Prediction with Model-based Neutrality

Kazuto Fukuchi[1], Jun Sakuma[1], and Toshihiro Kamishima[2]

[1] University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8577 Japan
kazuto@mdl.cs.tsukuba.ac.jp and jun@cs.tsukuba.ac.jp
[2] National Institute of Advanced Industrial Science and Technology (AIST), AIST Tsukuba Central 2, Umezono 1-1-1, Tsukuba, Ibaraki, 305-8568 Japan
mail@kamishima.net

Abstract. 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, 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 deemed 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.

Keywords: 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 treatments.

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

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 [2]	CV Score	multiple	multiple	×
PR [8]	mutual information	any	multiple	×
Lipschitz property [4]	statistical parity	multiple	multiple	×
η-neutral logistic regression (proposal)	η-neutrality	multiple	multiple	√
η-neutral linear regression (proposal)	η-neutrality	continuous	continuous	√

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 [10]. 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 cause 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 [4][6][12]. 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 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 [2][11].

Calders et al. presented the *Calders–Verwer* 2 *Naive Bayes method* (CV2NB), which proactively removes the redlining effect [2]. 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 $\mathrm{CV}(\mathcal{D}) = p(y_+|v_+) - p(y_+|v_-)$. The CV2NB modifies the naive 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. Related to the CV2NB, it has been shown [14] that positive CV scores do not necessarily cause discrimination in some situations. There is also a method [9] 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. have introduced the *prejudice remover regularizer* (PR) for fairness-aware classification [8]. The PR regularizer penalizes the loss function if there is a high correlation between the target variable and the viewpoint variable. The penalty 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. [7]. Continuous viewpoint variables, however, cannot be treated by the PR method.

Dwork et al. have presented a classification method that uses a fairness-aware framework, in which statistical parity is used as the measure of fairness [4]. 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.

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 (Section 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 (Section 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 (Section 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 with logistic regression (Section 3) and linear regression (Section 4). The effectiveness of our methods is examined by both artificial and real datasets in Section 5.

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.

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. In the following discussion, 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

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

2.1 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.

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 and V, the discussion below is 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, 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*

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

Noting $\sum_{y \in \mathcal{Y}, v \in \mathcal{V}} \Pr(y, v) = 1$ holds, the dependency represented by $\Pr(v, y)/\Pr(v)\Pr(y) < 1$ is no need to consider.

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

$$M(X, Y, V; \theta, \phi) = \Pr(X) f(Y|X;\theta) g(V|X;\phi). \tag{3}$$

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. 3 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 $\forall v \in \mathcal{V}, y \in \mathcal{Y}$,*

$$\int_x \Pr(x) f(y|x;\theta) \left[g(v|x) - (1 + \eta)\bar{g}(v) \right] dx \leq 0, \tag{4}$$

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

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

$$\Pr(y, v) = \int_x \Pr(x, y, v)dx = \int_x \Pr(x)f(y|x;\theta)g(v|x)dx,$$
$$\Pr(y) = \int_x \int_v \Pr(x, y, v)dvdx = \int_x \Pr(x)f(y|x;\theta)dx,$$
$$\Pr(v) = \int_x \int_y \Pr(x, y, v)dydx = \int_x \Pr(x)g(v|x)dx = \bar{g}(v).$$

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

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

Thus, M is η-neutral if Eq. 4 holds.

2.2 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. 3 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*

$$\forall y, v, \sum_{i=1}^{N} f(y|x_i;\theta)\left[g(v|x_i) - (1+\eta)\tilde{g}(v)\right] \leq 0, \tag{5}$$

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

Proof. Theorem 1 states that $\Pr(X, Y, V)$ is η-neutral if Eq. 4 holds. By substituting $\tilde{\Pr}(X)$ into Eq. 4, the neutrality condition is rewritten as

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

Thus, M is empirically η-neutral if Eq. 5 holds.

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

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

2.3 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 [2] or mutual information [8]. 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 because the neutrality is evaluated with respect to each combination of possible target value y and viewpoint value v.

2.4 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.

2.5 Equivalence of η-neutrality and Statistical Parity

In this subsection, in order to discuss the equivalence of η-neutrality and the statistical parity[4], we assume examples $\mathcal{D}$ contains the viewpoint values. The statistical parity defines the neutrality considering the difference of two probabilistic distribution of target y, $P(y)$ and $Q(y)$,

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

Given $\epsilon \geq 0$ as a neutrality parameter, the statistical parity is defined by

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

where $\Pr(Y|V)$ is empirically evaluated with the given example set $\mathcal{D}$.

If the empirical distribution of the example set $\mathcal{D}$ is used as the model of the viewpoint in the empirical η-neutrality, and letting the distance function of the statistical parity

$$D_\eta(P,Q) = \max_{y\in\mathcal{Y}} \frac{\max\{P(y), Q(y)\}}{\tilde{\Pr}(v_+)P(y) + \tilde{\Pr}(v_-)Q(y)},$$

the statistical parity with parameter η is equivalent to the η-neutrality. The proof of the equivalence will be presented in the journal version of this manuscript.

In the following two sections, 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 η-neutral Logistic Regression

In this section, 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

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

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,

$$\begin{aligned}
\nabla L(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \Big(\sigma(\boldsymbol{\theta}^T \boldsymbol{x}_i) - y_i \Big) \boldsymbol{x}_i, \\
\nabla^2 L(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_i)(1 - \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_i)) \boldsymbol{x}_i \boldsymbol{x}_i^T .
\end{aligned}$$

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

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

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 Shor's r-algorithm based on adaptive space dilation [13]. Shor's r-algorithm can be initialized with any solution. We set the initial solution to the result of the logistic regression without the neutrality constraint. Although the constraint is nonconvex, in Section 5 we show by experiment η-neutrality can be achieved without sacrificing too much of the accuracy of the prediction. 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.

4 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

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

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

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

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:

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

where $\boldsymbol{X} = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, ..., \boldsymbol{x}_N^T)^T$ is the matrix of input vectors and $\boldsymbol{y} = (y_1, y_2, ..., y_N)^T$ 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,

$$\begin{aligned}\nabla_{\boldsymbol{w}} f(y|\boldsymbol{x}; \boldsymbol{w}, \beta) =& \nabla_{\boldsymbol{w}} \exp(-\ln f(y|\boldsymbol{x}; \boldsymbol{w}, \beta)) = -\beta(\boldsymbol{w}^T\boldsymbol{x} - y) f(y|\boldsymbol{x}; \boldsymbol{w}, \beta)\boldsymbol{x},\\ \nabla^2_{\boldsymbol{w}} f(y|\boldsymbol{x}; \boldsymbol{w}, \beta) =& \alpha(\boldsymbol{x}, y, \boldsymbol{w}, \beta)\beta f(y|\boldsymbol{x}; \boldsymbol{w}, \beta)\boldsymbol{x}\boldsymbol{x}^T,\end{aligned}$$

where $\alpha(\boldsymbol{x}, y, \boldsymbol{w}, \beta) = \beta(\boldsymbol{w}^T\boldsymbol{x} - y)^2 - 1$.

Since, depending on $\boldsymbol{w}$, $f(y|\boldsymbol{x}; \boldsymbol{w}, \beta) \geq 0$ and $\alpha(\boldsymbol{x}, y, \boldsymbol{w}, \beta) \in \mathbb{R}$ can take negative values, the Hessian is not positive definite. Hence, unfortunately, f is not convex with respect to $\boldsymbol{w}$. For this nonlinear constraint optimization, we again use Shor's r-algorithms for the experiments.

5 Experiments

5.1 Classification

Settings. In order to examine and compare the classification performance and the neutralization effect of the η-neutral logistic regression with other methods,

we performed experiments on two real data sets, Adult [5] and the Dutch Census [3]. Table 2 summarizes the specifications of each dataset. In both datasets, the target and viewpoint variables are set to "income (large/small)" and "gender (male/female)", respectively. Our method does not necessarily require that the viewpoint value (gender, in this case) be explicitly provided in the given dataset, but for comparison with other methods, it was chosen from the input variable of the dataset.

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 [2], logistic regression that uses the PR [7], and η-neutral logistic regression with viewpoint neutrality (VN, 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 $\eta \in \{0.00, 0.01, ..., 0.40\}$

As neutrality indices of prediction models, the 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 [8]:

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

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. 3 is empirically $\hat{\eta}$-neutral with respect to the given examples.

We compared the three measures: the accuracy, the normalized prejudice index (NPI), and the $\hat{\eta}$ of η-neutrality. These indices was evaluated with five-fold cross validation.

The values used for the learning of $f(y|x)$, the guarantee of neutrality, and the measurement of neutrality are summarized in Table 3. 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

Table 2. Specification of datasets. $\#y_+$ and $\#v_+$ represent the number of positive target and viewpoint values, respectively. The prediction accuracy (logistic regression) of the target (Acc(y) w.r.t. income) and viewpoint (Acc(v) w.r.t. gender) variables are also shown.

dataset	Adult	Dutch Census
#Instances	16281	60420
#Attributes	13	10
$\#y_+$	3846(23.6%)	31657(52.4%)
$\#v_+$	10860(66.7%)	30273(50.1%)
Acc(y)	0.851	0.835
Acc(v)	0.842	0.665

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 1 shows the experimental results. In the graphs, the best result is at the left top. Comparing the results of NB and NBns in Case 1, we can see that the improvement of neutrality by elimination of the viewpoint variable is limited. The same applies to LR and LRns.

In Case 1, CV2NB achieves a neutrality of nearly 0 in terms of both NPI and $\hat{\eta}$ in both datasets. In addition, the decrease in the accuracy of the prediction is less than 1% in the Adult dataset and 5% in the Dutch Census. Thus, neutralization by CV2NB works successfully in Case 1. On the other hand, neutralization by CV2NB does not work well in Case 2; the neutralization level is almost the same as it is for NBns. CV2NB modifies the target prediction model so that the CV score with respect to the given examples becomes zero. This can cause the prediction model to overfit the given examples. Hence, the NPI and $\hat{\eta}$ of CV2NB with respect to the unseen values of the viewpoint are large, as seen in the results of Case 2.

In Case 1, PR successfully balances the NPI and the accuracy for the Adult dataset, but it fails to balance the accuracy and $\hat{\eta}$. This is because NPI evaluates neutrality with respect to the average of the given examples, while $\hat{\eta}$ evaluates the lowest neutrality for all values of y and v as the worst case. This result indicates that the dependency of the predictions of target Y to the predictions of viewpoint V can be strong for some y and v, even when the NPI is kept small. In Case 2, neutralization with PR did not work well in either dataset. This was again due to overfitting; this can be confirmed by the fact that the NPI of v and

Table 3. 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	$\hat{y}, v$
	ours	$\boldsymbol{x}, v$	$g(v\|\boldsymbol{x})$	$\hat{y}, v$
Case 2	others	$\boldsymbol{x}, \hat{v}$	$\hat{v}$	$\hat{y}, v$
	ours	$\boldsymbol{x}$	$g(v\|\boldsymbol{x})$	$\hat{y}, v$

y is large in Case 2 even when the NPI of $\hat{v}$ and y is kept small (these results are omitted due to space limitations).

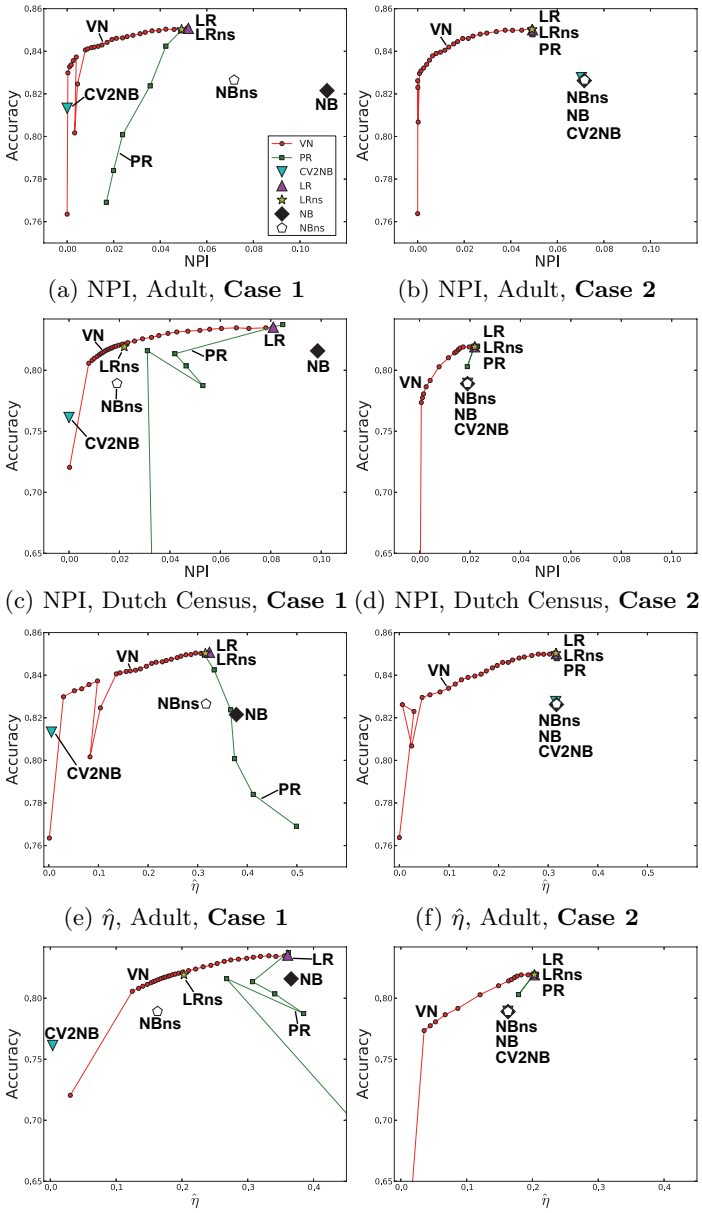
In both cases, our proposal, VN, successfully balances neutralization and accuracy of the predication by changing η. Furthermore, the decrease in the accuracy of the prediction was at most 5%, even after strong neutralization with small η. In some cases, the accuracy of VN becomes unstable with small η. The reason is thought to be the nonconvexity of the neutrality constraint. VN always guarantees neutrality of the prediction model, but the accuracy of the prediction can suddenly drop if the solution is captured by a local optimum.

5.2 Regression

Settings. In order to investigate the behaviors of neutralization in linear regression, we performed experiments of η-neutral linear regression with the Housing dataset [5]. This dataset contains 506 examples with 14 attributes; the MEDV (median value of owner-occupied homes, in $1000s) and the LSTAT (% lower status of the population) were used as the target and viewpoint values, respectively. 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 accuracy of the prediction was measured by the root-mean-square error (RMSE), and $\hat{\eta}$ was used as the measure of neutrality.

Results. Figure 2 shows the scatter plots of $(\hat{y}, y)$ (the top row) and $(\hat{y}, \hat{v})$ (the bottom row). From left to right, the neutrality parameter η was varied as $\eta \in \{1.0, 3.0, 10.0\}$. 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, any correlation between $\hat{y}$ and $\hat{v}$ appears in the $(\hat{y}, \hat{v})$ plot.

In Figure 2 (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 Figure 2, 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. This result indicates that our method can use η to successfully control the neutralization level of the regression model. The RMSE increases as η is decreases. In Figure 2 (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



(a) NPI, Adult, **Case 1** (b) NPI, Adult, **Case 2**

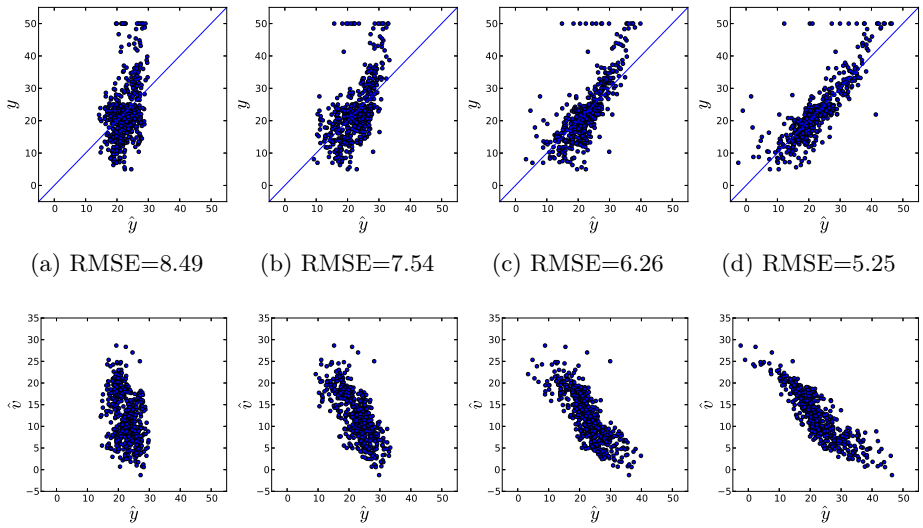
(c) NPI, Dutch Census, **Case 1** (d) NPI, Dutch Census, **Case 2**

(e) $\hat{\eta}$, Adult, **Case 1** (f) $\hat{\eta}$, Adult, **Case 2**

(g) $\hat{\eta}$, Dutch Census, **Case 1** (h) $\hat{\eta}$, Dutch Census, **Case 2**

Fig. 1. Accuracy vs. neutrality measure.

model is well neutralized. Thus, tuning of η is important to obtain a neutralized regression model with high accuracy.



(a) RMSE=8.49 (b) RMSE=7.54 (c) RMSE=6.26 (d) RMSE=5.25

(e) $\eta = 1.0$ (f) $\eta = 3.0$ (g) $\eta = 10.0$ (h) no neutralization

Fig. 2. 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.

6 Conclusion

In this paper, we proposed a framework in which to use 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.

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 Section 3 and Section 4, 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.

Acknowledgments. The work is supported by FIRST program "Development of the Fastest Database Engine for the Era of Very Large Database and Experiment and Evaluation of Strategic Social Services Enabled by the Database Engine" and is partially supported by JSPS KAKENHI, Grant Number 24500194, 25540094.

References

1. Boyd, D.: Privacy and publicity in the context of big data. In: Keynote Talk of The 19th Intl Conf. on World Wide Web (2010)
2. Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery 21(2), 277–292 (Sep 2010)
3. Dutch Central Bureau for Statistics: "Volkstelling" (2001)
4. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. pp. 214–226. ACM (2012)
5. Frank, A., Asuncion, A.: UCI machine learning repository (2010), `http://archive.ics.uci.edu/ml`
6. Kamiran, F., Calders, T., Pechenizkiy, M.: Discrimination aware decision tree learning. In: Data Mining (ICDM), 2010 IEEE 10th International Conference on. pp. 869–874. IEEE (2010)
7. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Enhancement of the neutrality in recommendation. In: Proceedings of the 2nd Workshop on Human Decision Making in Recommender Systems (Decisions@RecSys). pp. 8–14 (2012)
8. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: in Proceedings of the ECML/PKDD2012, Part II. vol. LNCS 7524, pp. 35–50. Springer (2012)
9. Luong, B.T., Ruggieri, S., Turini, F.: k-nn as an implementation of situation testing for discrimination discovery and prevention. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data mining. pp. 502–510. KDD '11, ACM, New York, NY, USA (2011)
10. Pariser, E.: The Filter Bubble: What The Internet Is Hiding From You. Viking, London (2011)
11. Pedreshi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 560–568. ACM (2008)
12. Ruggieri, S., Pedreschi, D., Turini, F.: Dcube: Discrimination discovery in databases. In: Proceedings of the 2010 international conference on Management of data. pp. 1127–1130. ACM (2010)
13. Shor, N.Z., Kiwiel, K.C., Ruszcayǹski, A.: Minimization methods for non-differentiable functions. Springer-Verlag New York, Inc., New York, NY, USA (1985)
14. Žliobaitė, I., Kamiran, F., Calders, T.: Handling conditional discrimination. In: Proceedings of the 2011 IEEE 11th International Conference on Data Mining. pp. 992–1001. ICDM '11, IEEE Computer Society, Washington, DC, USA (2011)