Pareto Optimal Linear Classification

Seung-Jean Kim SJKIM@STANFORD.ORG
Alessandro Magnani ALEM@STANFORD.EDU
Sikandar Samar SIKANDAR@STANFORD.EDU
Stephen Boyd BOYD@STANFORD.EDU
Department of Electrical Engineering, Stanford University, Stanford, CA 94304 USA

Johan Lim JOHANLIM@STAT.TAMU.EDU
Department of Statistics, Texas A&M University, College Station, TX 77843-3143 USA

Abstract

We consider the problem of choosing a linear classifier that minimizes misclassification probabilities in two-class classification, which is a bi-criterion problem, involving a trade-off between two objectives. We assume that the class-conditional distributions are Gaussian. This assumption makes it computationally tractable to find Pareto optimal linear classifiers whose classification capabilities are inferior to no other linear ones. The main purpose of this paper is to establish several robustness properties of those classifiers with respect to variations and uncertainties in the distributions. We also extend the results to kernel-based classification. Finally, we show how to carry out trade-off analysis empirically with a finite number of given labeled data.

1. Introduction

We consider two-class (binary) classification in which the input (or instance) space $\mathcal{X}$ is $\mathbb{R}^n$, and the output (or class label) set $\mathcal{Y}$ is $\{-1, +1\}$. An input-output pair (x, y) where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ is called an example. An example is called negative (positive) if its label is -1 $(+1)$. We assume that examples (x, y) are drawn randomly and independently according to a fixed probability distribution D over $\mathcal{X} \times \mathcal{Y}$. We use D_- (D_+) to denote the marginal distribution of inputs in the negative (positive) class.

Appearing in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006.

1.1. Trade-off in Two-Class Classification

A (binary) classifier $h : \mathcal{X} \to \mathcal{Y}$ assigns a binary class label to each instance from $\mathcal{X}$. The classification performance of a classifier h can be summarized by the pair $(P_{\rm tn}(h), P_{\rm tp}(h))$ of correct classification capabilities. Here, $P_{\rm tn}(h)$ is the *true negative rate* (the probability that a negative example is classified correctly) and $P_{\rm tp}(h)$ is the *true positive rate* (the probability that a positive example is classified correctly):

$$\begin{aligned} P_{\rm tn}(h) &= \Pr(h(x) = -1 \mid y = -1), \\ P_{\rm tp}(h) &= \Pr(h(x) = +1 \mid y = +1). \end{aligned}$$

The performance can also be summarized by the misclassification capabilities: the *false positive rate* $P_{\rm fp}(h) = 1 - P_{\rm tn}(h)$ and the *false negative rate* $P_{\rm fn}(h) = 1 - P_{\rm tp}(h)$. The *(expected) error rate* of h (the probability that an example drawn randomly from $\mathcal{X} \times \mathcal{Y}$ according to D is misclassified by h) is $\pi_- P_{\rm fp}(h) + \pi_+ P_{\rm fn}(h)$, where $\pi_- = \Pr(y = -1)$ and $\pi_+ = \Pr(y = 1)$ are prior class probabilities.

Suppose we are given a family $\mathcal{H}$ of classifiers. A standard classification problem is to find a classifier over this family that minimizes the error rate. In many applications, however, we want to know the possible trade-off of misclassification costs associated with the false positive and negative rates (Bach et al., 2005), which is called *cost-sensitive learning*. There has been a growing interest in taking into account skewed or asymmetric misclassification costs in the literature (Bach et al., 2005; Wu et al., 2005; Zadrozny et al., 2003; Zhu & Wu, 2004).

Cost-sensitive learning with the family $\mathcal{H}$ is a bi-criterion problem, involving a trade-off between two classification probabilities. We say that a given pair (α, β) of correct classification probabilities is *achievable* (by $\mathcal{H}$) if there is a classifier $h \in \mathcal{H}$ such that

$P_{\mathrm{tn}}(h) \geq \alpha$ and $P_{\mathrm{tp}}(h) \geq \beta$. The collection of all achievable pairs (α, β) defines a region in $[0,1] \times [0,1]$. We call the curve along the upper boundary of the region the *optimal trade-off curve*. We say that a classifier $h \in \mathcal{H}$ is *Pareto optimal* if the pair $(P_{\mathrm{tn}}(h), P_{\mathrm{tp}}(h))$ is on the optimal trade-off curve. We should mention that the definitions introduced above can be extended to optimal trade-off analysis with two objectives which are not associated with specific distributions, such as Chebyshev bounds on the true negative and positive rates.

We can carry out optimal trade-off analysis with other combinations of classification and misclassification probabilities that involve a trade-off, *e.g.*, the true and false positive rates. (The optimal trade-off curve between these rates is called the *receiver operating characteristic* (ROC).) The Pareto optimal linear classifiers remain the same, regardless of the combination used.

Optimal trade-off analysis with all classifiers amounts to performing a log-likelihood ratio test, which is known as the Neyman-Pearson lemma (Lehmann & Romano, 2005). We call the resulting optimal classifiers *Neyman-Pearson (NP) optimal*. In general, such classifiers are nonparametric. On the other hand, Pareto optimal classifiers are parametric if the family $\mathcal{H}$ is parameterized by a finite number of parameters.

1.2. Pareto Optimal Linear Classification with Gaussian Data

A classifier of the form $h(x) = \mathrm{sgn}\left(a^T x - b\right)$, where $\mathrm{sgn}(\cdot)$ is the sign function, is called *linear*. This classifier has parameters a (the slope or weight vector) and b (the threshold). We identify it with (a, b). The family of linear classifiers is given by $\mathcal{H}$

$$\mathcal{H} = \{(a,b) \mid a \in \mathbb{R}^n \backslash \{0\},\ b \in \mathbb{R}\}.$$

(We rule out linear classifiers with zero slope, since they lack classification capabilities.)

We consider optimal trade-off analysis with linear classifiers in the generative setting, in which we estimate the class-conditional distributions from given training inputs and find Pareto optimal classifiers with the estimates. We assume that the class-conditional distributions are Gaussian $D_- = N(\mu_-, \Sigma_-)$ and $D_+ = N(\mu_+, \Sigma_+)$. Here, we use $N(\mu, \Sigma)$ to denote the Gaussian distribution with mean μ and covariance Σ. We assume that Σ_- and Σ_+ are positive definite.

When the class-conditional distributions are Gaussian, NP optimal classifiers have quadratic decision boundaries (Hastie et al., 2001, §4.3). In the case of $\Sigma_- = \Sigma_+$, the quadratic decision boundaries become linear. In this case, Pareto optimal linear classification generates NP optimal classifiers.

1.3. Related Work

Recently, Bach et al. (2005) have proposed a method for carrying out an approximate trade-off analysis with linear classifiers. The method finds the linear classifier that minimizes the convex function

$$\begin{aligned} & f(a,b) \\ &= C_- \mathbf{E}_{x \sim D_-} \phi\left(b - a^T x\right) + C_+ \mathbf{E}_{x \sim D_+} \phi\left(a^T x - b\right), \end{aligned}$$

where C_- (C_+) represents the misclassification cost associated with the false positive (negative) rate. Here, ϕ is a convex loss function (*e.g.*, the hinge loss or logistic loss) that approximates the true loss function

$$\phi_{0\text{–}1}(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z < 0, \end{cases}$$

leading to a convex approximation to the true misclassification cost $C_- P_{\mathrm{fp}}(a,b) + C_+ P_{\mathrm{fn}}(a,b)$. By changing the ratio of C_- and C_+, we can find a parameterized family of linear classifiers with different misclassification costs. The method can be viewed as a *discriminative* approach to an approximate optimal trade-off analysis with linear classifiers (Bach et al., 2005).

1.4. Brief Overview

Pareto optimal linear classification with Gaussian class-conditional distributions is computationally tractable. Moreover, it requires the estimation of only the first two moments of the class-conditional distributions, *i.e.*, the means and covariances. The main purpose of this paper is to show that the linear classifiers found with the Gaussian assumption have other desirable features than the two mentioned above:

- The classifiers remain Pareto optimal, although the true distributions are mixtures of normal distributions with the same means and scaled covariances.
- The classifiers with true negative and positive rates greater than 0.5 remain Pareto optimal, although we judge the classification probabilities of a classifier with Chebyshev bounds on the rates.
- The classifiers remain Pareto optimal, although we judge the classification probabilities with their worst-case values over the distributions whose maximum allowable deviations from the Gaussian distributions in the Kullback-Liebler divergence are known.

The true distributions considered in the first result are called scale mixtures of normal distributions and can exhibit heavy-tail phenomena. The result implies that Pareto optimal linear classifiers are robust with respect to possible heavy-tail phenomena.

The last two results show that Pareto optimal linear classification has some desirable worst-case robustness properties with respect to variations and uncertainties in the distributions. In particular, the second result is related to the minimax probability machine (MPM) (Lanckriet et al., 2002) and its extension, the minimum error minimax probability machine (MEMPM) (Huang et al., 2004), which find Pareto optimal linear classifiers with the Chebyshev bounds.

Before describing in detail the three results listed above in §3–§5, we show in §2 how the optimal trade-off curve can be found efficiently via convex optimization. We extend the results to kernel-based classification in §6. In §7, we show how one can carry out empirical trade-off analysis with a finite number of examples with known labels. In §8, we give our conclusions.

2. Pareto Optimal Linear Classifiers

With the Gaussian assumption, we can compute the true negative and positive rates of any linear classifier (a, b) as

$$\begin{aligned} \Pr(a^T x < b \mid y = -1) &= \Phi\left(\frac{b - a^T\mu_-}{\sqrt{a^T\Sigma_- a}}\right), \\ \Pr(a^T x > b \mid y = +1) &= \Phi\left(\frac{a^T\mu_+ - b}{\sqrt{a^T\Sigma_+ a}}\right), \end{aligned} \tag{1}$$

where Φ is the cumulative distribution function (CDF) of the standard normal distribution. Optimal trade-off analysis with linear classifiers is a bi-criterion optimization problem with the two objectives above. We denote as $\mathcal{L}(\mu_-, \Sigma_-, \mu_+, \Sigma_+)$ the set of Pareto optimal linear classifiers found via solving this bi-criterion problem. (See Boyd and Vandenberghe (2004, §4.7.5) for more on bi-criterion optimization.)

2.1. Optimal Trade-off Curve

The pair $(0.5, 0.5)$ is always achievable and hence is below or on the optimal trade-off curve of true negative and positive rates. Let A be the point on the optimal trade-off curve with true negative rate 0.5, and let B be the point with true positive rate 0.5. The two points A and B are given by

$$A = (0.5, \Phi(r_1)), \qquad B = (\Phi(r_2), 0.5),$$

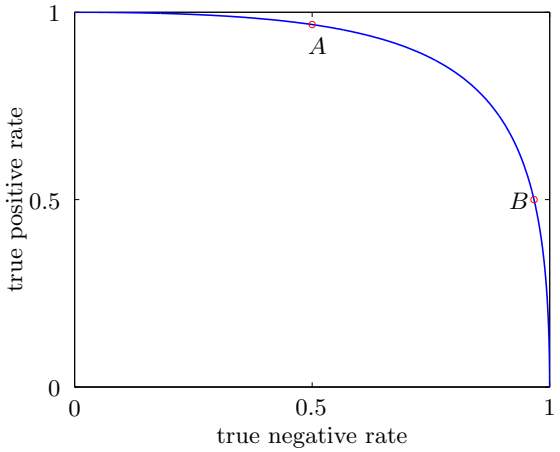

Figure 1. Optimal trade-off curve with Gaussian class-conditional distributions.

where

$$\begin{aligned} r_1 &= \sup_{b=a^T\mu_-, a\in\mathbb{R}^n\setminus\{0\}} \frac{a^T\mu_+ - b}{\sqrt{a^T\Sigma_+ a}} \\ &= \sup_{a\in\mathbb{R}^n\setminus\{0\}} \frac{a^T\mu_+ - a^T\mu_-}{\sqrt{a^T\Sigma_+ a}} \\ &= \left[(\mu_+ - \mu_-)^T\Sigma_+^{-1}(\mu_+ - \mu_-)\right]^{1/2}, \\ r_2 &= \sup_{b=a^T\mu_+, a\in\mathbb{R}^n\setminus\{0\}} \frac{b - a^T\mu_-}{\sqrt{a^T\Sigma_- a}} \\ &= \sup_{a\in\mathbb{R}^n\setminus\{0\}} \frac{a^T\mu_+ - a^T\mu_-}{\sqrt{a^T\Sigma_- a}} \\ &= \left[(\mu_+ - \mu_-)^T\Sigma_-^{-1}(\mu_+ - \mu_-)\right]^{1/2}. \end{aligned} \tag{2}$$

Suppose $\mu_+ \neq \mu_-$. Then, the point $(0.5, 0.5)$ is below the optimal trade-off curve. We can divide the curve into three segments: one from $(0, 1)$ to A, one between A and B, and one from B to $(0, 1)$. (See Figure 1.) The segment from $(0, 1)$ to A corresponds to Pareto optimal linear classifiers (a, b) with $P_{\text{tn}}(a.b) \leq 0.5$, and the one from B to $(1, 0)$ corresponds to Pareto optimal linear classifiers (a, b) with $P_{\text{tp}}(a.b) \leq 0.5$. Those classifiers sacrifice one objective significantly in favor of the other. The middle segment corresponds to Pareto optimal linear classifiers with true negative and positive rates greater than 0.5, and so are of practical interest, unless the misclassification costs are highly skewed.

2.2. Trade-off Analysis via Convex Optimization

The segment between A and B can be found via solving a convex problem of the form

$$\begin{array}{ll} \text{minimize} & \sqrt{a^T\Sigma_+ a} + \lambda\sqrt{a^T\Sigma_- a} \\ \text{subject to} & a^T(\mu_+ - \mu_-) = 1, \end{array} \tag{3}$$

in which the variable is $a \in \mathbb{R}^n$ and $\lambda > 0$ is a (varying) parameter. Since the objective of (3) is strictly convex, this problem has a unique solution. Let a_λ be the unique solution and define

$$b_\lambda = \mu_+^T a_\lambda - d_\lambda \left(a_\lambda^T \Sigma_+ a_\lambda\right)^{1/2},$$

where d_λ is the inverse of the optimal value of (3). The family of Pareto optimal linear classifiers $(a^\star, b^\star)$ with $P_{\mathrm{tn}}(a^\star, b^\star), P_{\mathrm{tp}}(a^\star, b^\star) > 0.5$ is given by $\{(a_\lambda, b_\lambda) \mid 0 < \lambda < \infty\}$. The proof is given in Kim et al. (2006), a longer version of this paper.

The other two segments can also be found via convex optimization; see Kim et al. (2006) for the details.

3. A General Condition for Pareto Optimality

The following proposition is instrumental in studying the robustness properties of the family $\mathcal{L}(\mu_-, \Sigma_-, \mu_+, \Sigma_+)$.

Proposition 1 *Suppose that the true negative and positive rates for any linear classifier (a, b) with $a \neq 0$ can be written as*

$$\begin{aligned} P_{\mathrm{tn}}(a, b) &= \kappa_- \left(\frac{b - a^T \mu_-}{\sqrt{a^T \Sigma_- a}}\right), \\ P_{\mathrm{tp}}(a, b) &= \kappa_+ \left(\frac{a^T \mu_+ - b}{\sqrt{a^T \Sigma_+ a}}\right), \end{aligned} \tag{4}$$

where κ_- and κ_+ are strictly increasing over $\mathbb{R}$. Then, the set of linear classifiers Pareto optimal with the two objectives in (4) is given by $\mathcal{L}(\mu_-, \Sigma_-, \mu_+, \Sigma_+)$.

The proof is given in Kim et al. (2006).

As a direct consequence of this proposition, for any positive scaling parameters λ_- and λ_+, Pareto optimal linear classification with $D_- = N(\mu_-, \lambda_- \Sigma_-)$ and $D_+ = N(\mu_+, \lambda_+ \Sigma_+)$ is the same as that with $D_- = N(\mu_-, \Sigma_-)$ and $D_+ = N(\mu_+, \Sigma_+)$. The usefulness of Proposition (1) is not limited to this simple case.

4. Classification with Scale Mixtures of Normal Distributions

A *scale mixture of normal distributions* has the probability density function (PDF)

$$p_X(x) = \int \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-(x-\mu)^T \lambda \Sigma (x-\mu)} p_\Lambda(\lambda)\, d\lambda,$$

where $|\Sigma|$ is the determinant of Σ. Here λ is called the mixing parameter, and p_Λ is called the mixing distribution. Note that x is drawn according to $N(\mu, \lambda\Sigma)$ where λ is in turn drawn according to the mixing distribution p_Λ. (If the mixing parameter has a point mass at $\lambda = 1$, then the distribution of x is equal to the normal $N(\mu, \Sigma)$.) The mean of x is μ, and the covariance is $\bar{\lambda}\Sigma$ if the mixing parameter has a finite mean $\bar{\lambda} = \mathbf{E}\,\lambda$. We denote the distribution as $S(\mu, \Sigma, p_\Lambda)$.

By varying the mixing distribution, we can generate a wide variety of heavy-tailed distributions, including multivariate t-distributions and multivariate Cauchy distributions (Andrew & Mallows, 1974; Genz & Bretz, 2001). Scale mixtures of normal distributions have been widely used to model heavy-tailed phenomena of multivariate data. For instance, those distributions have been used in statistical image modeling (Wainwright & Simoncelli, 2001).

The following lemma shows how to evaluate analytically classification probabilities of a linear classifier with a scale mixture of normal distributions.

Lemma 1 *Suppose that $x \sim S(\mu, \Sigma, p_\Lambda)$. Then,*

$$\Pr(a^T x > b) = \kappa\left(\frac{a^T \mu - b}{\sqrt{a^T \Sigma a}}\right),$$

where

$$\kappa(u) = \int_0^\infty \Phi\big(u/\sqrt{\lambda}\big) p_\Lambda(\lambda)\, d\lambda. \tag{5}$$

The proof is given in Kim et al. (2006).

The function κ defined above is strictly increasing, since Φ is. The following corollary then follows from Proposition 1.

Corollary 1 *For any mixing distributions p_- and p_+, the set of Pareto optimal linear classifiers for the class-conditional distributions $D_- = S(\mu_-, \Sigma_-, p_-)$ and $D_+ = S(\mu_+, \Sigma_+, p_+)$ is given by $\mathcal{L}(\mu_-, \Sigma_-, \mu_+, \Sigma_+)$.*

NP optimal classifiers for scale mixtures of normal distributions depend on the mixing distributions, unlike Pareto optimal linear classifiers.

5. Robust Linear Classification

In the previous section, we have assumed that the class-conditional distributions D_- and D_+ are fixed and from the family of scale mixtures of normal distributions. In this section we consider the case where these distributions are not known, but certain prior information about them is given. We assume that $D_- \in \mathcal{D}_-$ and $D_+ \in \mathcal{D}_-$, where $\mathcal{D}_-$ and $\mathcal{D}_+$ are the sets of possible distributions. We will judge the classification probabilities of a classifier h by their worst-case

values, over $D_- \in \mathcal{D}_-$ and $D_+ \in \mathcal{D}_+$,

$$\begin{aligned} P_{\text{tn}}^{\text{wc}}(h) &= \inf\{\Pr(h(x) < 0) \mid x \sim D_- \in \mathcal{D}_-\}, \\ P_{\text{tp}}^{\text{wc}}(h) &= \inf\{\Pr(h(x) > 0) \mid x \sim D_+ \in \mathcal{D}_+\}. \end{aligned}$$

We seek Pareto optimal classifiers which are inferior to no other classifiers with the worst-case values above.

5.1. Classification with Chebyshev Bounds

As a first example, we consider the case in which the means and covariances of the class-conditional distributions are known exactly but otherwise arbitrary:

$$D_- \in \mathcal{D}_- = \mathcal{D}(\mu_-, \Sigma_-), \qquad D_+ \in \mathcal{D}_+ = \mathcal{D}(\mu_+, \Sigma_+).$$

Here, we use $\mathcal{D}(\mu, \Sigma)$ to denote the set of distributions with mean μ and covariance Σ and otherwise arbitrary. The nominal distributions are the Gaussian $N(\mu_-, \Sigma_-)$ and $N(\mu_+, \Sigma_+)$.

Using the Chebyshev bound (Marshall & Olkin, 1960), we can compute the worst-case true negative and positive rates as

$$\begin{aligned} P_{\text{tn}}^{\text{wc}}(a,b) &= \inf\left\{\Pr(a^T x < b) \mid x \sim D_- \in \mathcal{D}_-\right\} \\ &= \Psi\left(\frac{b - a^T\mu_-}{\sqrt{a^T\Sigma_- a}}\right), \\ P_{\text{tp}}^{\text{wc}}(a,b) &= \inf\left\{\Pr(a^T x > b) \mid x \sim D_+ \in \mathcal{D}_+\right\} \\ &= \Psi\left(\frac{a^T\mu_+ - b}{\sqrt{a^T\Sigma_+ a}}\right), \end{aligned} \tag{6}$$

where Ψ is defined by

$$\Psi(u) = u_+^2/(1 + u_+^2), \qquad u_+ = \max\{u, 0\}.$$

See, *e.g.*, Lanckriet et al. (2002) for the proof.

The function Ψ is strictly increasing over $(0, \infty)$. This property allows us to carry out Pareto optimal linear classification with the worst-case classification probabilities in (6) without solving the associated bicriterion optimization problem.

Proposition 2 *A linear classifier is Pareto optimal with the worst-case classification probabilities in (6) if and only if it is Pareto optimal linear with (1) and its true negative and positive rates are greater than* 0.5.

The proof is given in Kim et al. (2006).

Figure 2 illustrates the link, established above, between Pareto optimal linear classification with the two objectives in (6) and that with the two in (1). Here, the dotted curve corresponds to the optimal trade-off curve of the two objectives in (1). The curve from C to D (excluding C and D) corresponds to the optimal trade-off curve of the worst-case probabilities

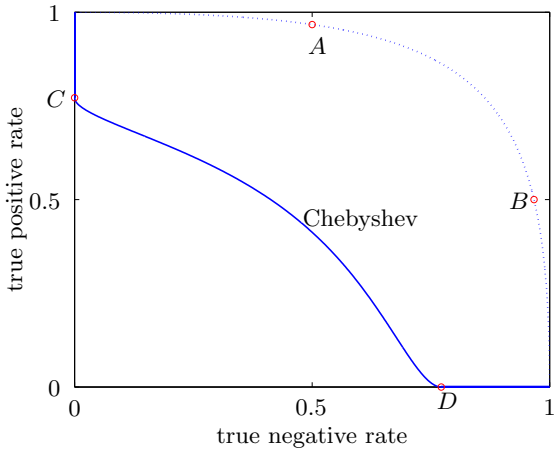

Figure 2. Optimal trade-off curve with Chebyshev bounds.

in (6). The two points C and D are $C = (0, \Psi(r_1))$ and $D = (\Psi(r_2), 0)$ with r_1 and r_2 in (2). (Any point on the line from D to $(0, 1)$ or from C to $(1, 0)$ is not on the optimal trade-off curve, since it is inferior to a point on the same line which is achievable.) Proposition 2 implies that the optimal trade-off curve of the worst-case probabilities in (6) can be easily computed with the classifiers found in §2.2.

We close by clarifying the link between the MEMPM and Pareto optimal linear classification described above. The MEMPM amounts to solving a problem of the form

$$\begin{aligned} \text{maximize} \quad & \theta\alpha + (1-\theta)\beta \\ \text{subject to} \quad & \Psi\left(\frac{b - a^T\mu_-}{\sqrt{a^T\Sigma_- a}}\right) \geq \alpha, \\ & \Psi\left(\frac{a^T\mu_+ - b}{\sqrt{a^T\Sigma_+ a}}\right) \geq \beta, \\ & a^T\mu_- < b < a^T\mu_+, \end{aligned} \tag{7}$$

where $\theta \in (0, 1)$ controls the weights of the worst-case classification probabilities. (The MPM corresponds to the special case of $\alpha = \beta$.) This problem is to find a Pareto optimal linear classifier with the worst-case classification probabilities in (6). Proposition 2 then tells us that the MEMPM in fact finds a linear classifier which is Pareto optimal for $D_- = N(\mu_-, \Sigma_-)$ and $D_+ = N(\mu_+, \Sigma_+)$.

5.2. Classification with KL Divergence Bounds

As another example, we consider the case in which the first and second moments of the true class-conditional distributions are uncertain, but their maximum allowable deviations from the nominal distributions are known in terms of a distance metric. We use, as

the distance metric between two distributions, the Kullback-Liebler (KL) divergence (negative relative entropy):

$$d_{\mathrm{KL}}(P \mid\mid Q) = \int \log \frac{dP}{dQ}\, dP,$$

where dP/dQ denotes the Radon-Nikodym derivative of P with respect to Q.

We assume that the true class-conditional distributions are subject to

$$D_- \in \mathcal{D}_{\delta_-}(\mu_-, \Sigma_-), \quad D_+ \in \mathcal{D}_{\delta_+}(\mu_+, \Sigma_+), \tag{8}$$

where δ_- and δ_+ are positive constants. (The choice of $\delta_- = \delta_+ = 0$ corresponds to the Gaussian setting considered in §2.) Here, we use $\mathcal{D}_\delta(\mu, \Sigma)$ to denote the set of distributions with mean μ and covariance Σ that satisfy $d_{\mathrm{KL}}(D||N(\mu, \Sigma)) \leq \delta$. The nominal distribution of the negative examples is $N(\mu_-, \Sigma_-)$, and that of the positive examples is $N(\mu_+, \Sigma_+)$.

The following lemma shows that we can evaluate analytically the worst-case true negative and positive rates over the families above.

Lemma 2 *For any $\mu \in \mathbb{R}^n$ and any symmetric positive definite $\Sigma \in \mathbb{R}^{n\times n}$, we have*

$$\begin{aligned}&\inf\left\{\Pr\left(a^T x \geq b\right) \mid x \sim P,\ d_{\mathrm{KL}}(P \mid\mid N(\mu, \Sigma)) \leq \delta\right\}\\ &= \kappa_\delta\left(\frac{\mu^T a - b}{\sqrt{a^T \Sigma a}}\right),\ \forall\, a \in \mathbb{R}^n \backslash \{0\},\ \forall\, b \in \mathbb{R},\ \forall\, \delta \geq 0\end{aligned}$$

where

$$\begin{aligned}\kappa_\delta(u) &= 1 - f_\delta^{-1}(\Phi(-u)),\\ f_\delta(\epsilon) &= \sup_{v>0} \frac{e^{-d}(v+1)^\epsilon - 1}{v}.\end{aligned} \tag{9}$$

The proof is given in Kim et al. (2006).

Now, we have

$$\begin{aligned}&P_{\mathrm{tn}}^{\mathrm{wc}}(a, b)\\ &= \inf\left\{\Pr(a^T x < b) \mid x \sim D_- \in \mathcal{D}_{\delta_-}(\mu_-, \Sigma_-)\right\}\\ &= \kappa_{\delta_-}\left(\frac{b - a^T \mu_-}{\sqrt{a^T \Sigma_- a}}\right),\\ &P_{\mathrm{tp}}^{\mathrm{wc}}(a, b)\\ &= \inf\left\{\Pr(a^T x > b) \mid x \sim D_+ \in \mathcal{D}_{\delta_+}(\mu_+, \Sigma_+)\right\}\\ &= \kappa_{\delta_+}\left(\frac{a^T \mu_+ - b}{\sqrt{a^T \Sigma_+ a}}\right).\end{aligned}$$

For any $\delta > 0$, κ_δ is strictly increasing. The following corollary now follows from Proposition 1.

Corollary 2 *The set of Pareto optimal linear classifiers with the worst-case classification probabilities above is given by $\mathcal{L}(\mu_-, \Sigma_-, \mu_+, \Sigma_+)$.*

6. Kernel-Based Classification

In this section, we show how to extend the results established above to kernel-based classification.

6.1. Trade-off Analysis with Kernel-Based Classifiers

In kernel-based binary classification, we seek a classifier $h : \mathcal{X} \to \mathcal{Y}$ of the form $h(x) = \mathrm{sgn}(a^T \phi(x) - b)$, where $a \in \mathcal{H}$ is a weight vector in a high-dimensional (possibly infinite) Hilbert space $\mathcal{H}$, ϕ is a map from $\mathcal{X}$ into $\mathcal{H}$, and $b \in \mathbb{R}$ is the threshold. The space $\mathcal{H}$ is called the *feature space*. The feature space and mapping are defined implicitly through a *kernel (function)* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that satisfies $K(x, z) = \phi(x)^T \phi(z)$ for all $x, z \in \mathcal{X}$. Here, $w^T v$ denotes the inner product between $w, v \in \mathcal{H}$. See, *e.g.*, Schölkopf and Smola (2002) for more on kernel-based classification.

For the extension, we associate two Gaussian $N(\tilde{\mu}_-, \tilde{\Sigma}_-)$ and $N(\tilde{\mu}_+, \tilde{\Sigma}_+)$ in $\mathcal{H}$ with the negative class and positive class, respectively. As in linear classification, kernel-based classification involves a trade-off between two objectives. All results established above can be readily extended to optimal trade-off analysis with linear classifiers in the feature space. For instance, the extension of the computational method described in §2.2 leads to a convex problem of the form

$$\begin{array}{ll}\text{minimize} & \left(a^T \tilde{\Sigma}_+ a\right)^{1/2} + \lambda \left(a^T \tilde{\Sigma}_- a\right)^{1/2}\\ \text{subject to} & a^T(\tilde{\mu}_+ - \tilde{\mu}_-) = 1,\end{array} \tag{10}$$

where $a \in \mathcal{H}$ is the variable and $\lambda > 0$ is fixed.

The data for the extension, *i.e.*, the means and covariances of $N(\tilde{\mu}_-, \tilde{\Sigma}_-)$ and $N(\tilde{\mu}_+, \tilde{\Sigma}_+)$, can be estimated from given training inputs. Let $\{x_1, \ldots, x_{m_+}\}$ be the training inputs from the positive class and let $\{x_{m_+ + 1}, \ldots, x_m\}$ be from the negative class. The sample means are given by

$$\tilde{\mu}_+ = \frac{1}{m_+} \sum_{i=1}^{m_+} \phi(x_i), \quad \tilde{\mu}_- = \frac{1}{m_-} \sum_{i=m_+ + 1}^{m} \phi(x_i),$$

where $m_- = m - m_+$. The (regularized) sample covariances are given by

$$\begin{aligned}\tilde{\Sigma}_+ &= \frac{1}{m_+} \sum_{i=1}^{m_+} (\phi(x_i) - \tilde{\mu}_+)(\phi(x_i) - \tilde{\mu}_+)^T + \delta_+ I,\\ \tilde{\Sigma}_- &= \frac{1}{m_-} \sum_{i=m_+ + 1}^{m} (\phi(x_i) - \tilde{\mu}_-)(\phi(x_i) - \tilde{\mu}_-)^T + \delta_- I,\end{aligned}$$

where δ_+ and δ_- are positive regularization parameters. Since the covariances may be singular, we add (small) regularization terms to the covariances.

6.2. Kernel Trick

We describe how the kernel trick (Schölkopf & Smola, 2002) can be extended to the kernel-based classification method described above, based on the sample means and covariances. The extension is nearly identical to that of the MPM to kernel-based classification described in Lanckriet et al. (2002).

Let $G \in \mathbb{R}^{n\times n}$ be the Gram matrix that contains as its entries the inner products in $\mathcal{H}$ between all pairs of the images of the training inputs $\{x_1, \ldots, x_m\}$:

$$G_{ij} = k(x_i, x_j).$$

Then, we can reformulate (10) as

$$\begin{array}{ll} \text{minimize} & \left(\alpha^T F_+ \alpha\right)^{1/2} + \lambda \left(\alpha^T F_- \alpha\right)^{1/2} \\ \text{subject to} & \alpha^T G(g_+ - g_-) = 1, \end{array} \tag{11}$$

where the variable is $\alpha \in \mathbb{R}^m$, and

$$\begin{aligned} g_+ &= \begin{bmatrix} (1/m_+)\mathbf{1}_{m_+} \\ 0_{m_-} \end{bmatrix}, \quad g_- = \begin{bmatrix} 0_+ \\ (1/m_-)\mathbf{1}_{m_-} \end{bmatrix}, \\ F_+ &= GJ_+J_+^TG + \delta_+ G, \quad F_- = GJ_-J_-^TG + \delta_- G, \\ J_+ &= \operatorname{diag}\left((1/m_+^{1/2})\left[I - (1/m_+)\mathbf{1}_{m_+}\mathbf{1}_{m_+}^T\right], 0_{m_-}\right), \\ J_- &= \operatorname{diag}\left(0_{m_+}, (1/m_-^{1/2})\left[I - (1/m_-)\mathbf{1}_{m_-}\mathbf{1}_{m_-}^T\right]\right). \end{aligned}$$

Here 0_n denotes the vector of all zeros in $\mathbb{R}^n$, and $\operatorname{diag}(A_1, \ldots, A_n)$ denotes the diagonal matrix whose diagonal block entries are A_i. The solution of (10) is given by $a^\star = \sum_{i=1}^m \alpha^\star \phi(x_i)$, where $\alpha^\star$ is the solution of (11).

The optimal classifier determined by this solution can be expressed as

$$f(z) = \operatorname{sgn}\left(\sum_{i=1}^m \alpha_i^\star k(x_i, z) - b_\lambda\right)$$

with the threshold $b_\lambda = \alpha^{\star T} G g_+ - d^\star \left(\alpha^{\star T} F_+ \alpha^\star\right)^{1/2}$, where $d^\star$ is the inverse of the optimal value of (11). Note that this expression requires us to evaluate the kernel function at the pairs (x_i, z), $i = 1, \ldots, m$, not the feature mapping.

7. Empirical Trade-off Analysis

7.1. Trade-off Analysis with Finite Samples

There are several ways of carrying out empirical trade-off analysis with a finite number of given labeled data.

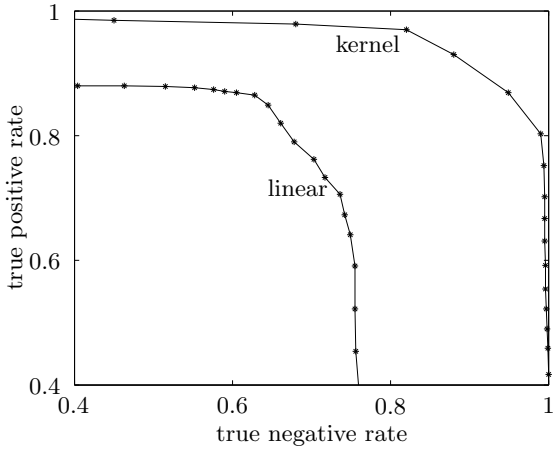

Figure 3. Empirical trade-off analysis results for the ionosphere benchmark data set.

We describe a procedure, based on the resampling technique (Efron & Tibshirani, 1993).

We first randomly partition the data set into a training set and a test set. We use the training set to estimate the sample means and covariances. We then find linear and kernel-based classifiers using the methods described in §2 and §6. For each Pareto optimal linear classifier found, we compute its true negative and positive rates with the test set. We repeat this procedure many times and collect the results. Finally, we compute the trade-off curve via constrained least-squares regression with the collected results, while taking into account the monotonicity of the trade-off curve.

7.2. An Illustrative Example

We illustrate empirical trade-off analysis with the ionosphere benchmark data set from the UCI repository (Newman et al., 1998). This data set consists of 351 points in $\mathbb{R}^{34}$. We used 70% of the data set as the training set.

Figure 3 shows the empirical trade-off analysis results for the ionosphere data set. Here, we used the Gaussian kernel ($e^{-\|x-y\|^2/\sigma}$), where the parameter σ was tuned via cross validation for equal prior class probabilities. For this benchmark data set, kernel-based classification is far superior to linear classification.

8. Conclusions

We have studied Pareto optimal linear classification under the Gaussian assumption on the class-conditional distributions, and studied its several robustness properties. We have also clarified the link

between this classification method and the MEMPM. The classification performance of these robust classification methods is comparable to that of support vector machines (SVMs) (Lanckriet et al., 2002; Huang et al., 2004), which are regarded as the state-of-the-art kernel methods. The numerical comparison result in conjunction with this link supports empirically Pareto optimal linear classifiers found under the Gaussian assumption.

The robustness analysis is based on the assumption that there is no estimation error in the estimates of the means and covariances. Pareto optimal linear classification may suffer from the so-called small sample problem with a small number of training inputs, since the covariances are hard to estimate accurately.

We mention two future research directions. One is to incorporate confidence band analysis in empirical trade-off analysis, which is in spirit similar to confidence band analysis in ROC curves (Macskassy et al., 2005). The other is to compare the generative approach described in this paper with the discriminative approach in Bach et al. (2005). To this end, the area under the optimal trade-off curve plays the same role as the area under the ROC curve (AUC) in ROC analysis.

Acknowledgments

The authors thank Kwangmoo Koh for helpful comments and suggestions. This material is supported in part by the National Science Foundation under grants #0423905 and (through October 2005) #0140700, by the Air Force Office of Scientific Research under grant #F49620-01-1-0365, by MARCO Focus center for Circuit & System Solutions contract #2003-CT-888, and by MIT DARPA contract #N00014-05-1-0700.

References

Andrew, A., & Mallows, C. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society-Series B*, *47*, 99–102.

Bach, F., Heckerman, D., & Horvitz, E. (2005). On the path to an ideal ROC curve: Considering cost asymmetry in learning classifiers. *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS)* (pp. 9–16).

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Efron, B., & Tibshirani, R. (1993). *An introduction to bootstrap*. London UK: Chapman and Hall.

Genz, A., & Bretz, F. (2001). Methods for the computation of multivariate t-probabilities. *Journal of Computational and Graphical Statistics*, *11*, 950–971.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning. data mining, inference, and prediction*. Springer.

Huang, K., Yang, H., King, I., Lyu, M., & Chan, L. (2004). The minimum error minimax probability machine. *Journal of Machine Learning Research*, *5*, 1253–1286.

Kim, S.-J., Magnani, A., Samar, S., Boyd, S., & Lim, J. (2006). Pareto optimal linear classification. Manuscript. Available from `www.stanford.edu/~boyd/pareto_opt_class.html`.

Lanckriet, G., El Ghaoui, L., Bhattacharyya, C., & Jordan, M. (2002). A robust minimax approach to classification. *Journal of Machine Learning Research*, *3*, 555–582.

Lehmann, E., & Romano, J. (2005). *Testing statistical hypotheses*. New York: Springer-Verlag. third edition.

Macskassy, S., Provost, F., & Rosset, S. (2005). ROC confidence bands: An empirical evaluation. *Proceedings of the 22nd International Conference on Machine Learning* (pp. 537–544). ACM.

Marshall, A., & Olkin, I. (1960). Multivariate Chebyshev inequalities. *Annals of Mathematical Statistics*, *32*, 1001–1014.

Newman, D., Hettich, S., Blake, C., & Merz, C. (1998). UCI repository of machine learning databases. Available from `www.ics.uci.edu/~mlearn/MLRepository.html`.

Schölkopf, B., & Smola, A. (2002). *Learning with kernels*. MIT Press, Cambride, MA.

Wainwright, M., & Simoncelli, E. (2001). Scale mixtures of Gaussians and the statistics of natural images. In *Advances in Neural Information Processing Systems*, 12, pp. 855-861, MIT Press.

Wu, J., Mullin, M., & Rehg, J. (2005). Linear asymmetric classifier for cascade detectors. *Proceedings of the 22nd International Conference on Machine Learning* (pp. 988–995). ACM.

Zadrozny, B., Langford, J., & Abe, N. (2003). Cost-sensitive learning by cost-proportionate example weighting. *Proceedings of the 3rd International Conference on Data Mining* (pp. 435–442). IEEE.

Zhu, X., & Wu, X. (2004). Cost-guided class noise handling for effective cost-sensitive learning. *Proceedings of the 4th International Conference on Data Mining* (pp. 297–304). IEEE.