ROBUST CLASSIFICATION OF SIGNAL ESTIMATES GIVEN A CHANNEL MODEL

Nathan Parrish, Maya R. Gupta

University of Washington
Department of Electrical Engineering
Seattle, WA

Hyrum S. Anderson

Sandia National Labs
Albuquerque, NM

ABSTRACT

In many signal processing applications, a signal to be classified has been corrupted by a channel and additive noise. A standard approach is to estimate the clean signal, then classify it. We consider two robust approaches that account for the estimation procedure. The first approach is an application of the MAP rule for noisy features, and the second is an approach for discriminative classifiers that treats that training points as random. An experiment confirms that the robust approaches offer performance gains.

Index Terms— Classification algorithms, machine learning algorithms, signal processing algorithms, multipath channels

1. INTRODUCTION

We consider the dataset shift problem of using training vectors to classify a test vector that is the output of a known linear system with additive noise. Specifically, we take as given n labeled training examples $\mathcal{T} = \{(x_i, g_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$ is a feature vector and $g_i \in \mathcal{G}$ is a class label. The goal is to classify a test sample x, where x and its class label and the n training example pairs are assumed drawn iid from some joint distribution over $\mathbb{R}^d \times \mathcal{G}$. However, one does not directly observe x, instead, one observes the corrupted and test sample $z = Hx + w$ for $z \in \mathbb{R}^n$, for a known $H \in \mathbb{R}^{n \times d}$, with w representing zero mean AWGN with $\text{Cov}[w] = \sigma_w^2 I$. Examples of this set-up are signal and image classification problems where the training data is collected under controlled conditions, but the test data is collected in a field environment in the presence of multipath, blur, or other convolutive noise [1, 2, 3, 4].

Given H and z, it is common to first form an estimate $\hat{x}$ of the true test sample x by deconvolution, and then classify $\hat{x}$ using a standard classifier trained on $\mathcal{T}$ as depicted in Figure 1. While this can reduce the severity of the dataset shift problem, it does not solve the problem because $\hat{x}$ is not iid with $\mathcal{T}$. Other approaches are to process the training samples through H to form *virtual* training samples $\{\hat{z}_i\}$ which are used to train a classifier for z [4], or to form random training samples $\{\hat{Z}_i\}$ to train a classifier for z [2, 3].

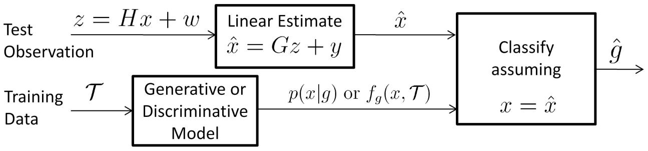

Fig. 1. Block diagram showing the non-robust classification of estimate $\hat{x}$. This classifier assumes that $p(\hat{x}, g) = p(x, g)$, which is not true in general.

In this work, we investigate how to best classify linear estimators of the form $\hat{x} = Gz + y$, such as the least-squares (LS) estimator or linear minimum mean-squared error estimator (LMMSE) [5]. One key question is whether it is possible to classify the estimate $\hat{x}$ with the same precision as the original test observation z. In Section II, we answer this question, showing that under certain conditions MAP classification of z is equivalent to MAP classification of $\hat{x}$. In Section III, we apply the noisy features MAP rule to quadratic discriminant analysis for linear systems. In Section IV, we summarize the recently proposed *expected kernel* classifier which treats the training feature vectors as random [2, 3], and investigate different expected kernels for the problem of classifying $\hat{x}$. Section V demonstrates the value of these robust classifiers through an illustrative experiment.

2. EQUIVALENCES OF MAP RULES

By the data processing theorem [6], $\hat{x} = Gz + y$ cannot contain more information about x's class label than the less-processed observation z. Thus, the optimal MAP rule for classification of z is $\arg\max_g p(z|g)$. The following proposition establishes conditions on the estimator so that $\arg\max_g p(\hat{x}|g)$ is equivalent to the optimal MAP rule.

Proposition: For $z = Hx + w$, $w \sim \mathcal{N}(0, \sigma_w^2 I)$ and some linear estimate $\hat{x} = Gz + y$, MAP classification of $\hat{x}$ is equivalent to MAP classification of z if $G^T(GG^T)^{-1}G$ is a projection matrix for H such that $G^T(GG^T)^{-1}GH = H$.

Proof. MAP classification of $\hat{x}$ can be expressed:

$$\begin{aligned}\hat{g} &= \arg\max_{g \in \mathcal{G}} p(g|\hat{x}) \\ &= \arg\max_{g \in \mathcal{G}} p(g) \int p(\hat{x}|x) p(x|g) dx. \qquad (1)\end{aligned}$$

MAP classification of z can be expressed:

$$\begin{aligned}\hat{g} &= \arg\max_{g \in \mathcal{G}} p(g|z) \\ &= \arg\max_{g \in \mathcal{G}} p(g) \int p(z|x) p(x|g) dx. \qquad (2)\end{aligned}$$

The decomposition of the MAP rule given in (2) is commonly referred to as the *noisy features rule* [7] and we believe was first applied to classifying noisy test samples by Aitchison and Lauder [8].

We will show that given the proposition's conditions $p(\hat{x}|x)$ is equivalent to $p(z|x)$ up to a constant that does not depend on g, and thus (1) is equivalent to (2). First,

$$\begin{aligned}p(z|x) &= \mathcal{N}(z; Hx, \sigma_w^2 I) \\ &= c_z e^{-\frac{1}{2\sigma_w^2}(x^T H^T - 2z^T)Hx}\end{aligned}$$

where c_z is a constant that does not depend on g or x.

Similarly,

$$\begin{aligned}p(\hat{x}|x) &= \mathcal{N}(\hat{x}; GHx + y, GG^T \sigma_w^2 I) \\ &= \mathcal{N}(Gz; GHx, GG^T \sigma_w^2 I) \\ &= c_{\hat{x}} \cdot e^{-\frac{1}{2\sigma_w^2}(x^T H^T - 2z^T)G^T (GG^T)^{-1} GHx} \\ &= c_{\hat{x}} e^{-\frac{1}{2\sigma_w^2}(x^T H^T - 2z^T)Hx} \\ &= \frac{c_{\hat{x}}}{c_z} p(z|x).\end{aligned}$$

□

It is easy to show that the linear LS estimator, $G = (H^T H)^{-1} H^T$, meets the projection requirement of the above proposition. Furthermore, if we assume that Σ_x is the covariance matrix of x and that σ_w^2 is the noise variance, then G for the LMMSE estimator is written as

$$G = \Sigma_x H (H \Sigma_x H^T + \sigma_w^2 I)^{-1}. \qquad (3)$$

By rewriting (3) using the Woodbury matrix identity [9] as $G = (\Sigma_x^{-1} + H^T H \frac{1}{\sigma_w^2})^{-1} H^T \frac{1}{\sigma_w^2}$, we can see that the LMMSE estimator also meets the projection requirement of the proposition.

3. ADAPTING THE QUADRATIC DISCRIMINANT ANALYSIS CLASSIFIER

A popular model for the class-conditional distribution $p(x|g)$ used in the MAP rule is the Gaussian [10], leading to linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), local Bayesian discriminant analysis (BDA) [11], and Gaussian mixture models. For example, the standard QDA classifier applied to this problem would first form the linear estimate $\hat{x} = Gz + y$, then classify as

$$\hat{g} = \arg\max_{g \in \mathcal{G}} p(g) \mathcal{N}(\hat{x}; \bar{x}_g, \Sigma_g) \qquad (4)$$

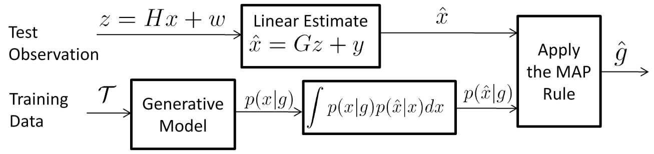

Fig. 2. Block diagram showing showing MAP classification of $\hat{x}$. This approach adapts the clean trained generative function $p(x|g)$ to $p(\hat{x}|g)$ in order to reconcile the difference between training and test distributions.

where $\bar{x}_g$ and Σ_g are the class conditional mean and covariance learned from $\mathcal{T}$. This approach is represented in Figure 1, and is suboptimal because it treats $\hat{x}$ as though it were x.

Better to use the MAP rule given in (1), derived as follows. We assume a Gaussian class-conditional distribution $p(x|g) = \mathcal{N}(x; \bar{x}_g, \Sigma_g)$, and note that $p(\hat{x}|x) = \mathcal{N}(\hat{x}; GHx + y, GG^T \sigma_w^2 I) = \mathcal{N}(GHx + y; \hat{x}, GG^T \sigma_w^2 I)$. Recall the product of Gaussians rule

$$\mathcal{N}(x; a, A)\mathcal{N}(Fx; b, B) = \mathcal{N}(Fa; b, FAF^T + B)\mathcal{N}(x; c, C) \qquad (5)$$

where the constant $C = \left[F^T B^{-1} G + A^{-1}\right]^{-1}$ and $c = C\left[F^T B^{-1} b + A^{-1} a\right]$.

Applying (5) to $p(\hat{x}|x)$ and $p(x|g)$ as in (1) produces the closed-form classifier:

$$\hat{g} = \arg\max_{g \in \mathcal{G}} p(g) \mathcal{N}(\hat{x}; GH\bar{x}_g + y, GH\Sigma_g (GH)^T + GG^T \sigma_w^2). \qquad (6)$$

This approach is depicted in Figure 2.

4. ROBUST DISCRIMINATIVE CLASSIFIERS

Discriminative classifiers classify a test point x by minimizing the empirical risk of a discriminative function over the training set. We write the discriminative function for class g and test point x as $f_g(x, \mathcal{T})$ to indicate that it is a function of the training data as well as the test point and, in some cases, the class label. The standard discriminative approach to classifying $\hat{x}$ would be to simply evaluate $f_g(\hat{x}, \mathcal{T})$ as depicted in Figure 1. This may be suboptimal because it treats $\hat{x}$ as though it were x.

Recent research into robust classifiers proposed an *expected kernel* [3] to make kernel classifiers such as the support vector machine more robust. Given a channel model and a kernel definition K, a random training sample $Z_i = h * x_i = N$ could be computed, and the expected kernel between two random training samples defined as

$$\begin{aligned} K_z(Z_i, Z_j) &= E_{Z_i|x_i, Z_j|x_j}\left[K(Z_i, Z_j)\right] \\ &= \int\int p(z_i|x_i)p(z_j|x_j)K(z_i, z_j)dz_i dz_j. \quad (7) \end{aligned}$$

We consider instead the random estimates $\{\hat{X}_i\}$ that would be formed from the training signals, with corresponding distributions

$$p(\hat{x}_i|x_i) = \mathcal{N}(\hat{x}_i; GHx_i + y, GG^T\sigma_w^2), \quad (8)$$

and then defining the expected kernel in terms of the random training estimated signals:

$$\begin{aligned} K_e(\hat{X}_i, \hat{X}_j) &= E_{\hat{X}_i|x_i, \hat{X}_j|x_j}\left[K(\hat{X}_i, \hat{X}_j)\right] \\ &= \int\int p(\hat{x}_i|x_i)p(\hat{x}_j|x_j)K(\hat{x}_i, \hat{x}_j)d\hat{x}_i d\hat{x}_j. \end{aligned} \quad (9)$$

This has the advantage over (7) of using the estimator.

Robust classification of $\hat{x}$ is then accomplished by training the discriminative function using the training set of random samples $\tilde{\mathcal{T}} = \{(\hat{X}_i, g_i)\}_{i=1}^M$. At test time, the expected kernel classifier is presented with a test sample that is not random, but is instead a determinstic $\hat{x}$. Therefore, the expected kernel between $\hat{x}$ and $\hat{X}_i$ is:

$$\begin{aligned} K_e(\hat{x}, \hat{X}_i) &= E_{\hat{X}_i|x_i}\left[K(\hat{x}, \hat{X}_i)\right] \\ &= \int p(\hat{x}_i|x_i)K(\hat{x}, \hat{x}_i)d\hat{x}_i. \quad (10) \end{aligned}$$

It was shown in [3] that if we use a Gaussian radial basis function as the base kernel with bandwidth parameter γ,

$$K(\hat{x}, \hat{x}_i) = \mathcal{N}(\hat{x}; \hat{x}_i, \gamma^{-1}I), \quad (11)$$

then we can again use the product of Gaussians rule (5) to develop a closed form solution for (9) and (10). Figure 3 gives a block diagram for expected kernel classification of $\hat{x}$.

The expected kernel rule can be used in any discriminative classifier that relies on a kernel function. One important distinction between the expected kernel classifier and the QDA classifier developed in section 3 is that the results for the expected kernel classifier depend on the choice of estimate $\hat{x}$, for instance, whether LS or MMSE is used. This will be apparent in the results section, where we apply this rule to a support vector machine (SVM).

5. EXPERIMENTS

We present experimental results to illustrate the advantages of using the robust generative and discriminative classifiers compared to the non-robust approach of classifying an $\hat{x}$ as though it were a true x. We used the benchmark optical dataset from the UCI Machine Learing Repository. The dataset consists of 8x8 pixel images of handwritten digits 0 through 9,with 3823 training and 1797 test images. The test and training images are adjusted so that each of the 64 pixels has mean 0 and standard deviation 1. We corrupt the test data with a Gaussian blur matrix with 4 by 4 pixel support and standard deviation 0.5 then add AWGN with varying standard deviation.

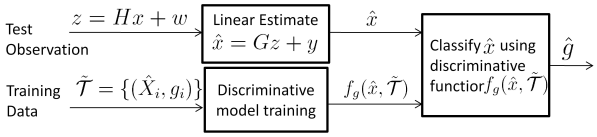

Fig. 3. Expected kernel classification of test feature vector estimate $\hat{x}$.

In the experiment, we compare robust and non-robust versions of a local Bayesian QDA (local BDA) classifier, described in [12]. We show results for the non-robust classification of $\hat{x}$ estimated using both the LS and LMMSE estimator. We also show results using the noisy features rule given in (6) applied to local BDA, which was shown in the Proposition to be the same whether classifying the LS or LMMSE $\hat{x}$ or z. In addition, we compared to non-robust versions of the SVM using the LS and LMMSE estimates $\hat{x}$ as well as the expected kernel rule for classification of LS and LMMSE $\hat{x}$ and z directly. In this case, the robust approaches are not equivalent, as is shown in the figure.

Figure 5 shows that the best performing method was the noisy features rule using local BDA. The best performing SVM method was classification of LMMSE $\hat{x}$ using the expected kernel for LMMSE $\hat{x}$. The results show that the performance of the expected kernel SVM is highly dependent on whether we define the expected kernel (and classify) in terms of LMMSE or LS $\hat{x}$ or z directly. We believe that one of the main reasons for this is the selection of the bandwidth parameter γ in the base kernel (11). As standard [10], we selected the bandwidth parameter by cross validation, and in Figure 5 we show the parameter selected by each of the different SVM methods. This figure shows that the bandwidth parameter is very sensitive to noise standard deviation in all cases other than the expected kernel with LMMSE $\hat{x}$, and this noise sensitivity makes the bandwidth parameter very hard to tune in practice.

6. CONCLUSIONS

We have proposed and compared robust classifiers based on the noisy features rule and the expected kernel, and illustrated that the naive approach of treating an estimate $\hat{x}$ as the true x is suboptimal and relies heavily on the accuracy of the estimate.

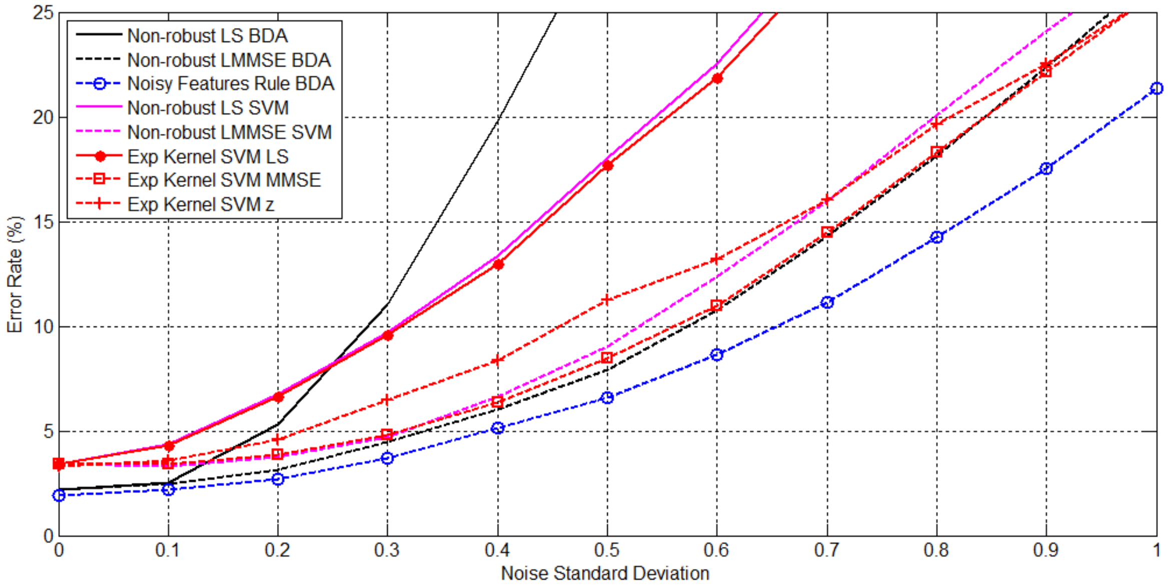

Fig. 4. Error rate of the robust and non-robust classification methods.

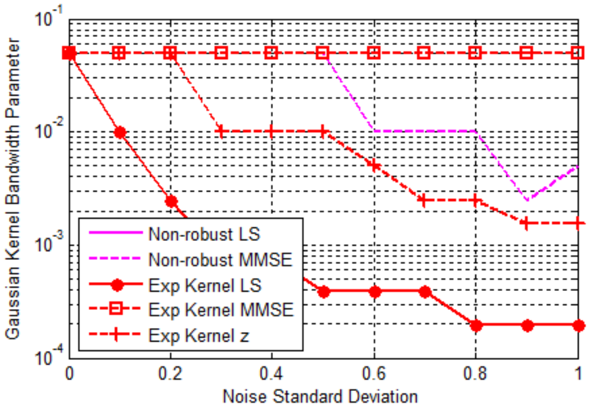

Fig. 5. Gaussian bandwidth parameter chosen during five-fold cross validation (CV) by each of the different SVM classification methods.

7. REFERENCES

[1] J. Flusser, T. Suk, and S. Saic, "Recognition of blurred images by the method of moments," *IEEE Trans. Image Proc.*, vol. 5, pp. 533–538, 1996.

[2] H. S. Anderson and M. R. Gupta, "Joint deconvolution and classification with applications to passive acoustic underwater multipath," *J. Acous. Soc. Am.*, vol. 124, no. 5, pp. 2973–2983, 2008.

[3] H. S. Anderson, M. R. Gupta, E. R. Swanson, and K Jamieson, "Channel-robust classifiers," *To appear in IEEE Trans. Signal Proc.*

[4] A. J. Llorens, T. L. Philip, I. W. Schurman, and C. R. Lorenz, "Enhancing passive automation performance using an acoustic propagation simulation," *J. Acous. Soc. Am.*, vol. 125, no. 4, pp. 2577–2577, 2009.

[5] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, 1993.

[6] T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley and Sons, United States of America, 1991.

[7] R. Duda, E. Hart, and D. Stork, *Pattern Classification*, Wiley-Interscience, 2nd edition, 2001.

[8] J. Aitchison and I. J. Lauder, "Statistical diagnosis from imprecise data," *Biometrika*, vol. 66, no. 3, pp. 475–483, 1979.

[9] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," Oct 2008.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, New York, 2001.

[11] E. K. Garcia, S. Feldman, M. R. Gupta, and S. Srivastava, "Completely lazy learning," *IEEE Trans. Knowledge Data Engineering*, vol. 22, no. 9, pp. 1274 – 1285, 2010.

[12] S. Srivastava, M. R. Gupta, and B. A. Frigyik, "Bayesian quadratic discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1287–1314, 2007.