Robust Classification of noisy data using Second Order Cone Programming approach

Chiranjib Bhattacharyya
Dept. Computer Science and Automation, Indian Institute of Science.
Bangalore, Karnataka, India
chiru@csa.iisc.ernet.in

Abstract

Assuming an ellipsoidal model of uncertainty a robust formulation for classifying noisy data is presented. The formulation is a convex optimization problem, in particular it is a instance of Second Order Cone Programmimg problem. The formulation is derived from a worst case consideration and the robustness properties hold for a large class of distributions. The equivalence of ellipsoidal uncertainty and Gaussian noise models is also discussed. The Generalized Optimal hyperplane is recovered as a special case of the robust formulation. **Ex**periments on real world datasets illustrates the efficacy of the formulation.

INTRODUCTION

Classifying observations into one of two classes is an important problem. In real life situations the observation vectors are noisy. We consider the problem of classifying such noisy signals, using hyperplanes. We will assume that only the second order statistics, the mean and covariance, of the noise are available. It is desired that the classifier should be robust to such uncertainty. The main contribution of this paper is to pose this problem as minimizing of a linear objective over ellipsoids.

As **a** special case of this robust formulation the generalized optimal hyperplane can be deduced. The generalized optimal hyperplane formulation was proposed by Vapnik[3] to tackle non-separable data. The support vector machine formulation is an modification to the generalized optimal hyperplane formulation. The generalized optimal hyperplane formulation relies on a user defined parameter which directly bounds the V-C dimension of the candidate hyperplanes. It also has an appealing geometric interpretation that of lower bounding the margin.

The robust formulation and the generalized optimal hyperplane are instances of Second Order Cone Programming(SOCP) problem. Recent advances in Interior point methods for convex nonlinear optimization[4] have made such problems feasible. As a special case of convex nonlinear optimization SOCPs have gained much attention in recent times. For a discussion of efficient algorithms and applications of SOCP see [8]. Various open source software packages are now available. but for the numerical experiments conducted in this paper we used Sedumi[2].

In the next section the problem of robust classification is approached assuming ellipsoidal uncertainty. In section 3 the stochastic interpretation of the robust formulation is studied. In section 4 the robust formulation is specialized to equal and diagonal covariance, yielding the generalized optimal hyperplane. It is contrasted with the support vector machine formulation. In section 5 numerical experiments are reported for generalized optimal hyperplane.

THE ROBUST PROBLEM FORMULATION

We will assume that the observation vector lies in an ellipsoid and is not known precisely. In other words it is known that the data-point can be any point in a specified ellipsoid. In many practical situations the such ellipsoidal specification of uncertainty may be useful. In a later section we will discuss the case of specific distribution on the noise model, e.g. gaussian, which is equivalent to the above ellipsoidal uncertainty assumption.

Let the set $\mathcal{B}(\bar{\mathbf{x}}, \mathbf{C}, \mathrm{y})$ denote

$$\mathcal{B}(\bar{\mathbf{x}}, \mathbf{C}, \mathrm{y}) = \{\mathrm{x} : (\mathbf{x} - \bar{\mathbf{x}})^T \Sigma^{-1} (\mathbf{x} - JT) \leq \gamma^2\}$$

an ellipsoid, Σ being positive definite matrix and $\gamma \geq 0$ is a scalar. Consider the dataset $D = \{(\mathcal{B}(\bar{\mathbf{x}}_i, \Sigma_i, \gamma_i), y_i) : i \in \{1, \ldots, N\}\}$, where $y_i \in \{1, -1\}$ are the class labels. The class label y_i is same for any $\mathbf{x} \in \mathcal{B}(\bar{\mathbf{x}}_i, \Sigma_i, \gamma_i)$. The classification problem is then to find a function to predict y given the ellipsoid $\mathcal{B}$. Let our classifier be the hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$, with the rule $y(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^T\mathbf{x} + b)$. For the hyperplane to be a valid classifier it must obey the following inequalities

$$y_i(\mathbf{w}^T\mathbf{x} + b) \geq 0 \; \forall \mathbf{x} \in \mathcal{B}(\bar{\mathbf{x}}_i, \Sigma_i, \gamma_i) \; i \in \{1, \ldots, N\}$$

If such a pair of $\{\mathrm{w}, b\}$ exists then the data is said to be separable. Datasets are often non-separable, hence it neccessary to relax the above inequalities. Slack variables ξ are introduced to formulate an optimization problem

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \textstyle\sum_{i=1}^{N} \xi_i \\ & y_i(\mathbf{w}^T\mathbf{x} + b) \geq -\xi_i \; \forall \mathbf{x} \in \mathcal{B}(\bar{\mathbf{x}}_i, \Sigma_i, \gamma_i) \\ & \xi_i \geq 0 \forall i \in \{1, \ldots N\} \end{aligned} \tag{1}$$

The constraints over the ellipsoids can be ensured by requiring that

$$\min_{\mathbf{x} \in \mathcal{B}(\bar{\mathbf{x}}_i, \Sigma_i, \gamma_i)} y_i(\mathbf{w}^T\mathbf{x} + b) \geq -\xi_i$$

The minimization on the left hand side is an instance of convex function minimization over **a** convex set. Using the KKT conditions, which are neccessary and sufficient for such

problems, it is easily seen that

$$\min_{\mathbf{x}\in\mathcal{B}(\bar{\mathbf{x}}_i,\Sigma_i,\gamma_i)} y_i\mathbf{w}^T\mathbf{x} = y_i\mathbf{w}^T\bar{\mathbf{x}}_i - \gamma_i\|\Sigma_i^{\frac{1}{2}}\mathbf{w}\|_2$$

where $\|.\|_2$ is the l^2 norm ($\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T\mathbf{x}}$). The following formulation

$$\begin{aligned}\min_{\mathbf{w},b,\xi} \quad & \textstyle\sum_{i=1}^N \xi_i \\ & y_i(\mathbf{w}^T\bar{\mathbf{x}}_i + b) \geq -\xi_i + \gamma_i\|\Sigma_i^{\frac{1}{2}}\mathbf{w}\|_2 \\ & \xi_i \geq 0 \forall i \in \{1,\ldots N\}\end{aligned} \quad (2)$$

is then equivalent to (1). Note that the constraints are positively homegenuos in $\{\mathbf{w},b,\xi\}$, which means that if we rescale the parameters by a positive constant the constraint is not affected. To deal with this extra degree of freedom, one may introduce the constraint $\|\mathbf{w}\|_2 \leq \mathbf{A}$, without any loss of generality. Also setting $\gamma_i = 1$ the formulation (2) can be restated as

$$\begin{aligned}\min_{\mathbf{w},b,\xi} \quad & \textstyle\sum_{i=1}^N \xi_i \\ & y_i(\mathbf{w}^T\bar{\mathbf{x}}_i + b) \geq -\xi_i + \|\Sigma_i^{\frac{1}{2}}\mathbf{w}\|_2 \\ & \|\mathbf{w}\|_2 \leq A \\ & \xi_i \geq 0 \forall i \in \{1,\ldots N\}\end{aligned} \quad (3)$$

We will refer to this formulation as the robust formulation. The robustness is due to the nonlinear term $\|\Sigma_i^{\frac{1}{2}}\mathbf{w}\|_2$ in the constraints.

The problem(3) is an instance of Second Order Cone Program(SOCP). A SOCP is defined[8] as

$$\begin{aligned}\min_{\mathbf{x}} \quad & \mathbf{c}^T\mathbf{x} \\ & \|A_i\mathbf{x} + b_i\|_2 \leq e_i^T\mathbf{x} + f_i\end{aligned} \quad (4)$$

The objective of **a SOCP is** linear, while the constraints are second order cones. Many optimization problems, linear programming, quadratic programming are special cases of SOCP. Recent advances in Interior point methods for solving such problems is a considerable source of excitement in the optimization litreature. Reliable open source softwares are now available to handle large scale SOCPs[2].

DISTRIBUTIONAL ASSUMPTIONS

In the previous section the case of ellipsoidal uncertainty was studied, In this section we first show that assuming ellipsoidal uncertainty is equivalent to assuming the observations are gaussian distributed. We also derive a worst case formulation, which is equivalent to solving the ellipsoidal uncertainty problem.

The Case of Gaussian Noise

Consider the problem

$$\begin{aligned}\min_{\xi,\mathbf{w},b} \quad & \textstyle\sum_{i=1}^N \xi_i \\ s.t \quad & Prob(y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq -\xi_i) \geq \kappa_i \\ & \mathbf{x}_i \sim \mathcal{N}(\bar{\mathbf{x}}_i, \Sigma_i)\end{aligned}$$

$$\|\mathbf{w}\|_2 \leq A\ ,\ \xi_i \geq 0\ \forall i \in \{1\ldots,N\} \quad (5)$$

This is a stochastic reformulation of (1). The stochastic constraints can be understood as requiring that

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq -\xi_i$$

be satisfied with high probability, say greater than some prespecified $\kappa_i \in (0,1]$. The variables κ_i provides the robustness, higher the κ_i more robust the hyperplane. In statistical litreature κ_i are often called confidence.

The formulation(5) can be restated as a deterministic optimization problem. The constraint

$$Prob(y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq -\xi_i) \geq \kappa_i\ \mathbf{x}_i \sim \mathcal{N}(\bar{\mathbf{x}}_i, \Sigma_i) \quad (6)$$

can be transformed into a second order cone constraint. Define the random variable $z_i = -y_i\mathbf{w}^T\mathbf{x}_i$, whose mean is $\bar{z}_i = -y_i\mathbf{w}^T\bar{\mathbf{x}}_i$ and standard deviation $\sigma_{z_i} = \sqrt{\mathbf{w}^T\Sigma_i\mathbf{w}}$. Using z_i the constraint (6) can be written as

$$P\left(\frac{z_i - \bar{z}_i}{\sigma_{z_i}} \leq \frac{y_i b + \xi_i - \bar{z}_i}{\sigma_{z_i}}\right) \geq \kappa_i$$

Since $\frac{z_i - \bar{z}_i}{\sigma_{z_i}}$ is a zero mean unit variance random variable the above equation can again be restated as

$$\phi\left(\frac{y_i b + \xi_i - \bar{z}_i}{\sigma_{z_i}}\right) \geq \kappa_i$$

where ϕ is the cumulative density function for the gaussian distribution

$$\phi(u) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} e^{-\frac{s^2}{2}} ds$$

and ϕ^{-1} is the inverse function. Then the stochastic constraint(6) is equivalent to the deterministic constraint

$$y_i(\mathbf{w}^T\bar{\mathbf{x}}_i + b) \geq -\xi_i + \phi^{-1}(\kappa_i)\sqrt{\mathbf{w}^T\Sigma_i\mathbf{w}} \quad (7)$$

If $\kappa_i \geq 0.5 (\phi^{-1}(\kappa_i) \geq 0)$ then the constraint is a SOCP constraint. Note the equivalence between (7) and the constraints in (2). The relation between the parameters is given by $\gamma_i = \phi^{-1}(\kappa_i)$. The matrices Σ_i are the covariances.

Throughout the derivation it was assumed that the uncertainty has a gaussian distribution. However the structural equivalence of the stochastic constraints(6) and deterministic constraints (7) hold as long as the distribution have the same mean and covariance. For each distribution the relationship between κ and y changes. One can even formulate a worst case problem using **a** chebychev bound.

A worst case formulation

Let $\mathbf{x} \sim (\bar{\mathbf{x}}, \Sigma)$ be a family of distributions having the same mean(%) and covariance(Σ). Then exploiting the Chebychev bound, [5, 6, 7] we can write

$$\begin{aligned} & Prob(y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq -\xi_i) \geq \tfrac{d^2}{1+d^2} \geq \kappa_i \\ & \mathbf{x}_i \sim (\bar{\mathbf{x}}_i, \Sigma_i) \\ & d^2 = \inf_{\mathbf{x}\in(y_i(\mathbf{w}^T\mathbf{x}_i+b)+\xi_i\leq 0)} (\mathbf{x}-\bar{\mathbf{x}})^T\Sigma^{-1}(\mathbf{x}-\bar{\mathbf{x}}) \end{aligned} \quad (8)$$

The equality is attained at the worst case distribution[5]. The computation of d is a convex function minimization over a convex set, hence a convex optimization problem. Cranking the convex optimization machinery an SOCP constraint similar to (7),

$$y_i(\mathbf{w}^T\bar{\mathbf{x}}_i + b) \geq -\xi_i + \psi^{-1}(\kappa_i)\sqrt{\mathbf{w}^T\Sigma_i\mathbf{w}}$$
$$\psi(u) = \tfrac{1}{1+u^2} \qquad (9)$$

is obtained. The above SOCP constraint allows us to cast robustness, in classification, problems without making distributional assumptions.

In this section the robust formulation(1) is rederived as an stochastic optimization problem. The problem is approached by making gaussian assumption on the data. The general case of distributions having the same first two moments are also discussed. Among such a family the worst case distribution is discussed by using a Chebychev like bound. In our remaining sections we specialize our results to the spherical covariance case.

GENERALIZED OPTIMAL HYPERPLANE

Let us specialize the formulation(3) to the case of equal and spherical covariance case $\Sigma_i = \sigma^2\mathbf{I}$, which yields

$$\begin{aligned} \min_{\xi,\mathbf{w},b} \quad & \textstyle\sum_{i=1}^N \mathbf{ti} \\ s.t \quad & y_i(\mathbf{w}^T\bar{\mathbf{x}}_i + b) \geq \sigma\|\mathbf{w}\|_2 - \xi_i \\ & \|\mathbf{w}\|_2 \leq A\ ,\ \xi_i \geq 0\ \ \forall i \in \{1\ldots,N\} \end{aligned} \qquad (10)$$

Note that at optimality $\|\mathbf{w}\|_2 = A$ and requiring that

$$\sigma A = 1 \qquad (11)$$

the equivalent problem

$$\begin{aligned} \min_{\xi,\mathbf{w},b} \quad & \textstyle\sum_{i=1}^N \xi_i \\ s.t \quad & y_i(\mathbf{w}^T\bar{\mathbf{x}}_i + b) \geq 1 - \xi_i \\ & \|\mathbf{w}\|_2 \leq A\ ,\ \xi_i \geq 0\ \ \forall i \in \{1\ldots,N\} \end{aligned} \qquad (12)$$

is formulated. This formulation was presented by Vapnik[3] and was referred to as the generalized optimal hyperplane. The objective function is an upper bound on the number of mistakes. A mistake occurs whenever the predicted class label $\text{sign}(\mathbf{w}^T\mathbf{x} + b)$ disagrees with the given class label, $\mathbf{y}$. The number of mistakes made by a hyperplane $\mathbf{w}^T\mathbf{x}+b=0$, is the same as number of $\xi_i \geq 1$. The optimization problem (12) then tries to find an hyperplane, which minimizes the error on the training set.

The interpretation of $\boldsymbol{A}$ is different for the robust formulation(3) and the generalized optimal hyperplane. For any value of $\mathbf{A}$ the robust formulation is equivalent; if the optimum solution be $\mathbf{z}_0 = [\mathbf{w}^T, b, \xi^T]^T$ for $\mathbf{A} = a_0$, then the optimum for any other $\mathbf{A} = a$ can be recovered from $\mathbf{z}_0$ by $\mathbf{z}(a) = a/a_0\mathbf{z}_0$. In short the same hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ is obtained for any $\mathbf{A} > 0$. However for the generalized optimal hyperplane formulation different values of $\mathbf{A}$ lead to different hyperplanes. This is largely because of equation (11). Choosing $\boldsymbol{A}$ automatically sets a value for $\sigma = \frac{1}{A}$. This then provides an insight why varying $\mathbf{A}$ might produce different results.

In the formulation(**12)** the parameter $\mathbf{A}$ is directly related to the V-C dimension of the hyperplanes examined. If A is small, (V-C dimension low) the candidate hyperplanes will make lot of errors on the training set and test set. While large $\boldsymbol{A}$ (corresponds to a higher V-C dimension), will result in low training set error, but large test set error, often referred to as overfitting. The choice of $\mathbf{A}$ can thus be guided from this overfitting scenario. The margin is defined as the distance between two hyperplanes $\mathbf{w}^T\mathbf{x} + b = 1\mathbf{w}^T\mathbf{x} + b = -1$ and is given by $\frac{2}{\|\mathbf{w}\|_2}$. Geometrically, fixing A corresponds to fixing the margin. The formulation(**12)** then finds a pair of hyperplanes, with fixed distance **A,** such that very few points are misclassified by the classifier, the hyperplane between the two parallel hyperplanes.

Let $\alpha_i(\alpha$ be the vector of all such $\alpha_i)$ be the lagrange multiplier corresponding to the inequality, $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$, and 6 be the lagrange multiplier corresponding to the inequality $\|\mathbf{w}\|_2 \leq \mathbf{A}$. The lagrange dual of the primal problem(12) can then be stated as

$$\begin{aligned} \max_{\alpha,\delta} \quad & \textstyle\sum_{i=1}^N \alpha_i - A6 \\ s.t \quad & \|\mathbf{K}^{\frac{1}{2}}\alpha\|_2 \leq \delta \\ & \textstyle\sum_{i=1}^N \alpha_i y_i = 0, 0 \leq \alpha_i \leq 1\ \forall i \in (1\ldots,N\} \end{aligned} \qquad (13)$$

The matrix $\mathbf{K}^{\frac{1}{2}}$ is the matrix square root of the gram matrix $\mathbf{K}_{ij} = y_i y_j \mathbf{x}_i^T\mathbf{x}_j$. The desired hyperplane parameters $\mathbf{w}, b$ are derived from the KKT conditions. It **is** worth mentioning that $\mathbf{w}$ is given by

$$\mathbf{w} = \tilde{C}\sum_{i=1}^N \alpha_i y_i \mathbf{x}_i,\ \tilde{C} = \frac{A}{\delta} \qquad (14)$$

Modifying (12) an alternate quadratic programming formulation was proposed[3], which now enjoys significant popularity as the Support Vector Machine(SVM) formulation.

$$\begin{aligned} \min_{\xi,\mathbf{w},b} \quad & \tfrac{1}{2}\|\mathbf{w}\|_2^2 + C\textstyle\sum_{i=1}^N \xi_i \\ s.t \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0\ \ \forall i \in \{1\ldots,N\} \end{aligned} \qquad (15)$$

However it is to be noted that (12) and (15) are different problems. They are same only when $C = \tilde{C}$ [3].

The formulation (15) uses an user defined parameter C, which has none of the nice properties of $\mathbf{A}$ discussed before. It is difficult to understand C but much easier to interpret **A.** Recognizing the fact that **12)** and its dual are SOCP, means that we can solve the problem as efficiently as a quadratic program.

Consider the case

$$\Sigma_i = \sigma_i\mathbf{I}.$$

Setting $A = 1$ without loss of generality, (3) specializes to

$$\begin{aligned} \min_{\xi,\mathbf{w},b} \quad & \textstyle\sum_{i=1}^{N} \xi_i \\ s.t \quad & y_i(\mathbf{w}^T\overline{\mathbf{x}}_i + b) \geq \sigma_i - \xi_i \\ & \|\mathbf{w}\|_2 \leq A\ ,\ \xi_i \geq 0\ \ \forall i \in \{1\ldots,N\} \end{aligned} \quad (16)$$

Again the dual of this problem can be stated as

$$\begin{aligned} \max_{\alpha,\delta} \quad & \textstyle\sum_{i=1}^{N} \alpha_i\sigma_i - \delta \\ s.t \quad & \|\mathbf{K}^{\frac{1}{2}}\alpha\|_2 \leq \delta \\ & \textstyle\sum_{i=1}^{N} \alpha_i y_i = 0, 0 \leq \alpha_i \leq 1\ Vi \in \{1\ldots,N\} \end{aligned} \quad (17)$$

The matrix $\mathbf{K}^{\frac{1}{2}}$ is the matrix square root of the gram matrix $\mathbf{K}_{ij} = y_i y_j \mathbf{x}_i^T \mathbf{x}_j$. The vector w is given by equation

$$\mathbf{w} = \frac{1}{\delta}\sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i, \quad (18)$$

The dual formulations(13),(17)always use the gram matrix $\mathbf{K}$. One can thus use nonlinear mercer kernels, to get more complex decision boundaries.

The formulation(12)is then robust to spherical noise, whose magnitude is specified by $\frac{1}{A}$. This is one of the main contributions of this paper. One can go beyond this equal spherical covariance assumption by using (3). This generalisation may be worth studying to handle noisty data.

EXPERIMENTS AND DISCUSSION

To measure the efficacy of the robust generalized optimal hyperplane formulation a performance index is needed. By noting that the gaussian density assumptions imply optimizing over ellipsoids we develop once such error measure. We also report experiments conducted to evaluate the performance of the generalized optimal hyperplane.

A measure of robustness

To measure the robustness of the robust generalized optimal hyperplane appropiate performance indices need to be defined. We propose a worst case error rate.

Again observe that the probabilistic constraint, with

$$\begin{aligned} Prob(y_i(\mathbf{w}^T\mathbf{x}_i + b) \quad & \geq 1 - \xi_i) \geq \kappa_i \\ \mathbf{x}_i & \sim N(\overline{\mathbf{x}}_i, \sigma_i\mathbf{I}) \end{aligned}$$

is equivalent to enforcing the constraint

$$\begin{aligned} & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \\ \forall \mathbf{x}_i \in \mathcal{E}_i = \{\mathbf{x}_i : \|\mathbf{x}_i - \overline{\mathbf{x}}_i\|_2 \leq \phi^{-1}(\kappa_i)\sigma_i\} \end{aligned} \quad (19)$$

By appropiate choice of κ_i set $\phi^{-1}(\kappa_i) = 1$The constraint can be seen as enforcing robustness by implying the entire sphere $\mathcal{E}_i$ to belong to the same side of the hyperplane. We wish to exploit this property to develop an criterion for measuring the performance of the robust generalized optimal hyperplane.

In light of the above arguments it can be said that robustness failed for those spheres which intersects the decision hyperplane. For such spheres a worst case prediction rule will be to take the negative of the predicted rule for the mean. We propose the following rule

$$\begin{aligned} y_i \quad & = sign(t)sign(|t| - \sigma_i) \\ t \quad & = \tfrac{\mathbf{w}^T\overline{\mathbf{x}}_i + b}{\|\mathbf{w}\|_2} \end{aligned} \quad (20)$$

Distance of $\overline{\mathbf{x}}_i$ from the hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ is given by $|t|$, while $sign(t)$ captures the prediction for the mean, $\overline{\mathbf{x}}_i$. If $|t| \leq \sigma_i$ then the hyperplane intersects the sphere $\mathcal{E}_i$. It is assumed that the mean of the hyperplane may be predicted correctly with a very high probability, and the worst case scenarios are those points in the same sphere but are on the different side of the hyperplane. This worst case rule(20) is used to construct a worst case error rate defined as number of mistakes using the worst case rule divided by total number of samples on the test set. This worst case error rate is our proposed performance metric.

Experiments

Experiments were conducted on three real world datasets, downloaded from UCI machine learning dataset website[9]. Ionosphere, sonar and wiconsin breast cancer were the three different datasets. The ionosphere dataset contains **34** dimensional observations, which are obtained from radar signals, while the sonar dataset contains 60 dimensional observation vectors. The wisconsin dataset contains 9 dimensional observations, obtained from processing of images. Each of the datasets contain observations which are possibly noisy, thus maybe a good testbed to measure the efficacy of the robust generalized optimal hyperplane.

In our experiments we assume the spherical noise model. Since there is no way to estimate the noise parameters, for our simulations we assumed that the noise was same $\sigma_i = \sigma$. The values specified in the datasets were treated as the means$(\overline{\mathbf{x}}_i)$. We studied the robustness, using 20, on the respective testsets, by varying σ. For comparison across various datasets we introduced a new parameter ρ, related to σ by

$$\sigma = \sqrt{M}r\rho$$

Let r be the smallest range among all the features for a specified dataset. More precisely

$$\begin{aligned} r = \min_i r_i \quad & r_i = \max_j x_{ji} - \min_j x_{ji}, \\ & j \in \{1\ldots N\}, i \in \{1\ldots M\}. \end{aligned}$$

N = number of training examples, M = number of features. The parameter $\rho \in [0, 1]$ can be seen as noise level, something akin to noise to signal ratio.

Each of the datasets are partitioned into training set (70%) and test set **(30%)**randomly. Let us call the classifier corresponding to $\rho = 0$ as the nominal classifier. It was obtained by solving**(13).** The worst case error rate was then evaluated for various values of ρ. For each ρ, the average worst case error rate is reported, average taken over 10 random partitions. The robust generalized optimal hyperplane was obtained by solving (17)for various values of ρ and again the

average worst case error rate for 10 different partitions was reported. Throughout all the above mentioned experiments A was arbitrarily set to 5.

We also conducted experiments to evaluate the generalized optimal hyperplane. The parameter A in (13) was varied and the error rate was evaluated. We again do 70%/30% split of training to test data randomly and report the average error rate for 10 such splits were reported. All experiments were done in matlab using the open source socp solver Sedumi[2]. A code will be made available upon request.

Discussion of results

On all the three datasets the error rates on the training set decreased monotonically as a function of A. The test set error bottomed out and showed an increase for higher values of A. This experiment clearly brings out the fact that very low A doesn't fit the data well enough, while very high A exhibits overfitting.

By design of the worst case error measure the robust generalized optimal hyperplane will have a lower error rate than the nominal classifier. But as noise increases both the rates should approach a common value. Also it should be monotonic increasing as a function of ρ. The experimental results agree with these facts. The robust formulation works best for the Wisconsin dataset. The sonar dataset looked most suseptible to noise.

CONCLUSIONS

A robust classification problem was formulated to deal with noisy observations. The problem is formulated as an SOCP, assuming that the mean and covariance of the noise model is known. As a special case of the robust formulation, the diagonal covariance case recovers the generalized optimal hyperplane. The analysis here throws light on the robustness properties of the support vector machine formulation. This then paves the way for using more elaborate noise model and hence is an useful generalization of the existing generalized optimal hyperplane formulation. The robustness properties of SVM regression[3], can be similarly studied.

Acknowledgements

REFERENCES

[1] Boyd, S. and Vandenberghe, L. 2002 Convex optimization Course notes Available at www.stanford.edu/class/ee364

[2] Sturm, J. Sedumi Matlab toolbox Available at www.unimaas.nl/~sturm

[3] Vapnik, V.(1998) The nature of statistical learning theory. Springer Verlag

[4] Nesterov, **Y.** and Nemirovsky, A. (1994) Interior Point Polynomial Methods in Convex Programming: Theory and Applications. Philadelphia, PA: SIAM.

[5] Marshall, **A.** W., and Olkin, I.,(1960) Multivariate Chebychev Inequalities Annals of Mathematical statistics, 31(4), pgs 1001-1014

[6] Bertsirnas, D. and Sethuraman, J.(2000) Moment Problems and Semidefinite OptimizationHandbook of Semidefinite Optimization, pgs 469-509 Kluwer Academic Press

[7] Lanckriet, G. R. G, El Ghaoui, L. , Bhattacharyya, C., and Jordan, M. I. (2002) Minimax probability machine Advances in Neural Information Processing Systems 14

[8] Lobo, M., Vandenbeghe, L., Boyd, S. and Lebret, H.(1998) Applications of second-order cone programming. Linear Algebra and its applications 284, Nov. Pgs 193-228

[9] Blake, C. L. and Merz, C. J. (1998) UCI Repository of Machine Learning Databases http://www.ics.uci.edu/ mlearn/MLRepository.html, Dept of Information and Computer Science, Irvine, California

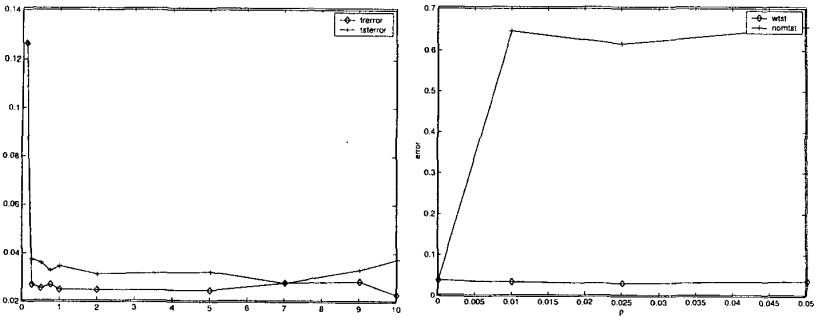

Figure 1. Wisconsin dataset: On the left average error rates for training set and test are plotted as a function of A. On the right average worst case error rate for both the robust and nominal hyperplane is plotted as function of p. The legend wtst corresponds to the robust generalized hyperplane.

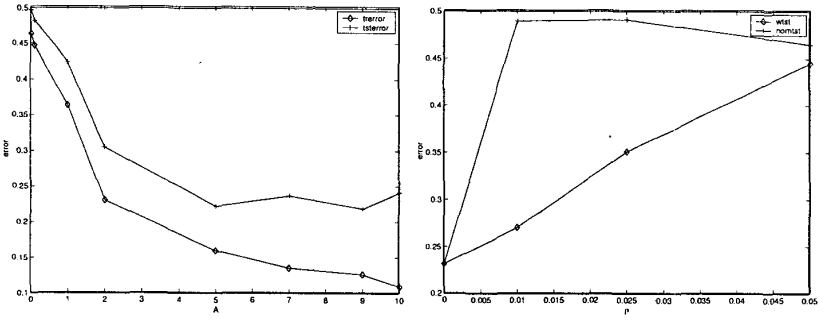

Figure 2. Sonar dataset: On the left average error rates for training set and test are plotted as a function of A. On the right average worst case error rate for both the robust and nominal hyperplane is plotted as function of p. The legend wtst corresponds to the robust generalized hyperplane.

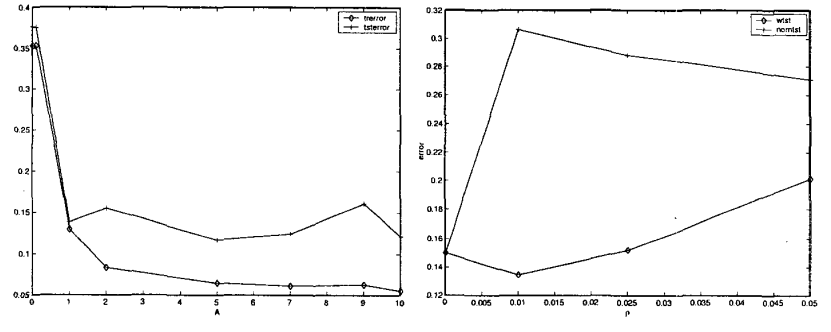

Figure 3. Ionosphere dataset: On the left average error rates for training set and test are plotted as a function of A. On the right average worst case error rate for both the robust and nominal hyperplane is plotted as function of ρ. The legend wtst corresponds to the robust generalized hyperplane.