On the Value of Pairwise Constraints in Classification and Consistency

Jian Zhang JIANZHAN@STAT.PURDUE.EDU
Department of Statistics, Purdue University - West Lafayette, IN 47907 USA

Rong Yan YANR@US.IBM.COM
IBM T.J. Watson Research Center, 19 Skyline Dr., Hawthorne, NY 10523 USA

Abstract

In this paper we consider the problem of classification in the presence of pairwise constraints, which consist of pairs of examples as well as a binary variable indicating whether they belong to the same class or not. We propose a method which can effectively utilize pairwise constraints to construct an estimator of the decision boundary, and we show that the resulting estimator is sign-insensitive consistent with respect to the optimal linear decision boundary. We also study the asymptotic variance of the estimator and extend the method to handle both labeled and pairwise examples in a natural way. Several experiments on simulated datasets and real world classification datasets are conducted. The results not only verify the theoretical properties of the proposed method but also demonstrate its practical value in applications.

1. Introduction

How to effectively learn a classifier with insufficient training data is an important problem in machine learning. One way to address this problem is to integrate new information sources that are complementary to the insufficient training data. Here we are interested in incorporating additional pairwise constraints to improve the classification performance. To be more specific, a pairwise constraint between two examples describes whether those two examples belong to the same class or not. In many real-world applications, pairwise constraints can be obtained automatically or using only a little human effort (Yan et al., 2006) . A large body of previous studies managed to empirically demonstrate the values of pairwise constraints in diverse domains, such as image segmentation (Yu & Shi, 2001), clustering (Wagstaff et al., 2001; Xing et al., 2002), video surveillance (Shental et al., 2003b; Yan et al., 2006) and text classification (Basu et al., 2003). However, most existing algorithms can only find a local-optimal solution for the learning problem, and furthermore, this multi-modality phenomenon in optimization often gets worse as more and more pairwise constraints are included. As a result, the performance of the learning method will not be further improved after getting many pairwise constraints, not to mention approaching the global optimal solution.

In this paper we propose a simple method which can effectively utilize those pairwise constraints in binary classification. The method can be very efficiently computed by first minimizing least squares and then applying some additional simple transformation. As a result, our method can easily handle a large number of pairwise constraints and does not suffer from the local minima problem as others (eg. EM) do. More importantly, we prove that the resulting estimator is *sign-insensitive* consistent[1] w.r.t. the optimal linear decision boundary, and its sign can be reliably estimated using very few labeled training examples. We also provide analysis in terms of asymptotic variance and discuss how to effectively combine information from both labeled examples and pairwise constraints.

2. Review

We first review the standard binary classification problem, some learning methods and their properties. Given training examples drawn from an underlying distribution $P_{X,Y}$ which is unknown, the objective of

Appearing in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007. Copyright 2007 by the author(s)/owner(s).

[1] We say that an estimator θ is sign-insensitive consistent if either θ or $-\theta$ is consistent.

classification is to find a predictor $h(x) : \mathbb{R}^p \mapsto \mathbb{R}$ so that the following generalization error is made as small as possible:

$$R = \mathbb{E}_{X,Y}[I(\mathsf{sign}(h(X)) \neq Y)], \quad (1)$$

where $I(.)$ is the set indicator function and $\mathbb{E}_{X,Y}$ is used to denote the expectation w.r.t. the distribution $P_{X,Y}$.

Given a set of i.i.d. samples $(x_1, y_1), \ldots, (x_m, y_m)$ drawn from $P_{X,Y}$ where $x_i \in \mathbb{R}^p$ and $y_i \in \{\pm 1\}$, the following criterion is often used in practice to find a predictor in a pre-specified hypothesis space $\mathcal{H}$:

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \left\{ \frac{1}{m} \sum_{j=1}^{m} L(h(x_j), y_j) + \lambda_m \Omega(h) \right\}, \quad (2)$$

where $L(.,.) : \mathbb{R} \times \{\pm 1\} \mapsto \mathbb{R}^+$ is some convex loss function, $\Omega(h)$ measures the complexity of the predictor h, and $\lambda_m \geq 0$ is used to control the trade-off between the empirical loss and model complexity. Note that equation (2) represents many popular classifiers such as logistic regression, Support Vector Machines, least square classifier, etc, by plugging in a suitable convex loss function $L(.,.)$.

From equation (1) to (2) several things are changed: (a) $P_{X,Y}$ is replaced by the empirical distribution which places weight $1/m$ over each observed example; (b) discontinuous classification error is replaced by a convex surrogate loss function so that the minimization of a convex function can be efficiently and reliably achieved ; (c) an additional term $\Omega(h)$ is added to penalize complex models, which helps to avoid overfitting caused by finite sample size.

More importantly, the replacement of the 0/1 classification error by a convex surrogate loss function L is also well-justified in theory. For example, it has been shown recently (Zhang, 2004; Bartlett et al., 2006) that for many convex loss functions such as the squared loss, hinge loss, logistic loss, exponential loss, etc, $\hat{h}$ (with $\lambda_m = o(1)$) is a consistent estimator of h^*, the Bayes classifier defined by

$$h^* = \underset{h \text{ measurable}}{\arg\inf} \; \mathbb{E}_{X,Y}[I(\mathsf{sign}(h(X)) \neq Y)], \quad (3)$$

and the minimum value is known as the Bayes risk. This result implies that we can minimize some empirical convex loss while still achieving Bayes-risk consistency, and the global minima can now be efficiently obtained since it is a convex optimization problem.

In this paper we consider the situation where we are also provided with n i.i.d.[2] pairwise constraints $(x_1^L, x_1^R, \tilde{y}_1), \ldots, (x_n^L, x_n^R, \tilde{y}_n)$ where $\tilde{y}_i \in \{\pm 1\}$ only indicates whether x_i^L and x_i^R belong to the same class or not. We are interested in whether we can obtain valuable information from those pairwise constraints and if so, how much information can we get out of them. In our setting, we focus on the situation where $n \gg m$, i.e. we assume that it is much easier and cheaper to get examples of pairwise constraints and meanwhile, the amount of labeled sample is quite limited.

We assume that the hypothesis space $\mathcal{H} = \{\mathbf{w} \mid \mathbf{w} \in \mathbb{R}^p\}$ to be the set of linear classifiers, and nonlinear decision boundary can be obtained by using the transformed example $\phi(x)$. We further assume that $\mathbf{w}^* \in \mathcal{H}$ achieves the optimal risk w.r.t. the hypothesis space $\mathcal{H}$.

The analysis in the rest of the paper will be based on the squared loss which is simple and easy to work with analytically. In particular, we have

$$L(h(X), Y) = (h(X) - Y)^2 = (1 - Y h(X))^2$$

where the second equality comes from the fact that $Y \in \{\pm 1\}$.

3. The Method

The basic idea of our method is actually quite simple. We first transform each pair of examples in the original space $\mathcal{X}$ to a single example in a new space $\tilde{\mathcal{X}}$ where a consistent estimator of the optimal decision boundary can be computed. The resulting estimator is then transformed back to $\mathcal{X}$ to obtain an optimal linear decision boundary in the original space.

In what follows we first present a simple method to find an estimator $\hat{\alpha}$ of $\mathbf{w}^*$ by just using examples of pairwise constraints, and in Section 4 we show that our estimator $\hat{\alpha}$ is sign-insensitive consistent to $\mathbf{w}^*$, i.e. either $\hat{\alpha}$ or $-\hat{\alpha}$ is consistent. Note that this is the best possible result we can achieve since only using pairwise constraints it is not possible to identify which side of the decision boundary is positive or negative. This result will then be combined with labeled examples to identify which side of the decision boundary is positive (or negative) with a high probability. As we will see, the method is very simple and does not require any additional computational package beyond a least square solver.

Given some samples of pairwise constraints $(x_1^L, x_1^R, \tilde{y}_1), \ldots, (x_n^L, x_n^R, \tilde{y}_n)$ where $x_1^L, x_1^R, \ldots, x_n^L, x_n^R$ are i.i.d. from P_X and $\tilde{y}_1, \ldots, \tilde{y}_n$ are also i.i.d. samples which are defined as

$$\tilde{y}_i = \begin{cases} +1, & y_i^l = y_i^r, \\ -1, & y_i^l \neq y_i^r. \end{cases}$$

[2] In this paper we assume that within each triple, x_i^R and x_i^L are also identically and independently distributed.

Note that $\tilde{y}_i$ can also be written as $y_i^L y_i^R$ and those true labels $y_1^L, y_1^R, \ldots, y_n^L, y_n^R$ are not observed. For each pair (x_i^L, x_i^R), we define a new vector

$$z_i \equiv z_i(x_i^L, x_i^R) = \overline{\mathsf{vech}}(x_i^L \circ x_i^R) \qquad (4)$$

where $x_i^L \circ x_i^R \in \mathbb{R}^{p \times p}$ is the outer product between two vectors x_i^L and x_i^R, and the operator $\overline{\mathsf{vech}}$ defined on a $p \times p$ square matrix will return a vector of size $p(p+1)/2$. For example, if $\mathbf{B} = [b_{ij}] \in \mathbb{R}^{3 \times 3}$, we have

$$\overline{\mathsf{vech}}(\mathbf{B}) = [b_{11}, b_{12} + b_{21}, b_{13} + b_{31}, b_{22}, b_{23} + b_{32}, b_{33}]^T.$$

Here we use the notation $\overline{\mathsf{vech}}$ since its definition is only slightly different from that of the **vech** operator used in the literature (Harville, 1997): recall that for a symmetric matrix $\mathbf{A} = [a_{ij}] \in \mathbb{R}^{3 \times 3}$, $\mathsf{vech}(\mathbf{A})$ is defined as

$$\mathsf{vech}(\mathbf{A}) = [a_{11}, a_{12}, a_{13}, a_{22}, a_{23}, a_{33}]^T.$$

Formally, for a symmetric $p \times p$ matrix $\mathbf{A} = [a_{ij}]$,

$$\mathsf{vech}(\mathbf{A}) = [a_{11}, a_{12}, \ldots, a_{1p}, a_{22}, \ldots, a_{2p}, \ldots, a_{pp}]^T,$$

i.e. it is a vector of length $p(p+1)/2$ which contains all elements of $\mathbf{A}$ except those "subdiagonal" ones. Now for a $p \times p$ matrix $\mathbf{B}$ the vech operator can be formally defined as

$$\overline{\mathsf{vech}}(\mathbf{B}) = \mathsf{vech}(\mathbf{B} + \mathbf{B}^T - \mathsf{diag}(\mathbf{B}))$$

where $\mathsf{diag}(\mathbf{B})$ returns a diagonal matrix whose i-th diagonal element is b_{ii}. Later we will also use the notation $\mathsf{vec}(\mathbf{B})$ which is a vector obtained by concatenating all column vectors of $\mathbf{B}$ together.

Given the i.i.d. samples of pairwise constraints $(x_1^L, x_1^R, \tilde{y}_1), \ldots, (x_n^L, x_n^R, \tilde{y}_n)$, we first apply the transformation defined in equation (4) to obtain the transformed pairwise samples $(z_1, \tilde{y}_1), \ldots, (z_n, \tilde{y}_n)$ where $z_i \in \mathbb{R}^{p(p+1)/2}$ and $\tilde{y}_i \in \{\pm 1\}$. We will then use the penalized least square method (a.k.a. ridge regression) to compute

$$\hat{\theta}(n, \lambda_n) = \underset{\theta \in \mathbb{R}^{p(p+1)/2}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} (\tilde{y}_i - \theta^T z_i)^2 + \lambda_n ||\theta||^2 \right\}. \qquad (5)$$

The resulting estimator $\hat{\theta}(n, \lambda_n)$ is applied to the inverse operator vech^{-1} to get a $p \times p$ symmetric matrix

$$\hat{\Theta} = \mathsf{vech}^{-1}(\hat{\theta}),$$

whose definition and uniqueness can be easily seen.

In Section 4 we will show that as long as $n \to \infty$ and $\lambda_n = o(1)$, the matrix $\hat{\Theta}$ converges in probability to

$$\hat{\Theta} \xrightarrow{p} \mathbf{w}^* \circ \mathbf{w}^*$$

for some $\mathbf{w}^* \in \mathcal{H}$ that achieves the optimal risk w.r.t. the hypothesis space $\mathcal{H}$. An sign-insensitive estimator of $\mathbf{w}^*$ can then be obtained by applying the eigendecomposition to the symmetric matrix $\hat{\Theta}$. Suppose the largest eigenvalue and its corresponding eigenvector of $\hat{\Theta}$ are s_1 and $\mathbf{u}_1 \in \mathbb{R}^p$ respectively (which can be computed very efficiently), we use

$$\hat{\alpha} = \sqrt{s_1}\mathbf{u}_1$$

as a sign-insensitive estimator of $\mathbf{w}^*$, since either $\hat{\alpha}$ or $-\hat{\alpha}$ will be close to $\mathbf{w}^*$.

To determine the correct sign of $\hat{\alpha}$, we use the following sign estimator which is computed using the set of labeled examples $(x_1, y_1), \ldots, (x_m, y_m)$:

$$\hat{s}(\hat{\alpha}) = \begin{cases} +1, & \sum_{i=1}^{m} I(y_i \hat{\alpha}^T x_i \geq 0) \geq \left\lceil \frac{m}{2} \right\rceil, \\ -1, & \text{otherwise,} \end{cases} \qquad (6)$$

where $\lceil t \rceil$ is the ceil function and it returns the smallest integer value that is greater than or equal to t. It can be found that $\hat{s}(\hat{\alpha})$ can be estimated correctly with a high probability which depends on both m and the optimal risk. The algorithm details are summarized in Algorithm 1.

Algorithm 1 Learning with pairwise constraints

```
1. Given the set of pairwise examples $\mathcal{P} = \{(x_1^L, x_1^R, \tilde{y}_1), \ldots, (x_n^L, x_n^R, \tilde{y}_n)\}$ and the set of labeled examples $\mathcal{L} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$.
2. Transform the pairwise examples to $(z_1, \tilde{y}_1), \ldots, (z_n, \tilde{y}_n)$ such that $z_i = \overline{\mathsf{vech}}(x_i^L \circ x_i^R)$.
3. Compute the (penalized) least square estimator $\hat{\theta}$ for the transformed pairwise examples based on equation (5) and let $\hat{\Theta} = \mathsf{vech}^{-1}(\hat{\theta})$.
4. Compute the largest eigenvalue and the corresponding eigenvector $s_1$ and $\mathbf{u}_1$ of $\hat{\Theta}$, and let $\hat{\alpha} = \sqrt{s_1}\mathbf{u}_1$.
5. Estimate the sign $\hat{s}(\hat{\alpha})$ using the labeled examples $(x_1, y_1), \ldots, (x_m, y_m)$ based on equation (6).
6. Output $\hat{s}(\hat{\alpha})\hat{\alpha}$ as the final classifier.
```

4. Analysis

Without loss of generality we assume that $X = [X_1, \ldots, X_p]^T$ is centered so that we have $\mathbb{E}[X] = \mathbf{0}$ and $\mathbb{E}[X_j^2] = 1$ for $j = 1, \ldots, p$. Consider the least square estimator for the labeled i.i.d. examples

$(x_1, y_1), \ldots, (x_m, y_m)$:

$$\hat{\beta}^{ls}(m, \lambda_m) = \arg\min_{\beta} \left\{ \frac{1}{m} \sum_{i=1}^{m} (y_i - \beta^T x_i)^2 + \lambda_m ||\beta||^2 \right\} \quad (7)$$

As long as $m \to \infty$ and $\lambda_m = o(1)$, it is easy to show that the population least square solution β^{ls} is

$$\beta^{ls} \;=\; \mathbb{E}[XX^T]^{-1}\mathbb{E}[XY] \quad (8)$$

and we know that $\beta^{ls} \in \mathcal{H}$ achieves the optimal risk w.r.t. $\mathcal{H}$.

On the other hand, consider the least square solution of the transformed pairwise examples $(z_1, \tilde{y}_1), \ldots, (z_n, \tilde{y}_n)$:

$$\hat{\theta}^{ls}(n, \lambda_n) = \arg\min_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (\tilde{y}_i - \theta^T z_i)^2 + \lambda_n ||\theta||^2 \right\} \quad (9)$$

Similarly, as long as $n \to \infty$ and $\lambda_n = o(1)$, we obtain the population solution θ^{ls} which can be written as

$$\theta^{ls} = \mathbb{E}[ZZ^T]^{-1}\mathbb{E}[Z\tilde{Y}] \in \mathbb{R}^{p(p+1)/2}. \quad (10)$$

The following key result establishes a simple relationship between β^{ls} and θ^{ls}.

Theorem 1. *Assume that (X, Y), (X^L, Y^L) and (X^R, Y^R) have the same distribution $P_{X,Y}$. Furthermore, assume that (X^L, Y^L) and (X^R, Y^R) are independent. Define $\tilde{Y} = Y^L Y^R$ and $Z = \overline{\mathsf{vech}}(X^L \circ X^R)$ and assume β^{ls} given by equation (8) uniquely. Then we have $\theta^{ls} = \mathsf{vech}(\beta^{ls} \circ \beta^{ls})$.*

Proof.[3] Define the $p^2 \times p(p+1)/2$ matrix $\mathbf{G}_p$ to be the duplication matrix such that $\mathsf{vec}(\mathbf{A}) = \mathbf{G}_p \mathsf{vech}(\mathbf{A})$ for any symmetric matrix $\mathbf{A}$, and define the $p(p+1)/2 \times p^2$ matrix $\mathbf{H}_p$ to be an arbitrary left inverse of $\mathbf{G}_p$ which satisfies $\mathsf{vech}(\mathbf{A}) = \mathbf{H}_p \mathsf{vec}(\mathbf{A})$ for any $p \times p$ symmetric matrix $\mathbf{A}$ (Obviously $\mathbf{H}_p$ is not unique due to the symmetry of $\mathbf{A}$).

Similarly we define the $p(p+1)/2 \times p^2$ **unique** matrix $\overline{\mathbf{H}}_p$ to be the matrix which satisfies $\overline{\mathsf{vech}}(\mathbf{B}) = \overline{\mathbf{H}}_p \mathsf{vec}(\mathbf{B})$ for any $p \times p$ matrix $\mathbf{B}$, and it is easy to verify that $\overline{\mathbf{H}}_p = \mathbf{G}_p^T$.

By the definition of β^{ls} and properties of the vech and vec operators we have

$$\begin{aligned}
&\mathsf{vech}(\beta^{ls} \circ \beta^{ls}) \\
&= \mathsf{vech}(\mathbb{E}[XX^T]^{-1}\mathbb{E}[XY]\mathbb{E}[XY]^T\mathbb{E}[XX^T]^{-1}) \\
&= \mathbf{H}_p \mathsf{vec}(\mathbb{E}[XX^T]^{-1}\mathbb{E}[XY]\mathbb{E}[XY]^T\mathbb{E}[XX^T]^{-1}) \quad (11) \\
&= \mathbf{H}_p (\mathbb{E}[XX^T]^{-1} \otimes \mathbb{E}[XX^T]^{-1}) \mathsf{vec}(\mathbb{E}[XY]\mathbb{E}[XY]^T)
\end{aligned}$$

where $\mathbf{A} \otimes \mathbf{B}$ denotes the Kronecker product of matrix $\mathbf{A}$ and $\mathbf{B}$, and the last equality comes from the fact that $\mathsf{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\mathsf{vec}(\mathbf{B})$.

Since $Z = \overline{\mathbf{H}}_p \mathsf{vec}(X^L \circ X^R)$, we have

$$\begin{aligned}
\mathbb{E}[Z\tilde{Y}] &= \mathbb{E}[\overline{\mathbf{H}}_p \mathsf{vec}(X^L \circ X^R) Y^L Y^R] \\
&= \overline{\mathbf{H}}_p \mathsf{vec}(\mathbb{E}[XY]\mathbb{E}[XY]^T) \quad (12)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[ZZ^T] &= \mathbb{E}[\overline{\mathbf{H}}_p \mathsf{vec}(X^L \circ X^R) \mathsf{vec}(X^L \circ X^R)^T \overline{\mathbf{H}}_p^T] \\
&= \overline{\mathbf{H}}_p \mathbb{E}[\mathsf{vec}(X^L \circ X^R) \mathsf{vec}(X^L \circ X^R)^T] \overline{\mathbf{H}}_p^T \\
&= \overline{\mathbf{H}}_p \mathbb{E}[(X^R \otimes X^L)((X^R)^T \otimes (X^L)^T)] \overline{\mathbf{H}}_p^T \\
&= \overline{\mathbf{H}}_p \mathbb{E}[(X^R (X^R)^T) \otimes (X^L (X^L)^T)] \overline{\mathbf{H}}_p^T \\
&= \overline{\mathbf{H}}_p (\mathbb{E}[XX^T] \otimes \mathbb{E}[XX^T]) \overline{\mathbf{H}}_p^T \quad (13)
\end{aligned}$$

where we used the fact that $\mathsf{vec}(\mathbf{a} \circ \mathbf{b}) = \mathbf{b} \otimes \mathbf{a}$ for column vectors $\mathbf{a}, \mathbf{b}$ and $(\mathbf{A} \otimes \mathbf{B})^T = \mathbf{A}^T \otimes \mathbf{B}^T$ in the third equation and $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$ in the fourth equation.

Next we show that $\mathsf{vech}(\beta^{ls} \circ \beta^{ls})$ is actually the solution of $\mathbb{E}[ZZ^T]\mathbf{w} = \mathbb{E}[Z\tilde{Y}]$. Since for any symmetric matrix $\mathbf{A}$ we have $\overline{\mathbf{H}}_p^T \mathbf{H}_p \mathsf{vec}(\mathbf{A}) = \mathbf{G}_p \mathbf{H}_p \mathsf{vec}(\mathbf{A}) = \mathsf{vec}(\mathbf{A})$ by definition, it follows from equations (11)-(13) that

$$\mathbb{E}[ZZ^T]\mathsf{vech}(\beta^{ls} \circ \beta^{ls}) = \overline{\mathbf{H}}_p \mathsf{vec}(\mathbb{E}[XY]\mathbb{E}[XY]^T) = \mathbb{E}[Z\tilde{Y}].$$

It is easy to see that $\mathbb{E}[XX^T]$ is singular if and only if $\mathbb{E}[ZZ^T]$ is singular, and thus $\mathsf{vech}(\beta^{ls} \circ \beta^{ls})$ is the unique solution. ■

Corollary 2. *Assume that the optimal risk with respect to $\mathcal{H}$ is err^*. Given m labeled examples, the estimator $\hat{s}(\hat{\alpha}(\infty, 0))\hat{\alpha}(\infty, 0)$ achieves the optimal risk err^* with probability $(1 - r)$ where $r = \sum_{k=\lceil m/2 \rceil}^{m} \begin{pmatrix} m \\ k \end{pmatrix} (1 - \mathsf{err}^*)^{m-k} (\mathsf{err}^*)^k$.*

Sketch of Proof. First we have $\hat{\theta} \xrightarrow{p} \theta^{ls}$. Since the function vech^{-1} is continuous we have $\hat{\Theta} \xrightarrow{p} (\beta^{ls} \circ \beta^{ls})$ by the continuous mapping theorem (van der Vaart, 2000). Also, define the *mapping*[4] $g(.) : \mathbb{S}^{p \times p} \mapsto \mathbb{R}^p$ over all symmetric matrices such that $g(\mathbf{A}) = \arg\min ||\mathbf{A} - \mathbf{w}\mathbf{w}^T||_F^2$, where $||.||_F$ denotes the Frobenius norm. Since g is also continuous, we have $\{\pm\hat{\alpha}\} = g(\hat{\Theta}) \xrightarrow{p} \{\pm\beta^{ls}\} = g(\beta^{ls} \circ \beta^{ls})$. The remaining follows from properties of the Binomial distribution. ■

Note that with fixed m, the resulting estimator achieves the optimal risk with a high probability. For example, with $\mathsf{err}^* = 0.1$ and only $m = 10$ labeled examples we have $1 - r = 0.9984$. Furthermore, it is not

[3] See Chapter 16 in (Harville, 1997) for detailed definitions and properties of the notations used here.

[4] Here $g(.)$ maps to a set and we take the distance between two sets to be the minimum distance of any two points from the two sets.

difficult to see that the sign-insensitive consistency result still holds ($\hat{\theta} \to \theta^{ls}$) when the pairwise constraints are not independent, as long as for each pair X_i^L and X_i^R are independent.

4.1. 2-d example

Since readers may not be familiar with the vech related operations used in Theorem 1, we illustrate the relationship between β^{ls} and θ^{ls} by considering the simple case when $p = 2$. Let $(X, Y) \sim P_{X,Y}$ and assume that $X = [X_1, X_2]^T$, $\mathbb{E}[X_1^2] = \mathbb{E}[X_2^2] = 1$ and $\rho = \mathbb{E}[X_1 X_2] \neq \pm 1$ so that $\mathbb{E}[XX^T]$ is non-singular. Let $r_1 = \mathbb{E}[X_1 Y]$ and $r_2 = \mathbb{E}[X_2 Y]$. Based on previous definition in equation (8), we have

$$\begin{aligned}\beta^{ls} &= \begin{bmatrix} \mathbb{E}[X_1X_1] & \mathbb{E}[X_1X_2] \\ \mathbb{E}[X_2X_1] & \mathbb{E}[X_2X_2] \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[X_1Y] \\ \mathbb{E}[X_2Y] \end{bmatrix} \\ &= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = \frac{1}{1-\rho^2} \begin{bmatrix} r_1 - \rho r_2 \\ -\rho r_1 + r_2 \end{bmatrix}.\end{aligned}$$

Similarly, by definition in equation (4) we have $Z = [X_1^L X_1^R, X_1^L X_2^R + X_2^L X_1^R, X_2^L X_2^R]^T$ and $\tilde{Y} = Y^L Y^R$. Furthermore, because $(X^L, Y^L), (X^R, Y^R) \sim P_{X,Y}$ are independent to each other, we have

$$\begin{aligned}\theta^{ls} &= \begin{bmatrix} 1 & 2\rho & \rho^2 \\ 2\rho & 2+2\rho^2 & 2\rho \\ \rho^2 & 2\rho & 1 \end{bmatrix}^{-1} \begin{bmatrix} r_1^2 \\ 2r_1r_2 \\ r_2^2 \end{bmatrix} \\ &= \frac{1}{(1-\rho^2)^2} \begin{bmatrix} 1 & -\rho & \rho^2 \\ -\rho & (1+\rho^2)/2 & -\rho \\ \rho^2 & -\rho & 1 \end{bmatrix} \begin{bmatrix} r_1^2 \\ 2r_1r_2 \\ r_2^2 \end{bmatrix} \\ &= \frac{1}{(1-\rho^2)^2} \begin{bmatrix} (r_1 - \rho r_2)^2 \\ (r_1 - \rho r_2)(-\rho r_1 + r_2) \\ (-\rho r_1 + r_2)^2 \end{bmatrix}.\end{aligned}$$

It can be easily verified that $\theta^{ls} = \mathsf{vech}(\beta^{ls} \circ \beta^{ls})$.

4.2. Asymptotic Variance

Consider the least square estimators $\hat{\beta}^{ls}(m, \lambda_m)$ and $\hat{\theta}^{ls}(n, \lambda_n)$, as defined in equation (7) and (9). We compare their asymptotic variances in this subsection. Assume that $\mathbb{E}[XX^T]$ and $\mathbb{E}[ZZ^T]$ are non-singular and $\lambda_m/\sqrt{m} \to \lambda_0 \geq 0$, from (Knight & Fu, 2000) we know that

$$\sqrt{m}\left(\hat{\beta}^{ls} - \beta^{ls}\right) \xrightarrow{d} N\Big(-\lambda_0 \mathbb{E}[XX^T]^{-1}\beta^{ls}, \sigma_x^2 \mathbb{E}[XX^T]^{-1}\Big),$$

where $\xrightarrow{d}$ stands for convergence in distribution. Similarly for $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ we have

$$\sqrt{n}\left(\hat{\theta}^{ls} - \theta^{ls}\right) \xrightarrow{d} N\Big(-\lambda_0 \mathbb{E}[ZZ^T]^{-1}\theta^{ls}, \sigma_z^2 \mathbb{E}[ZZ^T]^{-1}\Big),$$

where σ_x^2 and σ_z^2 are the variance of the error variables $e_x = y - x^T\beta^{ls}$ and $e_z = \tilde{y} - z^T\theta^{ls}$, respectively. We assume that $\lambda_0 = 0$ and only consider the variance components. Since the true distribution of Y is unknown we are not able to calculate the exact values σ_x^2 and σ_z^2 of e_x and e_z. However, they can be estimated empirically by using the residuals $\hat{e}_x$ and $\hat{e}_z$ for each labeled and pairwise constraint. Furthermore, it is interesting to notice that the variance does not get much worse as the input dimension is increased from p to $p(p+1)/2$.

4.3. Combined Estimator

When m is moderate or large, we may also want to utilize the information contained in $(x_1, y_1), \ldots, (x_m, y_m)$, not just for estimating $\hat{s}(\hat{\alpha}^{ls})$, the sign of $\hat{\alpha}^{ls}$. Consider the least square estimator $\hat{\alpha}^{ls}(n, \lambda_n)$ using the n pairwise samples and the least square estimator $\hat{\beta}^{ls}(m, \lambda_m)$ with the m labeled samples. We construct a new estimator $\hat{\gamma}$ as

$$\hat{\gamma} = \pi \hat{s}(\hat{\alpha}^{ls})\hat{\alpha}^{ls} + (1-\pi)\hat{\beta}^{ls}$$

where $\pi \in [0, 1]$ is the combination weight. The choice of π for such a combination can be determined according to the fact that the optimal (with smallest variance) linear combination of two independent unbiased estimators should have weights inversely proportional to their variances, which is easy to verify.

Here we have $\hat{s}(\hat{\alpha}^{ls})\hat{\alpha}^{ls}$ and $\hat{\beta}^{ls}$ almost unbiased. And also, since $\hat{\alpha}^{ls}$ and $\hat{\beta}^{ls}$ are independent, the two estimators are only weakly dependent through the sign estimator $\hat{s}(\hat{\alpha}^{ls})$. As a result, we can simply select π to be inversely proportional to their estimated variances. Due to space limitation we only investigate the cross validation method to determine π's value here.

5. Experiments

5.1. Simulations

We first use simulated datasets to verify the theoretical properties presented in Section 4. Suppose the true parameter vector is $\mathbf{w} = [0.5, 0, -0.5, -1.0, 2.5]^T$ and we generate $X \sim N(\mathbf{0}, \Sigma)$ where we have $\Sigma_{kl} = \rho^{|k-l|}$ with $\rho = 0.2$ for $k, l = 1, \ldots, p$. Given x_i and $\mathbf{w}$, the true label is generated according to

$$y_i \sim \mathsf{Bernoulli}\left((1 + \exp(-\mathbf{w}^T x_i))^{-1}\right).$$

Note that here we use -1 for negative and $+1$ for positive instead of the default $0, 1$ generated by the Bernoulli distribution. In this case, the population version of $\beta^{ls} = [0.127, 0.0, -0.127, -0.254, 0.635]^T$, which is the same as $\mathbf{w}$ after some rescaling and also achieves the Bayes risk. The number of labeled examples n and the number of pairwise examples m are set to be 50×2^k for $k = 0, 1, \ldots, 8$.

Since we only use m pairwise examples to estimate $\hat{\alpha}$, we assume that we know its correct sign. In our experiments this is done by choosing from $\{\hat{\alpha}, -\hat{\alpha}\}$ which gives better result. We evaluate the two estimators $\hat{\beta}(n, 0)$ and $\hat{\alpha}(m, 0)$ in terms of both the mean square errors

$$||\beta^{ls} - \hat{\beta}(n, 0)||^2, \; ||\beta^{ls} - \hat{\alpha}(m, 0)||^2$$

and the classification errors which are computed by using another 100k random samples. For each value of m, n, the experiment is repeated 100 times to compute the average result.

Figure 1 shows the result of both mean squared error and classification error as we vary m and n. There are several observations from this experiment: (1) both estimators are consistent ($\hat{\alpha}$ is sign-insensitive consistent) and the classification errors converge quickly to the Bayes risk; (2) the value of each pair $(X^L, X^R, \tilde{Y})$ is less than that of a single labeled example, but slightly more than that of a half example.

5.2. People Identification in Video

To examine the performance of the proposed algorithms, we collected two different datasets from a geriatric nursing home surveillance video. One dataset was extracted from a 6 hour long, single day and single view video. The other dataset was extracted from video across 6 consecutive days from the same camera view. Both collections were sampled at a resolution of 320×240 and a rate of 30 frames per second. The moving sequences of subjects were automatically extracted using a background subtraction tracker. The silhouette images, each of which corresponds to the extracted silhouette of a moving subject, are sampled from the tracking sequence every half second. We only keep the images that did not have any foreground segments containing two or more people. Finally, we obtain the single day dataset with 363 silhouette images for 6 subjects, and the multiple day dataset with 1118 silhouette images for 5 subjects.

Because of the relative robustness of color histograms to appearance variations, we represent the silhouette images using a histogram of HSV color spaces in all experiments, where each color channel has a fixed number of 16 bins. Thus we have a total of 48 one-dimensional features in the histogram. For each subject, the color representation is relatively stable in the single day dataset, but it is more diverse in the multiple day dataset which makes learning more difficult.

In the following experiments, we examine whether the proposed algorithm can help to distinguish the most frequently appeared persons from the other persons. In other words, we consider the most common persons as positive examples and all other persons as negative data. We process the dataset as follows: each dataset is first split into two disjoint sets based on temporal order, where the training set contains 25% of all the video sequences. The remaining images are used as test images. Therefore, we have 39 positive examples and 80 negative examples for the single-day dataset, as well as 65 positive examples and 108 negative examples for the multi-day dataset. In addition, we also generate pairwise constraints on the testing set. The number of pairwise constraints is increased from 0 to N, where N was chosen to be 20 in the single day dataset and 40 in the multiple day dataset. These constraints are obtained from users' feedback. Typically, the system gives users the most ambiguous pairs of examples and users provide the label of positive/negative constraints as feedback. The details of identifying pairwise constraints can be found in our previous work (Same Author). For evaluation, the classification error on testing data is reported. We use 3-fold cross validation to choose the combination weight π, which is 0.25 in our case.

Figure 2 compares the effectiveness of the proposed pairwise learning algorithms using a growing amount of pairwise constraints as well as the baseline classifier which only uses the labeled data. We can observe that the classification results in both settings outperform the baseline performance even with a small number of constraints available. In more detail, the classification error is reduced from 23% to 19% with 20 additional constraints in the single-day dataset, and the error is reduced from 34% to 30% with 40 additional constraints in the multi-day dataset. Generally speaking, we can achieve better classification results by adding more pairwise constraints. These observations have confirmed that our method can bring consistent performance improvement by leveraging more constraints.

6. Related Work

Pairwise constraints have been used to improve both supervised and unsupervised algorithms in previous work (Yu & Shi, 2001; Wagstaff et al., 2001; Shental et al., 2003b; Shental et al., 2003a; Basu et al., 2003; Kumar & Hebert, 2003; Kwok & Tsang, 2003; Xing et al., 2002; Yan et al., 2006). In the context of graph partitioning, Yu and Shi (2001) have successfully integrated pairwise constraints into a constrained grouping framework, leading to improved segmentation results. Wagstaff et al. (2001) introduced pairwise constraints into the k-means clustering algorithm for unsupervised learning problems. In more closely related work proposed by Xing et al. (2002), a distance metric learning method is proposed to incorporate pairwise information and solved by convex optimization. However, the

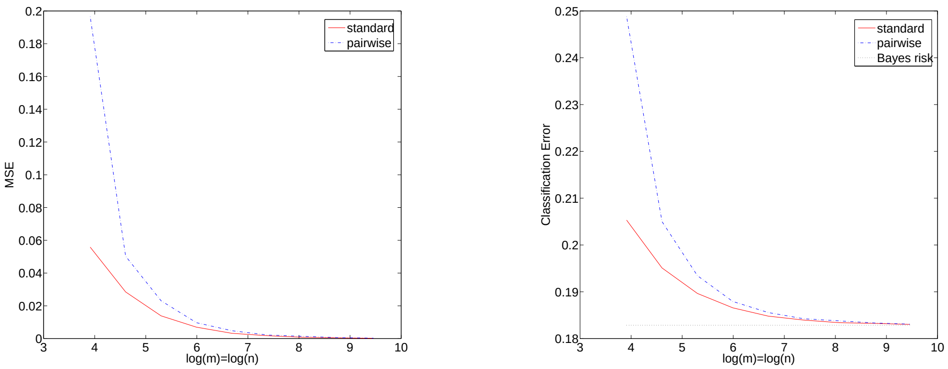

Figure 1. Empirical verification of consistency and efficiency results. “Standard” is for $\hat{\beta}$ which is obtained using labeled examples only; “Pairwise” is for $\hat{\alpha}$ or $-\hat{\alpha}$ which is obtained using pairwise constraints only.

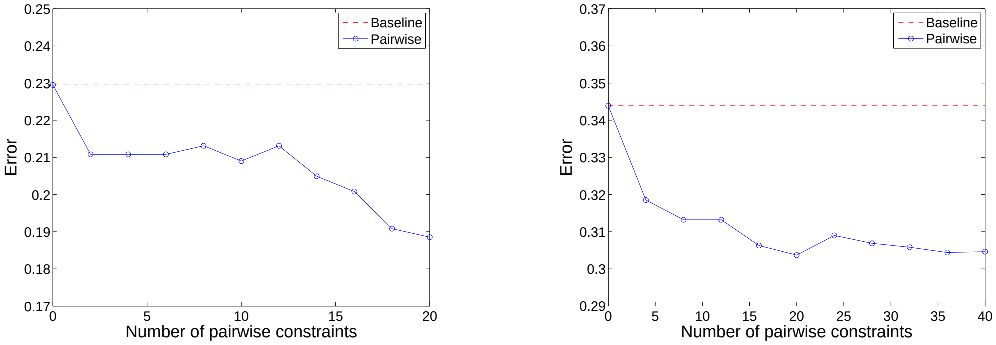

Figure 2. Results on a person identification task. Left: The classification error of pairwise learning algorithms against number of constraints in the single day dataset. The number of constraints is growing from 0 to 20. Right: similar except results are reported on the multiple day dataset with the constraints growing from 0 to 40.

method contains an iterative procedure with projection and eigenvalue decomposition which is computationally expensive and sensitive to parameter tuning. By comparison, relevant component analysis (RCA) (Shental et al., 2003b) is a simpler and more efficient approach for learning a full Mahalanobis metric. A whitening transformation of the covariance matrix of all the center-points in the chunklets is computed as a Mahalanobis distance. However, only positive constraints can be utilized in this algorithm. In (Shental et al., 2003a), Shental et al. propose a constrained Gaussian mixture model which incorporates both positive and negative pairwise constraints into a GMM model using the expectation-maximization (EM) algorithm. Basu et al. (2003) studied a new approach for semi-supervised clustering by adding additional penalty terms into the objective function. They also proposed an approach to actively select the most informative constraints rather than selecting them at random. In (Basu et al., 2004), they also used pairwise constraints to learn more advanced metrics such as parameterized Bregman distances or directional distances. Kumar and Hebert (2003) presented a discriminative learning framework for the classification of the image regions by incorporating interactions from neighborhood nodes together with the observed examples. Pairwise constraints have also been found useful in the context of kernel learning. Kwok and Tsang (2003) formulated the kernel adaptation problem as a distance metric learning problem that searches for a suitable linear transform in the kernel-induced feature space, even if it is of infinite dimensionality. Yan et al. (2006) propose a discriminative learning approach by incorporating pairwise constraints into a conventional margin-based learning framework.

Our paper is also related to the previous work on spectral clustering. Blatt et al. (Blatt et al., 1997) proposed a graph-cut method using a simple probabilistic model, which the cut criterion depends on a free parameter called “temperature”. Shental et al. (Shental et al., 2003c) extended this approach to a new spectral clustering approach based on a undirected graphical model, so that it can provide a probabilistic interpretation for the typical minimal cut approach. Zass and Shashua (Zass & Shashua, 2005) proposed a unified approach to probabilistic clustering by a completely positive factorization, which includes spectral clustering, normalized cuts and kernel K-means as special instances. They also considered clustering with pairwise constraints by using a linear combination of the affinity matrix and the constraint matrix. Our method differs from those spectral clustering methods in several ways: First, it results in an optimal, out-of-sample classifier (instead of just clustering existing examples); Second, the transformation used in our method is based on tensor product and involves both examples within a constraint, which is different from the transformation used in the above spectral clustering methods.

7. Conclusion and Future Work

In this paper we study the problem of classification in the presence of pairwise constraints, and propose a simple method which can effectively utilize pairwise constraints to construct an estimator of the decision boundary. Our method can be efficiently computed and we show that the resulting estimator is a sign-insensitive consistent estimator of the optimal linear decision boundary. We also study the asymptotic variance of the estimator and extend the method so that it can handle both labeled and pairwise examples. Our results confirm the theoretical analysis and show that the proposed method is effective on several datasets. As future work, we would like to extend the linear function class to the kernel class which is dense w.r.t. measurable functions. This will both simplify the computation and lead to an estimator which can approximately achieve the Bayes risk.

References

Bartlett, P. L., Jordan, M. I., & McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association, 101(473)*, 138–156.

Basu, S., Banerjee, A., & Mooney, R. J. (2003). Active semi-supervision for pairwise constrained clustering. *Proceedings of the 20th Intl. Conf. on Machine Learning.* Washington, DC.

Basu, S., Bilenko, M., & Mooney, R. J. (2004). A probabilistic framework for semi-supervised clustering. *Proc. of SIGKDD* (pp. 59–68).

Blatt, M., Wiseman, S., & Domany, E. (1997). Data clustering using a model granular magnet. *Neural Computation, 9*, 1805–1842.

Harville, D. (1997). *Matrix algebra from a statistician's perspective.* Springer.

Knight, K., & Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics, 28(5)*, 1356–1378.

Kumar, S., & Hebert, M. (2003). Discriminative random fields: A discriminative framework for contextual interaction in classification. *IEEE International Conference on Computer Vision (ICCV).*

Kwok, J. T., & Tsang, I. W. (2003). Learning with idealized kernel. *Proceedings of the 20th Intl. Conf. on Machine Learning.* Washington, DC.

Shental, N., Bar-Hillel, A., Hertz, T., & Weinshall, D. (2003a). Computing gaussian mixture models with em using side information. *Workshop on 'The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining', ICML 2003.* Washington, DC.

Shental, N., Bar-Hillel, A., Hertz, T., & Weinshall, D. (2003b). Enhancing image and video retrieval: Learning via equivalence constraints. *Proc. of CVPR.* Madison, WI.

Shental, N., Zomet, A., Hertz, T., & Weiss, Y. (2003c). Pairwise clustering and graphical models. *NIPS 03.*

van der Vaart, A. (2000). *Asymptotic statistics.* Cambridge University Press.

Wagstaff, K., Cardie, C., Rogers, S., & Schroedl, S. (2001). Constrained k-means clustering with background knowledge. *Proc. of the 18th Intl. Conf. on Machine Learning* (pp. 577–584). Morgan Kaufmann Publishers Inc.

Xing, E. P., Ng, A. Y., Jordan, M. I., & Russel, S. (2002). Distance metric learning with applications to clustering with side information. *Advances in Neural Information Processing Systems.*

Yan, R., Zhang, J., Yang, J., & Hauptmann, A. (2006). A discriminative learning framework with pairwise constraints for video object classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(4)*, 578–593.

Yu, S. X., & Shi, J. (2001). Grouping with directed relationships. *Lecture Notes in Computer Science, 2134*, 283–291.

Zass, R., & Shashua, A. (2005). A unifying approach to hard and probabilistic clustering. *Proc. of ICCV.*

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics, 32*, 56–85.