Abstract

Classifiers and rating scores are prone to implicitly codifying biases, which may be present in the training data, against protected classes (i.e., age, gender, or race). So it is important to understand how to design classifiers and scores that prevent discrimination in predictions. This paper develops computationally tractable algorithms for designing accurate but fair support vector machines (SVM’s). Our approach imposes a constraint on the covariance matrices conditioned on each protected class, which leads to a nonconvex quadratic constraint in the SVM formulation. We develop iterative algorithms to compute fair linear and kernel SVM’s, which solve a sequence of relaxations constructed using a spectral decomposition of the nonconvex constraint. Its effectiveness in achieving high prediction accuracy while ensuring fairness is shown through numerical experiments on several data sets.

1 INTRODUCTION

The increasing prevalence of machine learning to automate decision-making systems has drawn societal scrutiny on the approaches used for verification and validation of these systems. In particular, two main concerns have been voiced with regards to the correctness and accuracy of decisions provided. The first is a general lack of interpretability for how predictions are produced by many learning techniques [Ridgeway et al., 1998; Lou et al., 2012; Caruana et al., 2015], and the second is the possibility of perpetuating inequities that may be present in the training data [Fodesta et al., 2014; Barocas and Selbst, 2016; Bhandari, 2016].

Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84. Copyright 2018 by the author(s).
Recall that the essence of the kernel trick is to replace $X(X')^T$ with $K(X,X')$ as part of the kernel trick.

Last, we define some additional notation. Let $|n| = \{1, \ldots, n\}$, and note $1(u)$ is the indicator function. A positive semidefinite matrix $U$ is denoted $U \succeq 0$. If $U, V$ are vectors of equal dimension, then the notation $U \circ V$ refers to their element-wise product: $(U \circ V)_i = U_i \cdot V_i$. Also, $e$ is the vector whose entries are all 1.

### 3 ROC Visualization of Fairness

#### 3.1 Demographic Parity

One popular notion of fairness is that predictions of the label $y$ are independent of the protected class $z$. This definition is typically stated in terms of a single threshold, though it can be generalized to multiple thresholds. We say that a classifier $h(x,t)$ has demographic parity at level $\Delta$ (abbreviated as DP-$\Delta$) if

$$\left| \mathbb{P}[h(x,t) = +1 | z = +1] - \mathbb{P}[h(x,t) = +1 | z = -1] \right| \leq \Delta, \forall t \in \mathbb{R}. \quad (2)$$

To understand this, note $\mathbb{P}[h(x,t) = +1 | z = +1]$ is the true positive rate when predicting the protected class at threshold $t$, while $\mathbb{P}[h(x,t) = +1 | z = -1]$ is the false positive rate when predicting the protected class at threshold $t$. So the intuition is that a classifier is DP-$\Delta$ if its true positive and false positive rates with respect to its ability to predict the protected class are approximately (up to $\Delta$ deviation) equal at all thresholds.

Reinterpreted, demographic parity requires that predictions of the classifier cannot reveal information about the protected class any better (up to $\Delta$ deviation) than random guessing. DP-$\Delta$ is in fact equivalent to requiring that the ROC curve for the classifier $h(x,t)$ in predicting $z$ is within $\Delta$ of the line of no-discrimination, which is the line that is achievable by biased random guessing. More visually, Figure 4 shows how DP-$\Delta$ can be seen using an ROC curve.

#### 3.2 Equal Opportunity

Demographic parity has been criticized as too strict, and so another notion of fairness has been proposed in which predictions of the label $y$ are independent of the protected class $z$, when the true label is positive (i.e., $y = +1$). In this definition, we must interpret $y = +1$ as a better label than $y = -1$; for instance, $y = -1$ may be a loan default, while $y = +1$ is full repayment...
of a loan. This definition is typically stated [Hardt et al., 2016] in terms of a single threshold, though it can be generalized to multiple thresholds. We say that a classifier \( h(x,t) \) has equal opportunity with level \( \Delta \) (abbreviated as EO-\( \Delta \)) if

\[
\left| \Pr[h(x,t) = +1 | z = +1, y = +1] - \Pr[h(x,t) = +1 | z = -1, y = +1] \right| \leq \Delta, \quad \forall t \in \mathbb{R}. \tag{3}
\]

To understand this, note \( \Pr[h(x,t) = +1 | z = +1, y = +1] \) is the true positive rate conditioned on \( y = +1 \) when predicting the protected class at threshold \( t \), while \( \Pr[h(x,t) = +1 | z = -1, y = +1] \) is the false positive rate conditioned on \( y = +1 \) when predicting the protected class at threshold \( t \). So the intuition is that a classifier is EO-\( \Delta \) if its false positive rates and true positive rates are approximately (up to \( \Delta \) deviation) equal at all threshold levels for the protected class, when conditioned on \( y = +1 \).

Reinterpreted, equal opportunity requires that predictions of the classifier cannot reveal information about the protected class any better (up to \( \Delta \) deviation) than random guessing, when the true label is positive. EO-\( \Delta \) is equivalent to requiring that the ROC curve for the classifier \( h(x,t) \) in predicting \( z \) conditioned on \( y = +1 \) is within \( \Delta \) of the line of no-discrimination. Figure 1 shows how DP-\( \Delta \) can be seen using an ROC curve.

4 FAIRNESS CONSTRAINTS

In this section, we derive several fairness constraints for linear SVM. The kernel trick is used to convert these constraints for use in kernel SVM. We will focus on presenting formulations for demographic parity, though all of our formulations easily generalize to equal opportunity by simply conditioning on \( y = +1 \).

4.1 Constraints for Linear SVM

We first study constraints that can be used to ensure fairness with respect to \( z \) for a linear SVM

\[
h(x,b) = \text{sign}(x^Tw + b), \tag{4}
\]

where \( w \in \mathbb{R}^p, b \in \mathbb{R} \) are coefficients to predict the label \( y \). Consider the following generic linear SVM formulation

\[
\begin{align*}
\min_{w,\theta} & \sum_{i=1}^{n} u_i \lambda \|w\|^2_2 \\
\text{s.t.} & \quad y_i(x_i^Tw + b) \geq 1 - u_i, \quad \text{for } i \in [n] \\
& \quad u_i \geq 0, \quad \text{for } i \in [n] \\
& \quad G(w,\theta) \leq 0.
\end{align*} \tag{5}
\]

where \( \lambda \in \mathbb{R} \) is a tuning parameter that can be chosen using cross-validation, and \( G(w,\theta) \leq 0 \) is a fairness constraint with the fairness level controlled by the (possibly vector-valued) parameter \( \theta \). We will next consider several possibilities for \( G(w,\theta) \leq 0 \).

4.1.1 Indicator Constraints

The definition of DP-\( \Delta \) in (2) uses probabilities, which are are not available as data. Fortunately, we can use empirical fractions of events as an approximation:

\[
\begin{align*}
\left| \frac{1}{\#P} \sum_{i \in P} \mathbf{1}(\text{sign}(x_i^Tw + t) = +1) - \\
\frac{1}{\#N} \sum_{i \in N} \mathbf{1}(\text{sign}(x_i^Tw + t) = +1) \right| \leq \Delta, \quad \forall t \in \mathbb{R}. \tag{6}
\end{align*}
\]

Standard arguments using VC dimension [Wainwright, 2017] can be used to show that (6) bounds (3) with high probability. However, (6) is also difficult to include in the linear SVM as the fairness constraint \( G(w,\theta) \leq 0 \) since it involves the discontinuous and nonconvex \( \text{sign}(\cdot) \) function, and is infinite-dimensional since it must hold for all \( t \in \mathbb{R} \).

4.1.2 Integer Constraints

An initial idea to incorporate (6) in the linear SVM is based on recent empirical results [Miyashiro and Takano, 2015] [Bertsimas et al., 2016] where mixed integer programming (MIP) was used to exactly solve cardinality constrained linear regression. Specifically,
we can approximate the indicator constraints [8] as the following mixed-integer linear inequalities
\[ -\Delta \leq \frac{1}{w^T} \sum_{i \in P} v_i(t) - \frac{1}{w^T} \sum_{i \in N} v_i(t) \leq \Delta \]
\[ -M \cdot (1 - v_i(t)) \leq x_i^T w + t \leq M \cdot v_i(t) \]  
for \( t \in \{t_1, \ldots, t_k\} \); where \( M > 0 \) is a large constant, and \( \{t_1, \ldots, t_k\} \) is a fixed set of values. This removes the \( \text{sign}(\cdot) \) function, and it ensures a finite number of constraints. However, we found in numerical experiments using [Gurobi, 2016] and [Mosek, 2017] that computing a fair linear SVM with the above constraints was prohibitively slow except for very small data sets.

### 4.1.3 Convex Relaxation

We next derive a convex relaxation of the indicator constraints [8]. Let \( M > 0 \) be a constant, and consider
\[ -M \leq u \leq M. \]
Then convex upper bounds for the indicators functions are \( 1(\text{sign}(u) = +1) \leq 1 + u/M \) and \( -1(\text{sign}(u) = +1) \leq -u/M \). Using these upper bounds on [8] leads to the following convex relaxation:
\[ -d \leq \left( \frac{1}{w^T} \sum_{i \in P} x_i - \frac{1}{w^T} \sum_{i \in N} x_i \right)^T w \leq d, \]  
where \( d = M \cdot (\Delta - 1) \). There are three important remarks about this convex relaxation. The first is that the threshold \( t \) does not appear, which means this is simply a linear constraint. The second is that the bound \( M \) can be subsumed into the parameter \( d \) used to control the fairness level, meaning that this relaxation is practically independent of the bound \( M \). The third is that this convex relaxation is equivalent to the fairness constraint proposed by [Zafar et al., 2017], though they derive this using a correlation argument.

### 4.1.4 Covariance Constraints

The convex relaxation [8] is fairly weak, and so it is relevant to ask whether constraints can be added to [8] to improve the fairness level \( \Delta \) of the resulting linear SVM. Instead of using convex lifts [Lasserre, 2001, Gouveia et al., 2013, Chandrasekaran and Jordan, 2013] to tighten the above constraint, our approach is inspired by an information-theoretic bound and has an intuitive geometric interpretation.

Specifically, consider the two conditional distributions \( D_+ = \{|x|z = +1\} \) and \( D_- = \{|x|z = -1\} \). Equation [2] can be interpreted as the Kolmogorov-Smirnov (KS) distance between the distributions of the “score” functions \( w^T D_+ \) and \( w^T D_- \). Since KS distance is upper-bounded by the total variation distance, Pinsker’s inequality [Massart, 2007] implies
\[ \Delta \leq \sqrt{\frac{1}{2} \mathcal{K}(w^T D_+||w^T D_-)}, \]
where \( \mathcal{K}(\cdot||\cdot) \) is the Kullback-Leibler divergence. In the special case where \( D_+ \sim \mathcal{N}(\mu_+, \Sigma_+) \) and \( D_- \sim \mathcal{N}(\mu-, \Sigma_-) \), we have
\[ \Delta \leq \sqrt{\frac{1}{4} \left( \frac{\sigma_-}{\sigma_+} + \frac{1}{\sigma_+} (m_+ - m_-)^2 + \ln \frac{\sigma_+}{\sigma_-} - 1 \right)}, \]  
where \( m_+ = w^T \mu_+, m_- = w^T \mu-, \sigma_+ = w^T \Sigma_+ w \) and \( \sigma_- = w^T \Sigma_- w \) [Kullback, 1997].

Note the above bound is minimized if \( m_+ = m_- \) and \( \sigma_+ = \sigma_- \). In fact, [8] can be interpreted as requiring
\[ -d \leq \mathbb{E}(w^T D_+^T w - \mathbb{E}(w^T D_-^T w) \leq d, \]  
or that \( -d \leq m_+ - m_- \leq d \) in the special case above. Subsequently, we propose constraints to control the difference between conditional variances of \( x^T w + t \), which would make \( \sigma_+ \) close to \( \sigma_- \) in the above special case. Let \( \Sigma_+ \) and \( \Sigma_- \) be the sample covariance matrices for \( D_+ \) and \( D_- \), respectively. Then the sample variances of \( w^T D_+ \) and \( w^T D_- \) are \( w^T \Sigma_+ w \) and \( w^T \Sigma_- w \), respectively. So we specify our covariance constraint as
\[ -s \leq w^T (\Sigma_+ - \Sigma_-) w \leq s. \]  
To our knowledge, this constraint has not been previously used to improve the fairness of classifiers. Unfortunately, it is nonconvex because \( \Sigma_+ - \Sigma_- \) is symmetric but typically indefinite (i.e., not positive or negative semidefinite). Hence, computing a linear SVM with this constraint requires further development.

One obvious approach is to lift the constraint [11] and then construct a semidefinite programming (SDP) relaxation [Goemans and Williamson, 1995, Luo et al., 2010]. Specifically, note that [11] is equivalent to
\[ -s \leq \text{trace}(W(\Sigma_+ - \Sigma_-)) \leq s \]
\[ U = \begin{bmatrix} W & w^T \\ w & 1 \end{bmatrix} \succeq 0 \]  
\[ \text{rank}(U) = 1 \]  
The above is nonconvex, but it can be convexified by dropping the \( \text{rank}(U) = 1 \) constraint. However, we found in numerical experiments using the [Mosek, 2017] solver that the SDP relaxation was weak and did not consistently affect the fairness or accuracy of the SVM. Despite this result, we believe that additional convexification techniques [Kocuk et al., 2016, Madani et al., 2017] can be used to strengthen the quality of the SDP relaxation; we leave the problem of how to design a strengthened SDP relaxation for future work.

### 4.2 Constraints for Kernel SVM

We next briefly present constraints analogous to [8] and [11] that can be used to ensure fairness with respect to \( z \) for a kernel SVM
\[ h(x, b) = \text{sign}(K(X, x)^T(Y \circ \alpha) + b), \]  

where \( K \) is the kernel function.
where \( \alpha \in \mathbb{R}^n \) are coefficients to predict the label \( y \), and \( b = \frac{1}{\#P} \sum_{i \in P} (y_i - K(X, x_i)\alpha)^T (Y \circ \alpha) \) with the set of indices \( I = \{ i : 0 < \alpha_i < \lambda \} \). Consider the following general kernel SVM formulation:

\[
\begin{align*}
\min & \quad (Y \circ \alpha)^T K(X, X)(Y \circ \alpha) - \sum_{i=1}^n \alpha_i \\
\text{s.t.} & \quad Y^T \alpha = 0 \\
& \quad 0 \leq \alpha_i \leq \lambda, \quad \text{for } i \in [n] \\
& \quad H(w, \theta) \leq 0.
\end{align*}
\]

(14)

where \( \lambda \in \mathbb{R} \) is a tuning parameter that can be chosen using cross-validation, and \( H(w, \theta) \leq 0 \) is a fairness constraint with the fairness level controlled by the (possibly vector-valued) parameter \( \theta \). We will next consider several possibilities for \( H(w, \theta) \leq 0 \).

First, we note that we may use a convex relaxation on indicator functions akin to that presented in Sections 4.1.2 and 4.1.3 to obtain the linear constraint:

\[
\begin{align*}
-d \leq \frac{1}{\#P} \sum_{i \in P} K(X, x_i)^T (Y \circ \alpha) + \\
& \quad -\frac{1}{\#N} \sum_{i \in N} K(X, x_i)^T (Y \circ \alpha) \leq d. \\
\end{align*}
\]

(15)

Note again that threshold \( t \) does not appear.

Next, our covariance constraint can be rewritten for the kernel SVM by first recalling that the kernel SVM with \( K(x, x') = x^T x' \) generates the same classifier as directly solving a linear SVM. Thus, we have the relationship \( w = X^T (Y \circ \alpha) \) in this special case. Next, observe that:

\[
\begin{align*}
\Sigma_+ &= \frac{1}{\#P} X^T (I - \frac{1}{\#P} ee^T) X_+ \\
\Sigma_- &= \frac{1}{\#N} X^T (I - \frac{1}{\#N} ee^T) X_- \\
\end{align*}
\]

(16)

So if apply the kernel trick to our covariance constraint with the above relationships, then the resulting covariance constraint for kernel SVM becomes:

\[
\begin{align*}
-s \leq (Y \circ \alpha)^T (S_+ - S_-)(Y \circ \alpha) \leq s,
\end{align*}
\]

(17)

where we have:

\[
\begin{align*}
S_+ &= \frac{1}{\#P} K(X, X_+)(I - \frac{1}{\#P} ee^T) K(X, X_+)^T \\
S_- &= \frac{1}{\#N} K(X, X_-)(I - \frac{1}{\#N} ee^T) K(X, X_-)^T.
\end{align*}
\]

(18)

Similar to (11), the constraint (17) is a nonconvex quadratic constraint because \( (S_+ - S_-) \) is symmetric but typically indefinite.

5 SPECTRAL ALGORITHM

The covariance constraints are conceptually promising, but they result in nonconvex optimization problems. Here, we describe an iterative algorithm to effectively solve the SVM with our covariance constraints.

5.1 Linear SVM

Though we could compute the linear SVM with the covariance constraint (11) using successive linearization, a better approach is possible through careful design of the algorithm: Our key observation regarding (11) is that \( (\Sigma_+ - \Sigma_-) \) is symmetric, which means it can be diagonalized by an orthogonal matrix:

\[
\begin{align*}
\Sigma_+ - \Sigma_- &= \sum_{i=1}^{p} \zeta_i v_i v_i^T, \\
\end{align*}
\]

(19)

where \( \zeta_i \in \mathbb{R} \) and the \( v_i \) form an orthonormal basis. Now let \( I_{\zeta+} = \{ i : \zeta_i > 0 \} \) and \( I_{\zeta-} = \{ i : \zeta_i < 0 \} \), and define the positive semidefinite matrices:

\[
\begin{align*}
U_{\zeta+} &= \sum_{i \in I_{\zeta+}} \zeta_i v_i v_i^T, \\
U_{\zeta-} &= -\sum_{i \in I_{\zeta-}} \zeta_i v_i v_i^T
\end{align*}
\]

(20)

This means that the function:

\[
\begin{align*}
w^T (\Sigma_+ - \Sigma_-) w &= w^T U_{\zeta+} w - w^T U_{\zeta-} w
\end{align*}
\]

(21)

in the covariance constraint (11) is the difference of the two convex functions \( w^T U_{\zeta+} w \) and \( w^T U_{\zeta-} w \).

There is an important point to note regarding the practical importance of the spectral decomposition we performed above. The function in the covariance constraint (11) can alternatively be written as the difference of two convex functions \( w^T \Sigma_+ w \) and \( w^T \Sigma_- w \). However, using this alternative decomposition yields an algorithm where the convexified subproblems are weaker relaxations than the convex subproblems generated using the spectral decomposition. As a result, our algorithm given below is ultimately more effective because it employs the spectral decomposition.

Consequently, the constrained convex-concave procedure (Thu, 1995; Yuille and Rangarajan, 2002; Smola et al., 2005) can be used to design an algorithm for our setting. We opt to use a penalized form of the quadratic constraint in our spectral algorithm to ensure feasibility always holds. Let \( w_k \in \mathbb{R}^p \) be a fixed point, and consider the optimization problem where the concave terms are linearized:

\[
\begin{align*}
\min & \quad \sum_{i=1}^n u_i + \lambda \|w\|^2_2 + \mu \cdot t \\
\text{s.t.} & \quad y_i (x_i^T w + b) \geq 1 - u_i, \quad \text{for } i \in [n] \\
& \quad u_i \geq 0, \quad \text{for } i \in [n] \\
& \quad -d \leq \left( \frac{1}{\#P} \sum_{i \in P} x_i - \frac{1}{\#P} \sum_{i \in N} x_i \right)^T w \leq d \\
& \quad w^T U_{\zeta+} + w^T U_{\zeta-} - w^T U_{\zeta+} w_k - 2w_k^T U_{\zeta-} (w - w_k) \leq t \\
& \quad w^T U_{\zeta-} - w^T U_{\zeta+} - w_k^T U_{\zeta+} (w - w_k) \leq t
\end{align*}
\]

(22)

Our spectral algorithm for computing a fair linear SVM consists of the constrained CCP adapted to the problem of computing a linear SVM with the linear
constraint \(S\) and covariance constraint \(U\): We initialize \(w_0\) by solving a linear SVM with only the linear constraint \(S\), and then compute successive \(w_k\) by solving \(\mathcal{K}\). This produces a local minimum.

**Theorem 1** (Smola et al. 2005). The spectral algorithm defined above for computing a fair linear SVM gives iterates \(w_k\) that converge to a local minimum.

This theorem is simply an application of a theorem by Smola et al. 2005, and the constraint qualification required by this theorem trivially holds in our case because all of our convex constraints are linear.

### 5.2 Kernel SVM

We can apply a similar argument to the kernel SVM formulation by instead defining the decomposition \(S_+ - S_- = \sum_{i=1}^p \xi_i \nu_i \nu_i^T\), where \(\xi_i \in \mathbb{R}\) and the \(\nu_i\) form an orthonormal basis. Letting \(I_{\xi+} = \{i : \xi_i > 0\}\) and \(I_{\xi-} = \{i : \xi_i < 0\}\), we may define the positive semidefinite matrices

\[
U_{\xi+} = \sum_{i \in I_{\xi+}} \xi_i \nu_i \nu_i^T, \quad U_{\xi-} = -\sum_{i \in I_{\xi-}} \xi_i \nu_i \nu_i^T.
\]

Finally, we may implement the constrained CCP [Tuy 1995, Yuille and Rangarajan 2002, Smola et al. 2005] with penalized quadratic term in the same manner as above. We use a penalized form of the quadratic constraint in our spectral algorithm to ensure that feasibility always holds.

Letting \(\alpha_k \in \mathbb{R}^n\) be a fixed point, we obtain the analogous kernel formulation

\[
\begin{align*}
\min \ (Y \circ \alpha)^T \mathcal{K}(X, X)(Y \circ \alpha) - \sum_{i=1}^n \alpha_i + \mu \cdot t \\
\text{s.t.} \ Y^T \alpha = 0 \\
0 \leq \alpha_i \leq \lambda, \quad \text{for } i \in [n] \\
- \frac{1}{n} \sum_{i \in P} \mathcal{K}(X, x_i)^T (Y \circ \alpha) + (Y \circ \alpha)^T U_{\xi+} (Y \circ \alpha) - (Y \circ \alpha_k)^T U_{\xi-} (Y \circ \alpha_k) + 2 (Y \circ \alpha_k)^T U_{\xi+} (Y \circ \alpha_k) + 2 (Y \circ \alpha_k)^T U_{\xi-} (Y \circ \alpha - \alpha_k) & \leq t \quad (24)
\end{align*}
\]

We initialize \(w_0\) by solving a kernel SVM with only the linear constraint \(S\), and then computing successive \(w_k\) by solving \(\mathcal{K}\). This produces a local minimum.

**Theorem 2** (Smola et al. 2005). The spectral algorithm defined above for computing a fair kernel SVM gives iterates \(w_k\) that converge to a local minimum.

### 5.3 Generalization to Linear Learners

Our spectral algorithm can be applied to any linear classifier or linear regression model. For instance, in logistic regression we model the conditional probability of \(y_i\) given \(x_i\), as \(P[y_i|x_i] = (1+\exp(-y_i(w^T x_i+b)))^{-1}\). The maximum likelihood estimate (MLE) is computed by solving \(\min \sum_{i=1}^n \exp(-y_i(w^T x_i+b))\), and we can improve the fairness of the MLE by adding our con-
strains (8) and (11). This leads to the following formulation for fair logistic regression:

\[
\begin{align*}
\min & \quad \sum_{i=1}^{n} \exp(-y_i(w^T x_i + b)) \\
\text{s.t.} & \quad -d \leq \left( \frac{1}{|P|} \sum_{i \in P} x_i - \frac{1}{|N|} \sum_{i \in N} x_i \right)^T w \leq d \\
& \quad -s \leq w^T (\Sigma_+ - \Sigma_-) w \leq s
\end{align*}
\] (25)

Since (11) is nonconvex, we use our spectral algorithm to find a local minimum of the above problem. Let \(w_k \in \mathbb{R}^p\) be a fixed point, and consider the optimization problem where the concave terms are linearized:

\[
\begin{align*}
\min & \quad \sum_{i=1}^{n} \exp(-y_i(w^T x_i + b)) + \mu \cdot t \\
\text{s.t.} & \quad -d \leq \left( \frac{1}{|P|} \sum_{i \in P} x_i - \frac{1}{|N|} \sum_{i \in N} x_i \right)^T w \leq d \\
& \quad w^T U_+ w - w^T U_- w_k - 2w_k^T U_-^T (w - w_k) \leq t \\
& \quad w^T U_- w - w_k^T U_+ w - 2w_k^T U_+^T (w - w_k) \leq t
\end{align*}
\] (26)

where \(U_+, U_-\) are as defined in Section 5.1. Our spectral algorithm is: We initialize \(w_0\) by solving the MLE for logistic regression with only the linear constraint (8), and then computing successive \(w_k\) by solving (26). This produces a local minimum as discussed above.

6 NUMERICAL EXPERIMENTS

We use synthetic and real datasets to evaluate the efficacy of our approach. We compare linear SVM’s computed using our spectral algorithm (SSVM) to a standard linear SVM (LSVM) and a linear SVM computed using the approach of Zafar et al. [2017] (ZSVM), since this is the only existing approach that to our knowledge is designed to ensure fairness at all thresholds. Numerical implementations are available at http://github.com/molfat66/FairML.

6.1 Synthetic Data

Experimental Design We seek to examine the case in which our \(y\) and \(z\) labels are correlated, and \(X_N\) and \(X_P\) have differing covariances. Thus, we generate 200 data points where the \(y\) and \(z\) labels are generated through logit models using two separate sets of randomly generated “true” parameters, with dot product between the logit parameters of \(y\) and \(z\) of 0.85. The singular values of the covariance matrix of \([X_i|z_i=1]\) were then skewed to generate the data seen in Figure 2.The empirical correlation of \(y_i\) and \(z_i\) is 0.45.

Results The results of the three methods using \(d = 0.075\) and \(\mu = 10\) are shown in the three columns of Figure 2. Here, points in \(N\) and \(P\) are differentiated by marker shape (“x” and “o”, respectively), and points with label \(y_i = -1\) and \(y_i = 1\) are differentiated by color (red and green, respectively). If \(w\) denotes the coefficients computed by each method, then the second row shows the empirical densities of \(x_i^Tw\) conditioned on the protected class \(z_i\), and the third row shows the empirical densities of \(x_i^Tw\) conditioned on the label \(y_i\). Fairness occurs if the conditional densities in the second row are similar, and prediction accuracy occurs if the densities in the third row are disparate. These results show that the densities of \([x_i^Tw|z_i=+1]\) and \([x_i^Tw|z_i=-1]\) are distinguishable for LSVM and ZSVM, while they are almost identical for SSVM. On the other hand, the densities of \([x_i^Tw|y_i=+1]\) and \([x_i^Tw|y_i=-1]\) are distinct for all three methods.

6.2 Real World Datasets

Data overview We next use a wine quality dataset [Cortez et al., 2009] and a dataset of German credit card customers [Lichman, 2013]. The first dataset is a compilation of 12 attributes of 6,497 wines (e.g., acidity, residual sugar, alcohol content, and color), as well as a ranking out of 10 that is provided by professional taste-testers. Here, we label \(y_i\) = 1 when a wine is rated as a 6 or above and \(y_i = -1\) otherwise, and we define \(z_i = 1\) for white wines and \(z_i = -1\) for reds. Notably, all explanatory variables are continuous. The second dataset is a compilation of 20 attributes (e.g., marriage, employment and housing status, number of existing credit lines, and age) of 1000 German applicants for loans. We label \(y_i = 1\) for applicants that...
defaulted and \( y_i = -1 \) for applicants that did not default, and let \( z_i = 1 \) for applicants that are renting a home and \( z_i = -1 \) for applicants that own their home. Note that a large number of variables are discrete. There is no missing data in either dataset.

**Metrics of comparison** We compare SSVM and ZSVM based on the tradeoffs that they make between predictive accuracy for \( y \), measured using the area under the ROC curve (AUC), and their fairness with respect \( z \), measured by DP-\( \Delta \) with respect to \( z \).

**Experimental Design** We conducted five rounds of cross-validation on each dataset and computed the average AUC and average \( \Delta \), using a 70-30 training-testing split. Within each round, we first apply 5-fold cross-validation on LSVM to choose the \( \lambda \) that maximizes AUC, and this value of \( \lambda \) was used with both SSVM and ZSVM to minimize the impact of cross-validation on the comparison between methods. We varied \( d \) over the values 0, 0.001, 0.002, 0.005, 0.01, 0.025, 0.05 and 0.1. And for SSVM we tried several values of \( \mu \), which are shown in our plots.

**Results** Figure 3 shows representative examples of ROC curves for both datasets from one instance of cross-validation. Both ZSVM and SSVM improve fairness with respect to LSVM while maintaining high accuracy, and SSVM ensures an even stricter level of fairness than LSVM while keeping high accuracy. The tradeoff curves between prediction accuracy and fairness are shown in Figure 4. Increasing \( d \) generally jointly decreases fairness and increases accuracy, while small increases in \( \mu \) for our SSVM can often improve both fairness and accuracy. Large increases in \( \mu \) generally increase fairness but decrease accuracy. Note that setting \( \mu = 0 \) leads to the curve for SSVM to align with the curve of ZSVM, since they are equivalent when \( \mu = 0 \). Also note \( \Delta \) is more sensitive than AUC to changes in \( d, \mu \), which implies we are able control fairness without losing much predictive accuracy.

**7 Conclusion**

We considered multi-threshold notions of fairness for classifiers, and designed a nonconvex constraint to improve the fairness of linear and kernel SVM’s under all thresholds. We developed an iterative optimization algorithm (that uses a spectral decomposition) to handle our nonconvex constraint in the resulting problem to compute the SVM, and empirically compared our approach to standard linear SVM and an SVM with a linear fairness constraint using both synthetic and real data. We found that our method can strictly improve the fairness of classifiers for all thresholding values with little loss in accuracy; in fact, some of our results even showed a slight increase in accuracy with increasing fairness. Our work opens the door for further research in a number of areas, including hierarchies of fairness constraints considering subsequent moments of the data, and theoretical guarantees on the fairness of such classification methods.

**Acknowledgements**

This work was supported by the UC Berkeley Center for Long-Term Cybersecurity.
References


E. Bhandari. Big data can be used to violate civil rights laws, and the FTC agrees, 2016.


