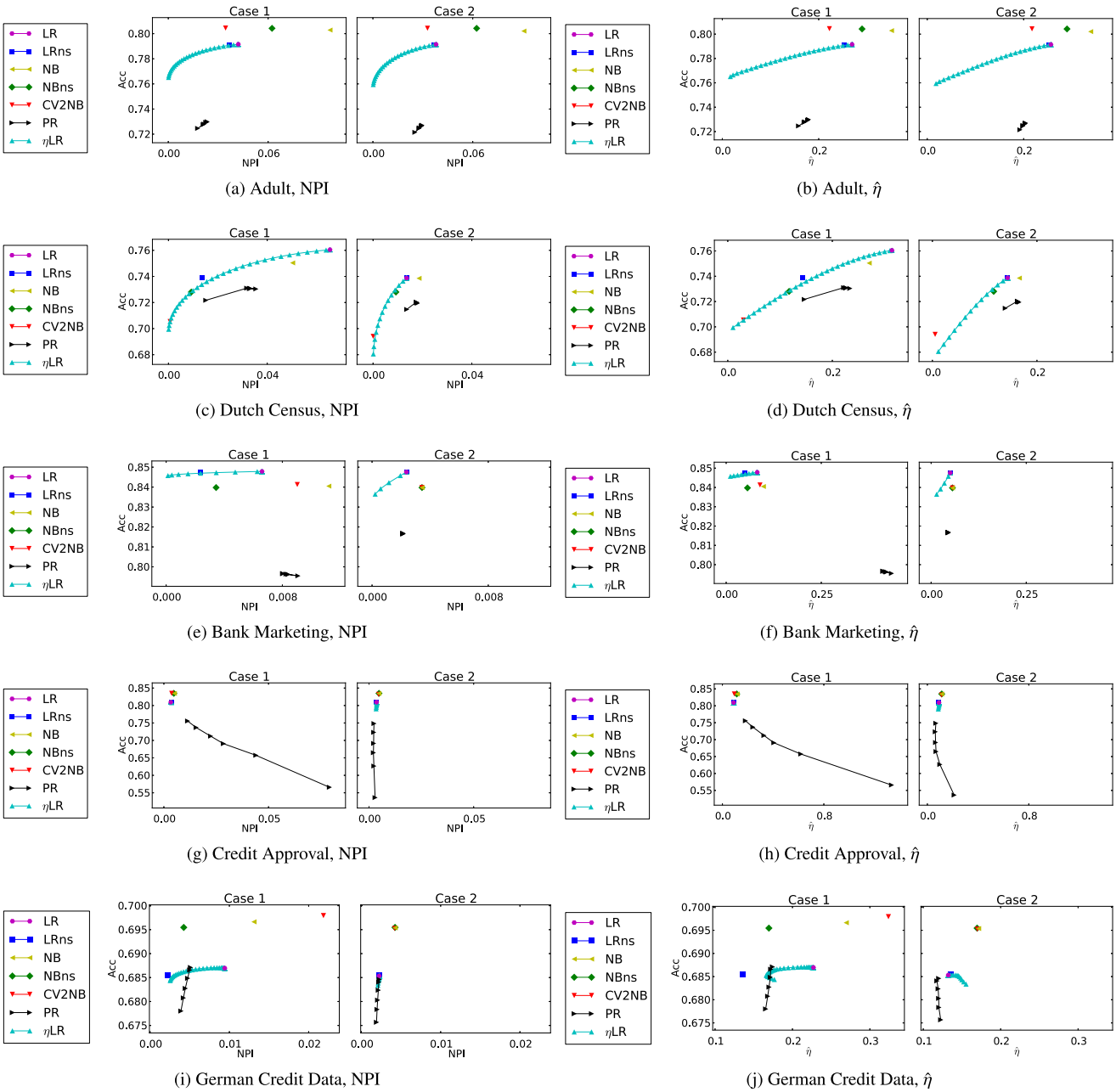
In both cases, our η-neutral logistic regression successfully balances neutralization and accuracy of the predication by changing η in Adult, Dutch Census and Bank Marketing. Particularly, in Bank Marketing, even though the neutralization level of the other methods is almost the same as its baseline (LR or NB), our ηLR can achieve the prediction model with low neutralization level. Furthermore, the decrease in the accuracy of the prediction was at most 5% in these datasets, even after strong neutralization with small η. Whereas both of the neutralization level and the accuracy of PR can be worsen due to the non-convexity of the neutrality term in the objective function, the neutralization by constraints guarantees the neutralization level even if the constraints is non-convex. Thus, ηLR empirically works well even if its constraints is non-convex.

In German Credit Data, the neutralization level of ηLR is lower than LRns in both cases. It is noteworthy that the neutralization level of ηLR is even lower than LR in Case 2. This was again due to imbalanced viewpoint labels of the dataset. The given model of the viewpoint is trained so that it ignores minor viewpoint label. Hence, due to the overfitting of the model learned by ηLR with such model of the viewpoint to major viewpoint label, the neutralization level of ηLR may be lower than LR.

In Credit Approval, all neutralization technique did not work well in both cases. From Table 3, the number of the example set of these datasets are up to seven hundred. This result can indicate that estimation of the neutrality measures for test dataset need sufficiently number of the example set.

5.2 Regression

Settings. In order to investigate the behaviors of neutralization in linear regression, we performed experiments of η-neutral linear regression on three real datasets specified in Table 5. As with the specification of the dataset for the classification, the table shows #Inst., #Attr., "Viewpoint" and "Target". In addition, the table also provides "Corr", the correlation coefficient between the target variable and the



(a) Adult, NPI (b) Adult, $\hat{\eta}$

(c) Dutch Census, NPI (d) Dutch Census, $\hat{\eta}$

(e) Bank Marketing, NPI (f) Bank Marketing, $\hat{\eta}$

(g) Credit Approval, NPI (h) Credit Approval, $\hat{\eta}$

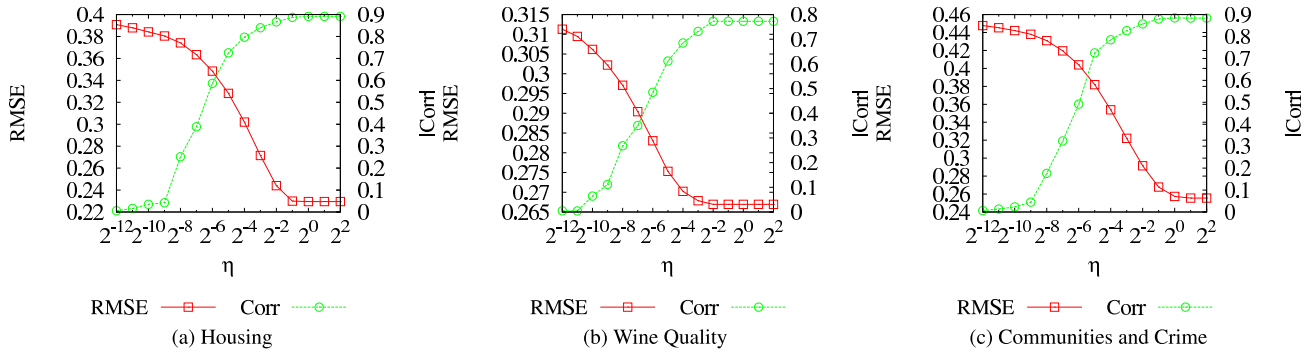
(i) German Credit Data, NPI (j) German Credit Data, $\hat{\eta}$

Fig. 2 Accuracy vs. neutrality measure. Each subplot displays the result of Case 1 (left) and the result of Case 2 (right) corresponding to the datasets and the neutrality measure ($\hat{\eta}$ or NPI).

Table 5 Specification of datasets for regression task. #Inst., #Attr., "Viewpoint" and "Target" denote the number of example sets, the number of attributes, the attribute used as the target variable and the attribute used as the viewpoint variable, respectively. "Corr" represents the correlation coefficient between the target variable and the viewpoint variable.

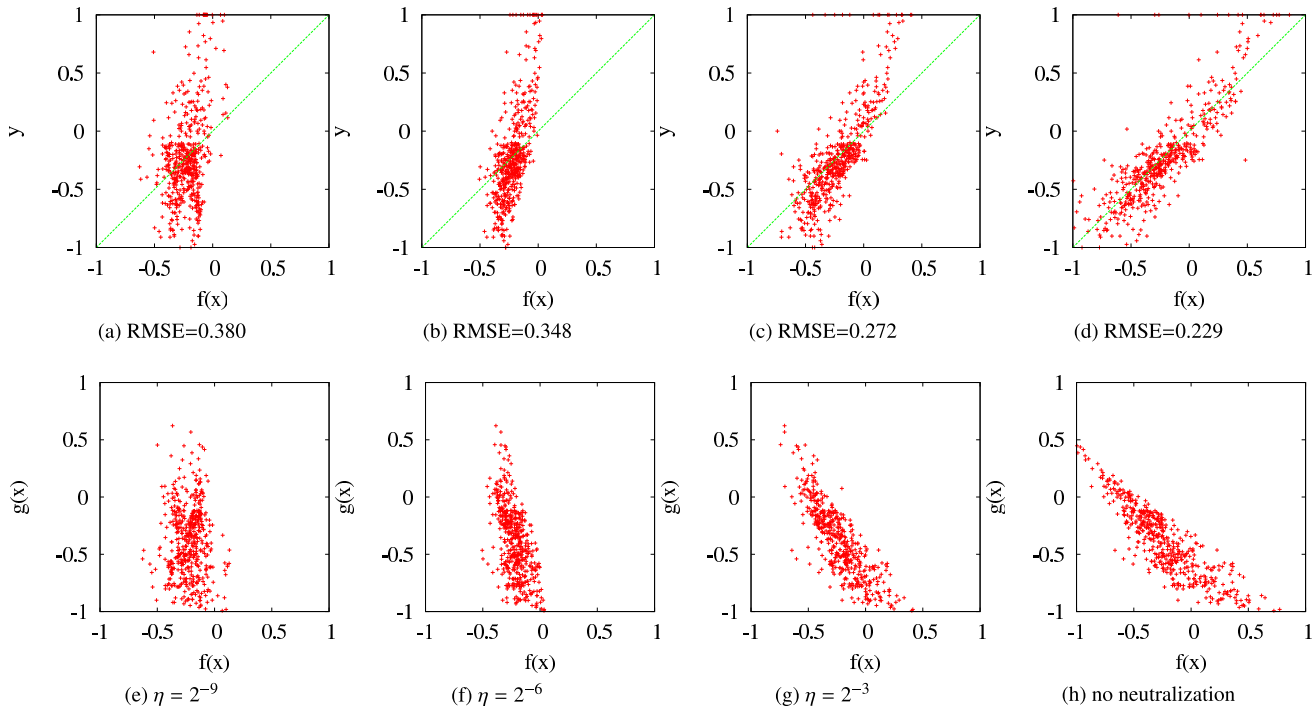
dataset	#Inst.	#Attr.	Viewpoint	Target	Corr
Housing [18]	506	14	LSTAT	MEDV	-0.738
Wine Quality (Red) [18]	1599	12	alcohol	quality	0.476
Communities and Crime [18]	1994	123	PctKids2Par	ViolentCrimesPerPop	-0.738

viewpoint variable. We chose the viewpoint variables for each dataset as the attribute of which the correlation coefficient with respect to the target variable maximizes. All the attributes, the target variable and the viewpoint variable were scaled into the range $[-1, 1]$. Letting the regression parameters of the target f and viewpoint g be $\boldsymbol{w}$ and $\boldsymbol{w}_v$, respectively, the predicted values were $\hat{y} = \boldsymbol{w}^T\boldsymbol{x}$ and $\hat{v} = \boldsymbol{w}_v^T\boldsymbol{x}$. The neutrality parameter η was varied as $\eta \in \{2^{-12}, 2^{-11}, \ldots, 2^2\}$.



(a) Housing (b) Wine Quality (c) Communities and Crime

Fig. 3 The plots show RMSE and the absolute value of the correlation coefficient between the predicted target value and the viewpoint value corresponding to the neutrality parameter η.



(a) RMSE=0.380 (b) RMSE=0.348 (c) RMSE=0.272 (d) RMSE=0.229

(e) $\eta = 2^{-9}$ (f) $\eta = 2^{-6}$ (g) $\eta = 2^{-3}$ (h) no neutralization

Fig. 4 Scatter plots with respect to Housing dataset. Top row: scatter plots of target prediction value $\hat{y}$ and true target value y. Bottom row: scatter plots of target prediction value $\hat{y}$ and viewpoint prediction value $\hat{v}$. Correlation in the $\hat{y} - \hat{v}$ plots means that the neutralization level of the regression model is low.

The accuracy of the prediction was measured by root-mean-square error (RMSE); $\hat{\eta}$ and the correlation coefficient between the target variable and the viewpoint variable were used as the measure of neutrality.

Results. Figure 3 shows RMSE and the absolute value of the correlation coefficient between the predicted target value $\hat{y}$ and the viewpoint value v corresponding to the neutrality parameter η. For all datasets, the plots explicitly show that the correlation coefficient becomes lower and RMSE becomes higher as η decreases. This results show that our η-neutral linear regression with low neutrality parameter η can obtain the neutral regression model in the sense of the correlation coefficient. Furthermore, this results indicate that our η-neutral linear regression can use η to successfully control the neutralization level of the regression model.

Figure 4 shows the scatter plots of $(\hat{y}, y)$ (the top row) and $(\hat{y}, \hat{v})$ (the bottom row) with $\eta \in \{2^{-9}, 2^{-6}, 2^{-3}\}$ on the Housing dataset. From left to right, the neutrality parameter η was varied as $\eta \in \{2^{-9}, 2^{-6}, 2^{-3}\}$. The most right figures show the results without neutralization. The $(\hat{y}, \hat{v})$ plot represents the prediction accuracy of the regression model. When the model achieves a better RMSE, the points in the $(\hat{y}, y)$ plot concentrate more along the diagonal line. At the same time, the $(\hat{y}, \hat{v})$ plot represents neutrality. If the neutrality is low, correlation between $\hat{y}$ and $\hat{v}$ appears in the $(\hat{y}, \hat{v})$ plot.

In Fig. 4 (h), a strong negative correlation between $\hat{y}$ and $\hat{v}$ can be found. Thus, this regression model has a low neutrality if no neutralization is performed. In Fig. 4, the level of neutralization increases from right to left. The plots

show that the dependency of $\hat{y}$ on $\hat{v}$ becomes weaker as η decreases. In Fig. 4 (e), we can see that the regression model of the target value that has high neutrality outputs almost constant values; such regression is useless even if the model is well neutralized. Thus, selection of η is important to obtain a neutralized regression model with high accuracy.

6. Conclusion

In this paper, we propose a framework for using a maximum likelihood estimation for learning probabilistic models with neutralization. There are two key points in which our proposal is different from existing methods.

First, our method guarantees neutrality of the target prediction model with respect to a given viewpoint prediction model. Due to this model-based neutralization, our method allows neutralization of target prediction models with respect to viewpoints arbitrarily defined by users, as long as the viewpoint prediction model is provided in the form of a probabilistic distribution.

Second, our neutrality measure, η-neutrality, is based on the principle that the model should guarantee neutrality with respect to every combination of target and viewpoint value that appears in the dataset.

In order to clarify the relationship between the neutrality measures, we define the neutrality factor. Then, we showed that all of the neutrality measures are represented with aggregation of the neutrality factors. We also show that η-neutrality can upper bound all of the other neutrality measures.

Experimental results show that our method with model-based neutralization achieves neutralization even when only a model of the viewpoint is provided. In addition, it balances the accuracy of the target prediction with the neutrality. As discussed in Sect. 3.1 and Sect. 3.2, likelihood maximization with the η-neutrality constraint is nonconvex optimization; this is due the nonconvexity of the constraint function. As an area of future work, we intend to find a way to convexify the constraints induced by the neutrality condition.

The privacy problem is strongly related to the neutrality problem. The difference between the privacy problem and the neutrality problem causes from the treatments of the sensitive information. The sensitive information in the privacy problem is individuals' information that they want not to be published. On the other hand, the sensitive information in the neutrality problem, which is equivalent to the viewpoint random variable, is individuals' information that they want not to be made decisions depending on this information. Following the treatment of the sensitive information, we can define the adversaries in the privacy problem as entities that can predict the sensitive information. The adversaries in the neutrality problem can be defined in the same manner as entities that make decisions depending on the sensitive information.

Acknowledgements

The work is supported by JST CREST program of Advanced Core Technologies for Big Data Integration, JSPS KAKENHI 12913388, and 25540094.

References

[1] D. Boyd, "Privacy and publicity in the context of big data," Keynote Talk of The 19th International Conference on World Wide Web, April 2010.
[2] E. Pariser, The Filter Bubble: What The Internet Is Hiding From You, Viking, London, 2011.
[3] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," Data Mining and Knowledge Discovery, vol.21, no.2, pp.277–292, Sept. 2010.
[4] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Part II, Lecture Notes in Computer Science, vol.7524, pp.35–50, Springer, 2012.
[5] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R.S. Zemel, "Fairness through awareness," Innovations in Theoretical Computer Science, pp.214–226, ACM, 2012.
[6] R.S. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," Proceedings of the 30th International Conference on Machine Learning (3), JMLR Proceedings, vol.28, pp.325–333, JMLR.org, 2013.
[7] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination aware decision tree learning," The 10th IEEE International Conference on Data Mining, pp.869–874, IEEE Computer Society, 2010.
[8] S. Ruggieri, D. Pedreschi, and F. Turini, "Dcube: discrimination discovery in databases," Proceedings of the ACM SIGMOD International Conference on Management of Data, pp.1127–1130, ACM, 2010.
[9] A. Romei and S. Ruggieri, "A multidisciplinary survey on discrimination analysis," The Knowledge Engineering Review, pp.1–57, 4 2013.
[10] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.560–568, ACM, 2008.
[11] I. Zliobaite, F. Kamiran, and T. Calders, "Handling conditional discrimination," The 11th IEEE International Conference on Data Mining, pp.992–1001, IEEE Computer Society, 2011.
[12] B.L. Thanh, S. Ruggieri, and F. Turini, "k-nn as an implementation of situation testing for discrimination discovery and prevention," Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.502–510, ACM, 2011.
[13] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Enhancement of the neutrality in recommendation," Proceedings of the 2nd Workshop on Human Decision Making in Recommender Systems, CEUR Workshop Proceedings, vol.893, pp.8–14, CEUR-WS.org, 2012.
[14] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Efficiency improvement of neutrality-enhanced recommendation," Proceedings of the 3rd Workshop on Human Decision Making in Recommender Systems in conjunction with the 7th ACM Conference on Recommender Systems (RecSys 2013), CEUR Workshop Proceedings, vol.1050, pp.1–8, CEUR-WS.org, 2013.
[15] A. Wächter and L.T. Biegler, "On the implementation of an interior–point filter line-search algorithm for large-scale nonlinear programming," Mathematical Programming, vol.106, no.1, pp.25–57, 2006.
[16] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," Journal of Machine Learning Research, vol.3, pp.1415–1438, 2003.

[17] T. Suzuki, M. Sugiyama, T. Kanamori, and J. Sese, "Mutual information estimation reveals global associations between stimuli and biological processes," BMC Bioinformatics, vol.10, no.S-1, 2009.
[18] A. Frank and A. Asuncion, "UCI machine learning repository," 2010.
[19] Dutch Central Bureau for Statistics, "Volkstelling," 2001.

Appendix: Proof of Lemma 2

Proof. Let $\mathcal{Y}^+ = \{y \in \mathcal{Y} | \tilde{\Pr}(y|v_+) \geq \tilde{\Pr}(y|v_-)\}$ and let $\mathcal{Y}^- = \{y \in \mathcal{Y} | \tilde{\Pr}(y|v_+) \leq \tilde{\Pr}(y|v_-)\}$. Then, we have

$$
\begin{aligned}
& D_{tv}(\tilde{\Pr}(Y|v_+), \tilde{\Pr}(Y|v_-)) \\
&= \frac{1}{2} \sum_{y \in \mathcal{Y}} |\tilde{\Pr}(y|v_+) - \tilde{\Pr}(y|v_-)| \\
&= \frac{1}{2}\Big[\sum_{y \in \mathcal{Y}^+} (\tilde{\Pr}(y|v_+) - \tilde{\Pr}(y|v_-)) \\
&\qquad + \sum_{y \in \mathcal{Y}^-} (\tilde{\Pr}(y|v_-) - \tilde{\Pr}(y|v_+)) \Big] \\
&= \frac{1}{2}\Big[\sum_{y \in \mathcal{Y}^+} (\tilde{\Pr}(y|v_+) - \tilde{\Pr}(y|v_-)) \\
&\qquad + \sum_{y \in \mathcal{Y}^+} (\tilde{\Pr}(y|v_+) - \tilde{\Pr}(y|v_-)) \Big] \\
&\quad (\because \sum_{y \in \mathcal{Y}^-} \tilde{\Pr}(y|v) = 1 - \sum_{y \in \mathcal{Y}^+} \tilde{\Pr}(y|v) \ \ \forall v \in \mathcal{V}) \\
&= \sum_{y \in \mathcal{Y}^+} (\tilde{\Pr}(y|v_+) - \tilde{\Pr}(y|v_-)). \qquad (\mathrm{A}\cdot 1)
\end{aligned}
$$

Similarly, we obtain

$$
D_{tv}(\tilde{\Pr}(Y|v_+), \tilde{\Pr}(Y|v_-)) = \sum_{y \in \mathcal{Y}^-} (\tilde{\Pr}(y|v_-) - \tilde{\Pr}(y|v_+)). \qquad (\mathrm{A}\cdot 2)
$$

Combining Eq. (A· 1) and Eq. (A· 2) and with the fact that $\tilde{\Pr}(v_+) + \tilde{\Pr}(v_-) = 1$, we have

$$
\begin{aligned}
& D_{tv}(\tilde{\Pr}(Y|v_+), \tilde{\Pr}(Y|v_-)) \\
&= \tilde{\Pr}(v_-) \sum_{y \in \mathcal{Y}^+} (\tilde{\Pr}(y|v_+) - \tilde{\Pr}(y|v_-)) \\
&\qquad + \tilde{\Pr}(v_+) \sum_{y \in \mathcal{Y}^-} (\tilde{\Pr}(y|v_-) - \tilde{\Pr}(y|v_+)) \\
&= \sum_{y \in \mathcal{Y}^+} (\tilde{\Pr}(v_-)\tilde{\Pr}(y|v_+) - \tilde{\Pr}(y, v_-)) \\
&\qquad + \sum_{y \in \mathcal{Y}^-} (\tilde{\Pr}(v_+)\tilde{\Pr}(y|v_-) - \tilde{\Pr}(y, v_+)) \\
&= \sum_{y \in \mathcal{Y}^+} ((1 - \tilde{\Pr}(v_+))\tilde{\Pr}(y|v_+) - \tilde{\Pr}(y, v_-)) \\
&\qquad + \sum_{y \in \mathcal{Y}^-} ((1 - \tilde{\Pr}(v_-))\tilde{\Pr}(y|v_-) - \tilde{\Pr}(y, v_+)) \\
&= \sum_{y \in \mathcal{Y}^+} (\tilde{\Pr}(y|v_+) - \tilde{\Pr}(y)) + \sum_{y \in \mathcal{Y}^-} (\tilde{\Pr}(y|v_-) - \tilde{\Pr}(y)) \\
&\quad (\because \tilde{\Pr}(y, v_+) + \tilde{\Pr}(y, v_-) = \tilde{\Pr}(y) \ \ \forall y \in \mathcal{Y}) \\
&= \sum_{y \in \mathcal{Y}^+} \tilde{\Pr}(y) \frac{\tilde{\Pr}(y|v_+)}{\tilde{\Pr}(y)} + \sum_{y \in \mathcal{Y}^-} \tilde{\Pr}(y) \frac{\tilde{\Pr}(y|v_-)}{\tilde{\Pr}(y)} - 1. \qquad (\mathrm{A}\cdot 3)
\end{aligned}
$$

From definition of $\mathcal{Y}^+$ and $\mathcal{Y}^-$,

$$
\tilde{\Pr}(y|v_+) = \max\{\tilde{\Pr}(y|v_+), \tilde{\Pr}(y|v_-)\} \text{ if } y \in \mathcal{Y}^+, \text{ and} \qquad (\mathrm{A}\cdot 4)
$$

$$
\tilde{\Pr}(y|v_-) = \max\{\tilde{\Pr}(y|v_+), \tilde{\Pr}(y|v_-)\} \text{ if } y \in \mathcal{Y}^- \qquad (\mathrm{A}\cdot 5)
$$

hold. By substituting Eqs. (A· 4) and (A· 5) into Eq. (A· 3), we have

$$
\begin{aligned}
& D_{tv}(\tilde{\Pr}(Y|v_+), \tilde{\Pr}(Y|v_-)) \\
&= \sum_{y \in \mathcal{Y}} \tilde{\Pr}(y) \frac{\max\{\tilde{\Pr}(y|v_+), \tilde{\Pr}(y|v_-)\}}{\tilde{\Pr}(y)} - 1 \\
&= \mathrm{E}_Y \left[\frac{\max\{\tilde{\Pr}(y|v_+), \tilde{\Pr}(y|v_-)\}}{\tilde{\Pr}(y)} \right] - 1 \\
&= \mathrm{E}_Y \left[\max \left\{ \frac{\tilde{\Pr}(y, v_+)}{\tilde{\Pr}(y)\tilde{\Pr}(v_+)}, \frac{\tilde{\Pr}(y, v_-)}{\tilde{\Pr}(y)\tilde{\Pr}(v_-)} \right\} \right] - 1 \\
&= \mathrm{E}_Y \left[\max_{v \in \{v_+, v_-\}} \frac{\tilde{\Pr}(y, v)}{\tilde{\Pr}(y)\tilde{\Pr}(v)} \right] - 1.
\end{aligned}
$$

□

Kazuto Fukuchi received a B.E. degree from the University of Tsukuba, Tsukuba, Japan in 2013. He is currently pursuing a Master in computer science at Department of Computer Science, School of System and Information Engineering, University of Tsukuba, Japan. His research interests include data mining, machine learning and their applications.

Toshihiro Kamishima was born in 1968. He has received the master's degree in Engineering at the Kyoto university in 1994, and received the degree of doctor of informatics at the Kyoto university in 2001. He has joined to Electrotechnical Laboratory in 1994, and it is reorganized into Advanced Industrial Science and Technology. He received Japanese Society for Artificial Intelligence Annual Conference Awards in 2003, 2008, 2011, 2014, and Japanese Society for Artificial Intelligence Distinguished Service Award in 2009. His research interests are recommender systems, data mining, and machine learning. He is a member of AAAI, ACM, and JSAI.

Jun Sakuma received a Ph.D. degree in Engineering from the Tokyo Institute of Technology, Tokyo Japan in 2003. He has been an associate professor in the Department of Computer Science, School of System and Information Engineering, University of Tsukuba, Tsukuba, Japan, since 2009. Prior to that, he worked as an assistant professor in the Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Tokyo, Japan (2004–2009). He worked as a researcher at Tokyo Research Laboratory, IBM, Tokyo Japan (2003–2004). His research interests include data mining, machine learning, data privacy and secruity.