Neutralized Empirical Risk Minimization with Generalization Neutrality Bound

Kazuto Fukuchi and Jun Sakuma

University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8577 Japan
kazuto@mdl.cs.tsukuba.ac.jp, jun@cs.tsukuba.ac.jp

Abstract. Currently, machine learning plays an important role in the lives and individual activities of numerous people. Accordingly, it has become necessary to design machine learning algorithms to ensure that discrimination, biased views, or unfair treatment do not result from decision making or predictions made via machine learning. In this work, we introduce a novel empirical risk minimization (ERM) framework for supervised learning, neutralized ERM (NERM) that ensures that any classifiers obtained can be guaranteed to be neutral with respect to a viewpoint hypothesis. More specifically, given a viewpoint hypothesis, NERM works to find a target hypothesis that minimizes the empirical risk while simultaneously identifying a target hypothesis that is neutral to the viewpoint hypothesis. Within the NERM framework, we derive a theoretical bound on empirical and generalization neutrality risks. Furthermore, as a realization of NERM with linear classification, we derive a max-margin algorithm, neutral support vector machine (SVM). Experimental results show that our neutral SVM shows improved classification performance in real datasets without sacrificing the neutrality guarantee.

Keywords: neutrality, discrimination, fairness, classification, empirical risk minimization, support vector machine.

1 Introduction

Within the framework of empirical risk minimization (ERM), a supervised learning algorithm seeks to identify a hypothesis f that minimizes empirical risk with respect to given pairs of *input* x and *target* y. Given an input x without the target value, hypothesis f provides a prediction for the target of x as $y = f(x)$. In this study, we add a new element, *viewpoint hypothesis* g, to the ERM framework. Similar to hypothesis f, which is given an input x without the viewpoint value, viewpoint hypothesis g provides a prediction for the viewpoint of the x as $v = g(x)$. In order to distinguish between the two different hypotheses, f and g, f will be referred to as the *target hypothesis*. Examples of the viewpoint hypothesis are given with the following specific applications.

With this setup in mind, we introduce our novel framework for supervised learning, *neutralized ERM* (NERM). Intuitively, we say that a target hypothesis is neutral to a given viewpoint hypothesis if there is low correlation between

T. Calders et al. (Eds.): ECML PKDD 2014, Part I, LNCS 8724, pp. 418–433, 2014.

the target $f(x)$ and viewpoint $g(x)$. The objective of NERM is to find a target hypothesis f that minimizes empirical risks while simultaneously remaining neutral to the viewpoint hypothesis g. The following two application scenarios motivate NERM.

Application 1 (Filter bubble). Suppose an article recommendation service provides personalized article distribution. In this situation, by taking a user's access logs and profile as input x, the service then predicts that user's preference with respect to articles using supervised learning as $y = f(x)$ (target hypothesis). Now, suppose a user strongly supports a policy that polarizes public opinion (such as nuclear power generation or public medical insurance). Furthermore, suppose the user's opinion either for or against the particular policy can be precisely predicted by $v = g(x)$ (viewpoint hypothesis). Such a viewpoint hypothesis can be readily learned by means of supervised learning, given users' access logs and profiles labeled with the parties that the users support. In such situations, if predictions by the target hypothesis f and viewpoint hypothesis g are closely correlated, recommended articles are mostly dominated by articles supportive of the policy, which may motivate the user to adopt a biased view of the policy [12]. This problem is referred to as the *filter bubble* [10]. Bias of this nature can be avoided by training the target hypothesis so that the predicted target is independent of the predicted viewpoint.

Application 2 (Anti-discrimination). Now, suppose a company wants to make hiring decisions using information collected from job applicants, such as age, place of residence, and work experience. While taking such information as input x toward the hiring decision, the company also wishes to predict the potential work performance of job applicants via supervised learning, as $y = f(x)$ (target hypothesis). Now, although the company does not collect applicant information on sensitive attributes such as race, ethnicity, or gender, suppose such sensitive attributes can be sufficiently precisely predicted from an analysis of the non-sensitive applicant attributes, such as place of residence or work experience, as $v = g(x)$ (viewpoint hypothesis). Again, such a viewpoint hypothesis can be readily learned by means of supervised learning by collecting moderate number of labeled examples. In such situations, if hiring decisions are made by the target hypothesis f and if there is a high correlation with the sensitive attribute predictions $v = g(x)$, those decisions might be deemed discriminatory [11]. In order to avoid this, the target hypothesis should be trained so that the decisions made by f are not highly dependent on the sensitive attributes predicted by g. Thus, this problem can also be interpreted as an instance of NERM.

The neutrality of a target hypothesis should not only be guaranteed for given samples, but also for unseen samples. In the article recommendation example, the recommendation system is trained using the user's past article preferences, whereas recommendation neutralization is needed for unread articles. In the hiring decision example, the target hypothesis is trained with information collected from the past histories of job applicants, but the removal of discrimination from hiring decisions is the desired objective.

Given a viewpoint hypothesis, we evaluate the degree of neutrality of a target hypothesis with respect to given and unseen samples as *empirical neutrality risk* and *generalization neutrality risk*, respectively. The goal of NERM is to show that the generalization risk is theoretically bounded in the same manner as the standard ERM [2,1,6], and, simultaneously, to show that the generalization neutrality risk is also bounded with respect to given viewpoint hypothesis.

Our Contribution. The contribution of this study is three-fold. First, we introduce our novel NERM framework in which, assuming the target hypothesis and viewpoint hypothesis output binary predictions, it is possible to learn a target hypothesis that minimizes empirical and empirically neutral risks. Given samples and a viewpoint hypothesis, NERM is formulated as a convex optimization problem where the objective function is the linear combination of two terms, the empirical risk term penalizing the target hypothesis prediction error and the neutralization term penalizing correlation between the target and the viewpoint. The predictive performance and neutralization can be balanced by adjusting a parameter, referred to as the neutralization parameter. Because of its convexity, the optimality of the resultant target hypothesis is guaranteed (in Section 4).

Second, we derive a bound on empirical and generalization neutrality risks for NERM. We also show that the bound on the generalization neutrality risk can be controlled by the neutralization parameter (in Section 5). As discussed in Section 2, a number of diverse algorithms targeting the neutralization of supervised classifications have been presented. However, none of these have given theoretical guarantees on generalization neutrality risk. To the best of our knowledge, this is the first study that gives a bound on generalization neutrality risk.

Third, we present a specific NERM learning algorithm for neutralized linear classification. The derived learning algorithm is interpreted as a *support vector machine* (SVM) [14] variant with a neutralization guarantee. The kernelized version of the neutralization SVM is also derived from the dual problem (in Section 6).

2 Related Works

Within the context of removing discrimination from classifiers, the need for a neutralization guarantee has already been extensively studied. Calders & Verwer [4] pointed out that elimination of sensitive attributes from training samples does not help to remove discrimination from the resultant classifiers. In the hiring decision example, even if we assume that a target hypothesis is trained with samples that have no race or ethnicity attributes, hiring decisions may indirectly correlate with race or ethnicity through addresses if there is a high correlation between an individual's address and his or her race or ethnicity. This indirect effect is referred to as a *red-lining effect* [3].

Calders & Verwer [4] proposed the Calders–Verwer 2 Naïve Bayes method (CV2NB) to remove the red-lining effect from the Naïve Bayes classifier. The CV2NB method is used to evaluate the Calders–Verwer (CV) score, which is a measure that evaluates discrimination of naïve Bayes classifiers. The CV2NB method learns the naïve Bayes classifier in a way that ensures the CV score is made as small as possible. Based on this idea, various situations where discrimination can occur have been discussed in other studies [16,7]. Since a CV score is empirically measured with the given samples, naïve Bayes classifiers with low CV scores result in less discrimination for those samples. However, less discrimination is not necessarily guaranteed for unseen samples. Furthermore, the CV2NB method is designed specifically for the naïve Bayes model and does not provide a general framework for anti-discrimination learning.

Zemel et al. [15] introduced the learning fair representations (LFR) model for preserving classification fairness. LFR is designed to provide a map, from inputs to prototypes, that guarantees the classifiers that are learned with the prototypes will be fair from the standpoint of statistical parity. Kamishima et al. [8] presented a prejudice remover regularizer (PR) for fairness-aware classification that is formulated as an optimization problem in which the objective function contains the loss term and the regularization term that penalizes mutual information between the classification output and the given sensitive attributes. The classifiers learned with LFR or PR are empirically neutral (i.e., fair or less discriminatory) in the sense of statistical parity or mutual information, respectively. However, no theoretical guarantees related to neutrality for unseen samples have been established for these methods.

Fukuchi et al. [5] introduced *η-neutrality*, a framework for neutralization of probability models with respect to a given viewpoint random variable. Their framework is based on maximum likelihood estimation and neutralization is achieved by maximizing likelihood estimation while setting constraints to enforce η-neutrality. Since η-neutrality is measured using the probability model of the viewpoint random variable, the classifier satisfying η-neutrality is expected to preserve neutrality for unseen samples. However, this method also fails to provide a theoretical guarantee for generalization neutrality.

LFR, PR, and η-neutrality incorporate a hypothesis neutrality measure into the objective function in the form of a regularization term or constraint; however, these are all non-convex. One of the reasons why generalization neutrality is not theoretically guaranteed for these methods is the non-convexity of the objective functions. In this study, we introduce a convex surrogate for a neutrality measure in order to provide a theoretical analysis of generalization neutrality.

3 Empirical Risk Minimization

Let X and Y be an input space and a target space, respectively. We assume $D_n = \{(x_i, y_i)\}_{i=1}^n \in Z^n$ $(Z = X \times Y)$ to be a set of i.i.d. samples drawn from an unknown probability measure ρ over $(Z, \mathcal{Z})$. We restrict our attention to binary classification, $Y = \{-1, 1\}$, but our method can be expanded to handle multi-valued classification via a straightforward modification. Given the i.i.d. samples, the supervised learning objective is to construct a target hypothesis $f : X \to \mathbb{R}$ where the hypothesis is chosen from a class of measurable functions $f \in \mathcal{F}$. We assume that classification results are given by $\mathrm{sgn} \circ f(x)$, that is,

$y = 1$ if $f(x) > 0$; otherwise $y = -1$. Given a loss function $\ell : Y \times \mathbb{R} \to \mathbb{R}^+$, the generalization risk is defined by

$$R(f) = \int \ell(y, f(x)) d\rho.$$

Our goal is to find $f^* \in \mathcal{F}$ that minimizes the generalization risk $R(f)$. In general, ρ is unknown and the generalization risk cannot be directly evaluated. Instead, we minimize the empirical loss with respect to sample set D_n

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)).$$

This is referred to as *empirical risk minimization (ERM).*

In order to avoid overfitting, a regularization term $\Omega : \mathcal{F} \to \mathbb{R}^+$ is added to the empirical loss by penalizing complex hypotheses. Minimization of the empirical loss with a regularization term is referred to as *regularized ERM (RERM).*

3.1 Generalization Risk Bound

Rademacher Complexity measures the complexity of a hypothesis class with respect to a probability measure that generates samples. The Rademacher Complexity of class $\mathcal{F}$ is defined as

$$\mathcal{R}_n(\mathcal{F}) = \mathrm{E}_{D_n, \boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right]$$

where $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_n)^T$ are independent random variables such that $\Pr(\sigma_i = 1) = \Pr(\sigma_i = -1) = 1/2$. Bartlett & Mendelson [2] derived a generalization loss bound using the Rademacher complexity as follows:

Theorem 1 (Bartlett & Mendelson [2]). *Let ρ be a probability measure on $(Z, \mathcal{Z})$ and let $\mathcal{F}$ be a set of real-value functions defined on X, with* $\sup\{|f(x)| : f \in \mathcal{F}\}$ *finite for all $x \in X$. Suppose that $\phi : \mathbb{R} \to [0, c]$ satisfies and is Lipschitz continuous with constant L_ϕ. Then, with probability at least $1 - \delta$, every function in $\mathcal{F}$ satisfies*

$$R(f) \leq R_n(f) + 2L_\phi \mathcal{R}_n(\mathcal{F}) + c\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

4 Generalization Neutrality Risk and Empirical Neutrality Risk

In this section, we introduce the viewpoint hypothesis into the ERM framework and define a new principle of supervised learning, neutralized ERM (NERM), with the notion of *generalization neutrality risk.* Convex relaxation of the neutralization measure is also discussed in this section.

4.1 +1/−1 Generalization Neutrality Risk

Suppose a measurable function $g : X \to \mathbb{R}$ is given. The prediction of g is referred to as the *viewpoint* and g is referred to as the *viewpoint hypothesis*. We say the target hypothesis f is neutral to the viewpoint hypothesis g if the target predicted by the learned target hypothesis f and the viewpoint predicted by the viewpoint hypothesis g are not mutually correlating. In our setting, we assume the target hypothesis f and viewpoint hypothesis g to give binary predictions by $\mathrm{sgn} \circ f$ and $\mathrm{sgn} \circ g$, respectively. Given a probability measure ρ and a viewpoint hypothesis g, the neutrality of the target hypothesis f is defined by the correlation between $\mathrm{sgn} \circ f$ and $\mathrm{sgn} \circ g$ over ρ. If $f(x)g(x) > 0$ holds for multiple samples, then the classification $\mathrm{sgn} \circ f$ closely correlates to the viewpoint $\mathrm{sgn} \circ g$. On the other hand, if $f(x)g(x) \leq 0$ holds for multiple samples, then the classification $\mathrm{sgn} \circ f$ and the viewpoint $\mathrm{sgn} \circ g$ are inversely correlating. Since we want to suppress both correlations, our neutrality measure is defined as follows:

Definition 1 (+1/-1 Generalization Neutrality Risk). *Let $f \in \mathcal{F}$ and $g \in \mathcal{G}$ be a target hypothesis and viewpoint hypothesis, respectively. Let ρ be a probability measure over $(Z, \mathscr{Z})$. Then, the* +1/-1 generalization neutrality risk *of target hypothesis f with respect to viewpoint hypothesis g over ρ is defined by*

$$C_{\mathrm{sgn}}(f, g) = \left| \int \mathrm{sgn}(f(x)g(x)) d\rho \right|.$$

When the probability measure ρ cannot be obtained, a +1/−1 generalization neutrality risk $C_{\mathrm{sgn}}(f, g)$ can be empirically evaluated with respect to the given samples D_n.

Definition 2 (+1/−1 Empirical Neutrality Risk). *Suppose that $D_n = \{(x_i, y_i)\}_{i=1}^n \in Z^n$ is a given sample set. Let $f \in \mathcal{F}$ and $g \in \mathcal{G}$ be the target hypothesis and the viewpoint hypothesis, respectively. Then, the* +1/−1 empirical neutrality risk *of target hypothesis f with respect to viewpoint hypothesis g is defined by*

$$C_{n,\mathrm{sgn}}(f, g) = \frac{1}{n} \left| \sum_{i=1}^{n} \mathrm{sgn}(f(x_i)g(x_i)) \right|. \tag{1}$$

4.2 Neutralized Empirical Risk Minimization (NERM)

With the definition of neutrality risk, a novel framework, the *Neutralized Empirical Risk Minimization* (NERM) is introduced. NERM is formulated as minimization of the empirical risk and empirical +1/−1 neutrality risk:

$$\min_{f \in \mathcal{F}} \; R_n(f) + \Omega(f) + \eta C_{n,\mathrm{sgn}}(f, g). \tag{2}$$

where $\eta > 0$ is the neutralization parameter which determines the trade-off ratio between the empirical risk and the empirical neutrality risk.

4.3 Convex Relaxation of +1/−1 Neutrality Risk

Unfortunately, the optimization problem defined by Eq (2) cannot be efficiently solved due to the nonconvexity of Eq (1). Therefore, we must first relax the absolute value function of $C_{\mathrm{sgn}}(f,g)$ into the max function. Then, we introduce a convex surrogate of the sign function, yielding a convex relaxation of the +1/−1 neutrality risk.

By letting I be the indicator function, the +1/−1 generalization neutrality risk can be decomposed into two terms:

$$\begin{aligned}C_{\mathrm{sgn}}(f,g) =& \Big| \underbrace{\int I(\mathrm{sgn}(g(x)) = \mathrm{sgn}(f(x)))d\rho}_{\text{prob. that } f \text{ agrees with } g} - \underbrace{\int I(\mathrm{sgn}(g(x)) \neq \mathrm{sgn}(f(x)))d\rho}_{\text{prob. that } f \text{ disagrees with } g} \Big| \\ \coloneqq& |C^{+}_{\mathrm{sgn}}(f,g) - C^{-}_{\mathrm{sgn}}(f,g)| \end{aligned} \tag{3}$$

The upper bound of the +1/−1 generalization neutrality risk $C_{\mathrm{sgn}}(f,g)$ is tight if $C^{+}_{\mathrm{sgn}}(f,g)$ and $C^{-}_{\mathrm{sgn}}(f,g)$ are close. Thus, the following property is derived.

Proposition 1. *Let $C^{+}_{\mathrm{sgn}}(f,g)$ and $C^{-}_{\mathrm{sgn}}(f,g)$ be functions defined in Eq (3). For any $\eta \in [0.5,1]$, if*

$$C^{\max}_{\mathrm{sgn}}(f,g) \coloneqq \max(C^{+}_{\mathrm{sgn}}(f,g), C^{-}_{\mathrm{sgn}}(f,g)) \leq \eta,$$

then

$$C_{\mathrm{sgn}}(f,g) = |C^{+}_{\mathrm{sgn}}(f,g) - C^{-}_{\mathrm{sgn}}(f,g)| \leq 2\eta - 1.$$

Proposition 1 shows that $C^{\max}_{\mathrm{sgn}}(f,g)$ can be used as the generalization neutrality risk instead of $C_{\mathrm{sgn}}(f,g)$. Next, we relax the indicator function contained in $C^{\pm}_{\mathrm{sgn}}(f,g)$.

Definition 3 (Relaxed Convex Generalization Neutrality Risk). *Let $f \in \mathcal{F}$ and $g \in \mathcal{G}$ be a classification hypothesis and viewpoint hypothesis, respectively. Let ρ be a probability measure over $(Z, \mathcal{Z})$. Let $\psi : \mathbb{R} \to \mathbb{R}^{+}$ be a convex function and*

$$C^{\pm}_{\psi}(f,g) = \int \psi(\pm g(x) f(x)) d\rho.$$

Then, the relaxed convex generalization neutrality risk *of f with respect to g is defined by*

$$C_{\psi}(f,g) = \max(C^{+}_{\psi}(f,g), C^{-}_{\psi}(f,g)).$$

The empirical evaluation of relaxed convex generalization neutrality risk is defined in a straightforward manner.

Definition 4 (Convex Relaxed Empirical Neutrality Risk). *Suppose $D_n = \{(x_i, y_i)\}_{i=1}^{n} \in Z^n$ to be a given sample set. Let $f \in \mathcal{F}$ and $g \in \mathcal{G}$ be the target hypothesis and the viewpoint hypothesis, respectively. Let $\psi : \mathbb{R} \to \mathbb{R}^{+}$ be a* *convex function and*

$$C^{\pm}_{n,\psi}(f,g) = \frac{1}{n}\sum_{i=1}^{n} \psi(\pm g(x_i)f(x_i)).$$

Then, relaxed convex empirical neutrality risk *of f with respect to g is defined by*

$$C_{n,\psi}(f,g) = \max(C^{+}_{n,\psi}(f,g), C^{-}_{n,\psi}(f,g)).$$

$C^{\pm}_{n,\psi}(f,g)$ is convex because it is a summation of the convex function ψ. Noting that $\max(f_1(x), f_2(x))$ is convex if f_1 and f_2 are convex, $C_{n,\psi}(f,g)$ is convex as well.

4.4 NERM with Relaxed Convex Empirical Neutrality Risk

Finally, we derive the convex formulation of NERM with the relaxed convex empirical neutrality risk as follows:

$$\min_{f\in\mathcal{F}} \; R_n(f) + \Omega(f) + \eta C_{n,\psi}(f,g). \tag{4}$$

If the regularized empirical risk is convex, then this is a convex optimization problem.

The neutralization term resembles the regularizer term in the formulation sense. Indeed, the neutralization term is different from the regularizer in that it is dependent on samples. We can interpret the regularizer as a prior structural information of the model parameters, but we cannot interpret the neutralization term in the same way due to its dependency on samples. PR and LFR have similar neutralization terms in the sense of adding the neutrality risk to objective function, and neither can be interpreted as a prior structural information. Instead, the neutralization term can be interpreted as a prior information of *data*. The notion of a prior data information is relevant to *transfer learning* [9], which aims to achieve learning dataset information from other datasets. However, further research on the relationships between the neutralization and transfer learning will be left as an area of future work.

5 Generalization Neutrality Risk Bound

In this section, we show theoretical analyses of NERM generalization neutrality risk and generalization risk. First, we derive a probabilistic uniform bound of the generalization neutrality risk for any $f \in \mathcal{F}$ with respect to the empirical neutrality risk $C_{n,\psi}(f,g)$ and the Rademacher complexity of $\mathcal{F}$. Then, we derive a bound on the generalization neutrality risk of the optimal hypothesis.

For convenience, we introduce the following notation. For a hypothesis class $\mathcal{F}$ and constant $c \in \mathbb{R}$, we denote $-\mathcal{F} = \{-f : f \in \mathcal{F}\}$ and $c\mathcal{F} = \{cf : f \in \mathcal{F}\}$. For any function $\phi : \mathbb{R} \to \mathbb{R}$, let $\phi \circ \mathcal{F} = \{\phi \circ f : f \in \mathcal{F}\}$. Similarly, for any function $g : X \to \mathbb{R}$, let $g\mathcal{F} = \{h : f \in \mathcal{F}, h(x) = g(x)f(x) \; \forall x \in X\}$.

5.1 Uniform Bound of Generalization Neutrality Risk

A probabilistic uniform bound on $C_\psi(f,g)$ for any hypothesis $f \in \mathcal{F}$ is derived as follows.

Theorem 2. *Let $C_\psi(f,g)$ and $C_{n,\psi}(f,g)$ be the relaxed convex generalization neutrality risk and the relaxed convex empirical neutrality risk of $f \in \mathcal{F}$ w.r.t. $g \in \mathcal{G}$. Suppose that $\psi : \mathbb{R} \to [0,c]$ satisfies and is Lipschitz continuous with constant L_ψ. Then, with probability at least $1-\delta$, every function in $\mathcal{F}$ satisfies*

$$C_\psi(f,g) \le C_{n,\psi}(f,g) + 2L_\psi \mathcal{R}_n(g\mathcal{F}) + c\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

As proved by Theorem 2, $C_\psi(f,g) - C_{n,\psi}(f,g)$, the approximation error of the generalization neutrality risk is uniformly upper-bounded by the Rademacher complexity of hypothesis classes $g\mathcal{F}$ and $O(\sqrt{\ln(1/\delta)/n})$, where δ is the confidence probability and n is the sample size.

5.2 Generalization Neutrality Risk Bound for NERM Optimal Hypothesis

Let $\hat{f} \in \mathcal{F}$ be the optimal hypothesis of NERM. We derive the bounds on the empirical and generalization neutrality risks achieved by $\hat{f}$ under the following conditions:

1. Hypothesis class $\mathcal{F}$ includes a hypothesis f_0 s.t. $f_0(x) = 0$ for $\forall x$, and
2. the regularization term of f_0 is $\Omega(f_0) = 0$. (A)

The conditions are relatively moderate. For example, consider the linear hypothesis $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$ and $\Omega(f) = \|\boldsymbol{w}\|_2^2$ (ℓ_2^2 norm) and let $W \subseteq \mathbb{R}^D$ be a class of the linear hypothesis. If $\mathbf{0} \in W$, the two conditions above are satisfied. Assuming that $\mathcal{F}$ satisfies these conditions, the following theorem provides the bound on the generalization neutrality risk.

Theorem 3. *Let $\hat{f}$ be the optimal target hypothesis of NERM, where the viewpoint hypothesis is $g \in \mathcal{G}$ and the neutralization parameter is η. Suppose that $\psi : \mathbb{R} \to [0,c]$ satisfies and is Lipschitz continuous with constant L_ψ. If conditions (A) are satisfied, then with probability at least $1-\delta$,*

$$C_\psi(\hat{f},g) \le \psi(0) + \phi(0)\frac{1}{\eta} + 2L_\psi \mathcal{R}_n(g\mathcal{F}) + c\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

For the proof of Theorem 3, we first derive the upper bound of the empirical neutrality risk of $\hat{f}$.

Corollary 1. *If the conditions (A) are satisfied, then the empirical relaxed convex neutrality risk of $\hat{f}$ is bounded by*

$$C_{n,\psi}(\hat{f},g) \le \psi(0) + \phi(0)\frac{1}{\eta}.$$

Theorem 3 is immediately obtained from Theorem 2 and Corollary 1.

5.3 Generalization Risk Bound for NERM

In this section, we compare the generalization risk bound of NERM with that of a regular ERM. Theorem 1 denotes a uniform bound of the generalization risk. This theorem holds with the hypotheses which are optimal in terms of NERM and ERM. However, the hypotheses which are optimal in terms of NERM and ERM have different empirical risk values. The empirical risk of NERM is greater than that of ERM since NERM has a term that penalizes less neutrality. More precisely, if we let $\bar{f}$ be the optimal hypothesis in term of ERM, we have

$$R_n(\hat{f}) - R_n(\bar{f}) \geq 0. \tag{5}$$

The reason for this is that empirical risk of any other hypothesis is greater than one of $\bar{f}$ since $\bar{f}$ minimizes empirical risk. Furthermore, due to $\hat{f}$ is a minimizer of $R_n(f) + \eta C_{n,\phi}(f, g)$, we have

$$\begin{aligned} R_n(\hat{f}) + \eta C_{n,\phi}(\hat{f}, g) - R_n(\bar{f}) - \eta C_{n,\phi}(\bar{f}, g) &\leq 0 \\ R_n(\hat{f}) - R_n(\bar{f}) &\leq \eta(C_{n,\phi}(\bar{f}, g) - C_{n,\phi}(\hat{f}, g)). \end{aligned} \tag{6}$$

Since the left term of this inequality is greater than zero due to Eq (5), the empirical risk becomes greater if the empirical neutrality risk becomes lower.

6 Neutral SVM

6.1 Primal Problem

SVMs [14] are a margin-based supervised learning method for binary classification. The algorithm of SVMs can be interpreted as minimization of the empirical risk with regularization term, which follows the RERM principle. In this section, we introduce a SVM variant that follows the NERM principle.

The soft-margin SVM employs the linear classifier $f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b$ as the target hypothesis. In the objective function, the hinge loss is used for the loss function, as $\phi(yf(x)) = \max(0, 1 - yf(x))$, and the ℓ_2 norm is used for the regularization term, $\Omega(f) = \lambda \|f\|_2^2 / 2n$, where $\lambda > 0$ denotes the regularization parameter. In our SVM in NERM, referred to as the neutral SVM, the loss function and regularization term are the same as in the soft-margin SVM. For a surrogate function of the neutralization term, the hinge loss $\psi(\pm g(x)f(x)) = \max(0, 1 \mp g(x)f(x))$ was employed. Any hypothesis can be used for the viewpoint hypothesis. Accordingly, following the NERM principle defined in Eq (4), the neutral SVM is formulated by

$$\min_{\boldsymbol{w}, b} \sum_{i=1}^{n} \max(0, 1 - y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b)) + \frac{\lambda}{2} \|\boldsymbol{w}\|_2^2 + \eta C_{n,\psi}(\boldsymbol{w}, b, g), \tag{7}$$

where

$$\begin{aligned} C_{n,\psi}(\boldsymbol{w}, b, g) &= \max(C_{n,\psi}^{+}(\boldsymbol{w}, b, g), C_{n,\psi}^{-}(\boldsymbol{w}, b, g)), \\ C_{n,\psi}^{\pm}(\boldsymbol{w}, b, g) &= \sum_{i=1}^{n} \max(0, 1 \mp g(\boldsymbol{x}_i)(\boldsymbol{w}^T \boldsymbol{x}_i + b)). \end{aligned}$$

Since the risk, regularization, and neutralization terms are all convex, the objective function of the neutral SVM is convex. The primal form can be solved by applying the subgradient method [13] to Eq (7).

6.2 Dual Problem and Kernelization

Next, we derive the dual problems of the problem of Eq (7), from which the neutral SVM kernelization is naturally derived. First, we introduce slack variables $\boldsymbol{\xi}, \boldsymbol{\xi}^{\pm}$, and ζ into Eq (7) to represent the primal problem:

$$\begin{aligned} \min_{\substack{\boldsymbol{w},b,\\ \boldsymbol{\xi},\boldsymbol{\xi}^{\pm},\zeta}} \quad & \sum_{i=1}^{n} \xi_i + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 + \eta\zeta \qquad (8) \\ \text{sub to} \quad & \sum_{i=1}^{n} \xi_i^+ \leq \zeta, \sum_{i=1}^{n} \xi_i^- \leq \zeta, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \leq \xi_i, \\ & 1 - v_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \leq \xi_i^+, 1 + v_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \leq \xi_i^-, \\ & \xi_i \geq 0, \xi_i^+ \geq 0, \xi_i^- \geq 0, \zeta \geq 0 \end{aligned}$$

where slack variables ξ_i, ξ_i^+, and ξ_i^- denote measures of the degree of misclassification, correlation, and inverse correlation, respectively. The slack variable ζ, derived from max function in $C_{n,\psi}(\boldsymbol{w}, b, g)$, measures the imbalance of the degree of correlation and inverse correlation. From the Lagrange relaxation of the primal problem Eq (8), the dual problem is derived as

$$\begin{aligned} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}^{\pm}} \quad & \lambda \sum_{i=1}^{n} b_i - \frac{1}{2}\sum_{i}^{n}\sum_{j}^{n} a_i a_i k(x_i, x_j) \qquad (9) \\ \text{sub to} \quad & \sum_{i}^{n} a_i = 0, 0 \leq \alpha_i \leq 1, 0 \leq \beta_i^{\pm}, \beta_i^+ + \beta_i^- \leq \eta \end{aligned}$$

where $b_i = \alpha_i + \beta_i^+ + \beta_i^-, a_i = \alpha_i y_i + \beta_i^+ v_i - \beta_i^- v_i$. As seen in the dual problem, the neutral SVM is naturally kernelized with kernel function $\boldsymbol{x}_i^T\boldsymbol{x}_j = k(x_i, x_j)$. The derivation of the dual problem and kernelization thereof is described in the supplemental document in detail. The optimization of Eq (9) is an instance of *quadratic programming (QP)* that can be solved by general QP solvers, although it does not scale well with large samples due to its large memory consumption. In the supplemental documentation, we also show the applicability of the well-known *sequential minimal optimization* technique to our neutral SVM.

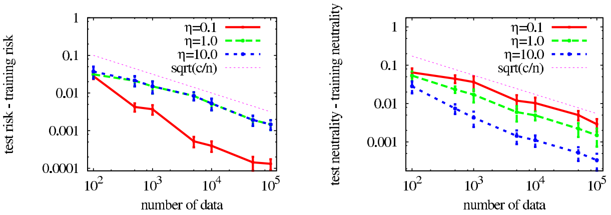
7 Experiments

In this section, we present experimental evaluation of our neutral SVM for synthetic and real datasets. In the experiments with synthetic data, we experimentally evaluate the change of generalization risk and generalization neutrality risk according to the number of samples, in which their relations are described in Theorem 2. In the experiments for real datasets, we compare our method with CV2NB [4], PR [8] and η-neutral logistic regression (ηLR for short) [5] in terms of risk and neutrality risk. The CV2NB method learns a naíve Bayes model, and then modifies the model parameters so that the resultant CV score approaches zero. The PR and ηLR are based on maximum likelihood estimation of a logistic regression (LR) model. These methods have two parameters, the regularizer parameter λ, and the neutralization parameter η. The PR penalizes the objective function of the LR model with mutual information. The ηLR performs maximum likelihood estimation of the LR model while enforcing η-neutrality as constraints. The neutralization parameter of neutral SVM and PR balances risk minimization and neutrality maximization. Thus, it can be tuned in the same manner used to determine the regularizer parameter. The neutralization parameter of ηLR determines the region of the hypothesis in which the hypotheses are regarded as neutral. The tuning strategy of the regularizer parameter and neutralization parameter are different in all these methods. We determined the neutralization parameter tuning range of these methods via preliminary experiments.

7.1 Synthetic Dataset

In order to investigate the change of generalization neutrality risk with sample size n, we performed our neutral SVM experiments for a synthetic dataset. First, we constructed the input $\boldsymbol{x}_i \in \mathbb{R}^{10}$ with the vector being sampled from the uniform distribution over $[-1,1]^{10}$. The target y_i corresponding to the input $\boldsymbol{x}_i$ is generated as $y_i = \mathrm{sgn}(\boldsymbol{w}_y^T \boldsymbol{x}_i)$ where $\boldsymbol{w}_y \in \mathbb{R}^{10}$ is a random vector drawn from the uniform distribution over $[-1,1]^{10}$. Noises are added to labels by inverting the label with probability $1/(1+\exp(-100|\boldsymbol{w}_y^T \boldsymbol{x}_i|))$. The inverting label probability is small if the input $\boldsymbol{x}_i$ is distant from a plane $\boldsymbol{w}_y^T \boldsymbol{x} = 0$. The viewpoint v_i corresponding to the input $\boldsymbol{x}_i$ is generated as $v_i = \mathrm{sgn}(\boldsymbol{w}_v^T \boldsymbol{x}_i)$, where the first element of $\boldsymbol{w}_v$ is set as $w_{v,1} = w_{y,1}$ and the rest of elements are drawn from the uniform distribution over $[-1,1]^9$. Noises are added in the same manner as the target. The equality of the first element of $\boldsymbol{w}_y$ and $\boldsymbol{w}_v$ leads to correlation between y_i and v_i. Set the regularizer parameter as $\lambda = 0.05n$. The neutralization parameter was varied as $\eta \in \{0.1, 1.0, 10.0\}$. In this situation, we evaluate the approximation error of the generalization risk and the generalization neutrality risk by varying sample size n. The approximation error of generalization risk is the difference of the empirical risk between training and test samples, while that of the generalization neutrality risk is the difference of the empirical neutrality risk between training and test samples. Five fold cross-validation was used for evaluation of the approximation error of the empirical risk and empirical neutrality; the average of ten different folds are shown as the results.

Results. Fig 1 shows the change of the approximation error of generalization risk (the difference of the empirical risks w.r.t. test samples and training samples), and the approximation error of generalization neutrality risk (the difference of the empirical neutrality risks w.r.t. test samples and training samples)



(a) risk (b) neutrality risk

Fig. 1. Change of approximation error of generalization risk (left) and approximation error of generalization neutrality risk (right) by neutral SVM (our proposal) according to varying the number of samples n. The horizontal axis shows the number of samples n, and the error bar shows the standard deviation across the change of five-fold division. The line “sqrt(c/n)” denotes the convergence rate of the approximation error of the generalization risk (in Theorem 1) or the generalization neutrality risk (in Theorem 2). Each line indicates the results with the neutralization parameter $\eta \in \{0.1, 1.0, 10.0\}$. The regularizer parameter was set as $\lambda = 0.05n$.

Table 1. Specification of Datasets

dataset	#Inst.	#Attr.	Viewpoint	Target
Adult	16281	13	gender	income
Dutch	60420	10	gender	income
Bank	45211	17	loan	term deposit
German	1000	20	foreign worker	credit risk

with changing sample size n. The plots in Fig 1 left and right show the approximation error of generalization risk and the approximation error of generalization neutrality risk, respectively.

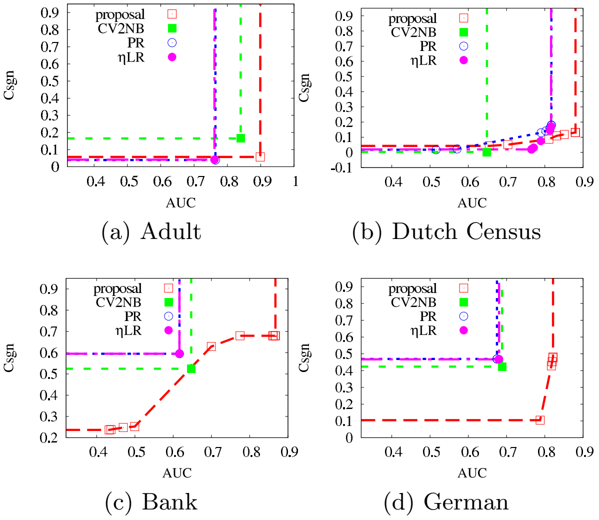
Recall that the discussions in Section 5.3 showed that the approximation error of generalization risk decreases with $O(\sqrt{\ln(1/\delta)/n})$ rate. As indicated by the Theorem 1, Fig 1 (left) clearly shows that the approximation error of the generalization risk decreases as sample size n increases. Similarly, discussions in Section 5.1 revealed that the approximation error of generalization neutrality risk also decreases with $O(\sqrt{\ln(1/\delta)/n})$ rate, which can be experimentally confirmed in Fig 1 (right). The plot clearly shows that the approximation error of the generalization neutrality risk decreases as the sample size n increases.

7.2 Real Datasets

We compare the classification performance and neutralization performance of neutral SVM with CV2NB, PR, and ηLR for a number of real datasets specified in Table 1. In Table 1, #Inst. and #Attr. denote the sample size and the number of attributes, respectively; “Viewpoint” and “Target” denote the attributes used as the target and the viewpoint, respectively. All dataset attributes were discretized by the same procedure described in [4] and coded by 1-of-K representation for PR, ηLR, and neutral SVM. We used the primal problem of neutral

Table 2. Range of neutralization parameter

method	range of neutralization parameter
PR	0, 0.01, 0.05, 0.1, ..., 100
ηLR	0, 5×10^{-5}, 1×10^{-4}, 5×10^{-4}, ..., 0.5
neutral SVM	0, 0.01, 0.05, 0.1, ..., 100



(a) Adult

(b) Dutch Census

(c) Bank

(d) German

Fig. 2. Performance of CV2NB, PR, ηLR, and neutral SVM (our proposal). The vertical axis shows the AUC, and horizontal axis shows $C_{n,sgn}(f,g)$. The points in these plots are omitted if they are dominated by others. The bottommost line shows limitations of neutralization performance, and the rightmost line shows limitations of classification performance, which are shown only as guidelines.

SVM (non-kernelized version) to compare our method with the other methods in the same representation. For PR, ηLR, and neutral SVM, the regularizer parameter was tuned in advance for each dataset in the non-neutralized setting by means of five-fold cross validation, and the tuned parameter was used for the neutralization setting. CV2NB has no regularization parameter to be tuned. Table 2 shows the range of the neutralization parameter used for each method.

The classification performance and neutralization performance was evaluated with *Area Under the receiver operating characteristic Curve* (AUC) and $+1/-1$ empirical neutrality risk $C_{n,\mathrm{sgn}}(f,g)$, respectively. Both measures were evaluated with five-fold cross-validation and the average of ten different folds are shown in the plots.

Results. Fig 2 shows the classification performance (AUC) and neutralization performance ($C_{n,\mathrm{sgn}}(f,g)$) at different setting of neutralization parameter η. In the graph, the best result is shown at the right bottom. Since the classification performance and neutralization performance are in a trade-off relationship, as indicated by Theorem Eq (6), the results dominated by the other parameter settings are omitted in the plot for each method.

CV2NB achieves the best neutrality in Dutch Census, but is less neutral compared to the other methods in the rest of the datasets. In general, the classification performance of CV2NB is lower than those of the other methods due to the poor classification performance of naíve Bayes. PR and ηLR achieve competitive performance to neutral SVM in Adult and Dutch Census in term of the neutrality risk, but the results are dominated in term of AUC. Furthermore, the results of PR and ηLR in Bank and German are dominated. The results of neutral SVM are dominant compared to the other methods in Bank and German dataset, and it is noteworthy that the neutral SVM achieves the best AUC in almost all datasets. This presumably reflects the superiority of SVM in the classification performance, compared to the naíve Bayes and logistic regression.

8 Conclusion

We proposed a novel framework, NERM. NERM provides a framework that learns a target hypothesis that minimizes the empirical risk and that is empirically neutral in terms of risk to a given viewpoint hypothesis. Our contributions are as follows: (1) We define NERM as a framework for guaranteeing the neutrality of classification problems. In contrast to existing methods, the NERM can be formulated as a convex optimization problem by using convex relaxation. (2) We provide theoretical analysis of the generalization neutrality risk of NERM. The theoretical results show the approximation error of the generalization neutrality risk of NERM is uniformly upper-bounded by the Rademacher complexity of hypothesis class $g\mathcal{F}$ and $O(\sqrt{\ln(1/\delta)/n})$. Moreover, we derive a bound on the generalization neutrality risk for the optimal hypothesis corresponding to the neutralization parameter η. (3) We present a specific learning algorithms for NERM, neutral SVM. We also extend the neutral SVM to the kernelized version.

Suppose the viewpoint is set to some private information. Then, noting that neutralization reduces correlation between the target and viewpoint values, outputs obtained from the neutralized target hypothesis do not help to predict the viewpoint values. Thus, neutralization realizes a certain type of privacy preservation. In addition, as already mentioned, NERM can be interpreted as a variant of transfer learning by regarding the neutralization term as data-dependent prior knowledge. Clarifying connection to privacy-preservation and transfer learning is remained as an area of future work.

Acknowledgments. The authors would like to thank Toshihiro Kamishima and the anonymous reviewers for their valuable suggestion. The work is supported by JST CREST program of Advanced Core Technologies for Big Data Integration, JSPS KAKENHI 12913388, and 25540094.

References

1. Bartlett, P.L., Bousquet, O., Mendelson, S.: Local rademacher complexities. The Annals of Statistics 33(4), 1497–1537 (2005)
2. Bartlett, P.L., Mendelson, S.: Rademacher and gaussian complexities: Risk bounds and structural results. Journal of Machine Learning Research 3, 463–482 (2002)
3. Calders, T., Kamiran, F., Pechenizkiy, M.: Building classifiers with independency constraints. In: Saygin, Y., Yu, J.X., Kargupta, H., Wang, W., Ranka, S., Yu, P.S., Wu, X. (eds.) ICDM Workshops, pp. 13–18. IEEE Computer Society (2009)
4. Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery 21(2), 277–292 (2010)
5. Fukuchi, K., Sakuma, J., Kamishima, T.: Prediction with model-based neutrality. In: Blockeel, H., Kersting, K., Nijssen, S., Železný, F. (eds.) ECML PKDD 2013, Part II. LNCS (LNAI), vol. 8189, pp. 499–514. Springer, Heidelberg (2013)
6. Kakade, S.M., Sridharan, K., Tewari, A.: On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (eds.) NIPS, pp. 793–800. Curran Associates, Inc. (2008)
7. Kamiran, F., Calders, T., Pechenizkiy, M.: Discrimination aware decision tree learning. In: 2010 IEEE 10th International Conference on Data Mining (ICDM), pp. 869–874. IEEE (2010)
8. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) ECML PKDD 2012, Part II. LNCS (LNAI), vol. 7524, pp. 35–50. Springer, Heidelberg (2012)
9. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22(10), 1345–1359 (2010)
10. Pariser, E.: The Filter Bubble: What The Internet Is Hiding From You. Viking, London (2011)
11. Pedreschi, D., Ruggieri, S., Turini, F.: Measuring discrimination in socially-sensitive decision records. In: Proceedings of the SIAM Int'l Conf. on Data Mining, pp. 499–514. Citeseer (2009)
12. Resnick, P., Konstan, J., Jameson, A.: Measuring discrimination in socially-sensitive decision records. In: Proceedings of the 5th ACM Conference on Recommender Systems: Panel on The Filter Bubble, pp. 499–514 (2011)
13. Shor, N.Z., Kiwiel, K.C., Ruszcayǹski, A.: Minimization Methods for Non-differentiable Functions. Springer-Verlag New York, Inc., New York (1985)
14. Vapnik, V.N.: Statistical learning theory (1998)
15. Zemel, R.S., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: ICML (3). JMLR Proceedings, vol. 28, pp. 325–333. JMLR.org (2013)
16. Zliobaite, I., Kamiran, F., Calders, T.: Handling conditional discrimination. In: 2011 IEEE 11th International Conference on Data Mining (ICDM), pp. 992–1001. IEEE (2011)