APPLICATION OF THE SAMPLE-CONDITIONED MSE TO NON-LINEAR CLASSIFICATION AND CENSORED SAMPLING

Lori A. Dalton

Department of Electrical and Computer Engineering and Department of Biomedical Informatics
The Ohio State University, Columbus, OH USA

ABSTRACT

Phenotype discrimination problems in biomedicine typically classify between types of pathology, stages of disease, response to treatment or survivability. In contrast to the usual heuristic classifier and error estimate computed from small sample data, recent work proposes a Bayesian modeling framework over an uncertainty class of feature-label distributions, which when combined with data facilitates optimal MMSE error estimation, optimal classifier design and a sample-conditioned MSE for error estimation analysis, all relative to uncertainty in the underlying distributions conditioned on the sample. Here we address application of the conditional MSE to non-linear classifiers and present an example with optimal Bayesian classification and censored sampling, an economical sampling procedure in which data are collected incrementally until desired criteria are met.

Index Terms— Bayesian estimation, minimum mean-square error, classification error, genomics, small samples

1. INTRODUCTION

The use of microarrays, next-generation sequencing and other high-throughput genomic and proteomic technologies is largely constrained by cost and the inherent difficultly of obtaining large biological samples for phenotype classifier training. It is precisely this kind of small-sample setting where classifier error estimation accuracy becomes a critical issue because it is the primary measure of the scientific validity of a classifier model, and accurate estimation is far from guaranteed. Indeed, classical classifier error estimation methods, such as cross-validation and bootstrap, are typically heuristic methods, with a number of studies demonstrating poor performance under small-sample high-throughput conditions [1, 2]. We focus on the MSE or its square root, the root-mean-square (RMS), as a measure of validity.

Classical error estimator analysis conditions on a fixed distribution and averages over the corresponding sampling distribution. Since the true distribution is usually unknown in practice, one may turn to "distribution free" bounds on error estimator accuracy. However, there are only a few cases where such bounds are available, and even when available these are typically too loose to be useful. For example, consider the following distribution free RMS bound for the leave-one-out error estimator with the discrete histogram rule and tie-breaking in the direction of class 0 [3]:

$$\mathrm{RMS}(\widehat{\varepsilon}_{\mathrm{loo}}|F) \leq \sqrt{\frac{1+6/e}{n} + \frac{6}{\sqrt{\pi(n-1)}}}, \qquad (1)$$

where F is the feature-label distribution and n is the sample size. One needs at least $n = 200$ points to achieve a bound of 0.506. Accurate distribution-free small-sample error estimation is an illusion [4].

Given the necessity of distributional assumptions, why not state them outright and fully integrate them into the analysis? This is precisely what is accomplished in a recent Bayesian framework for classification [5]. Rather than condition on an unknown distribution with uncertainty relative to the sampling distribution, we condition on the observed sample in hand with uncertainty now relative to the unknown feature-label distribution. The theory facilitates the optimization and analysis of both error estimation and classification, resulting in the *Bayesian MMSE error estimator* (BEE) and the *optimal Bayesian classifier* (OBC) [5]. Perhaps more importantly, we can assess the accuracy of an arbitrary error estimator via the *sample-conditioned RMS* [6], which addresses exact performance given the observed data, trained classifier and computed error estimate in hand.

Although closed form solutions for the conditional RMS have been found for Gaussian models under linear classification [6], OBCs found under the same models are generally non-linear, necessitating efficient methods to evaluate the conditional RMS under classifiers of arbitrary form. In this work, we present a novel method to address this problem. It thus becomes practical to evaluate not only the optimal classifier and BEE of any classifier under Gaussian models, but also the conditional RMS of any classifier and error estimator pair. A thorough review of recent work based on the Bayesian framework is given in Section 2, followed by an overview of the Gaussian model in Section 3. Section 4 outlines the proposed method. Finally, in Section 5 we provide an example application in censored sampling, where sample points are acquired incrementally until the estimated error and conditional RMS reach desired levels.

2. REVIEW OF THE BAYESIAN FRAMEWORK

Consider a binary classification problem with sample space $\mathcal{X}$ and class labels 0 and 1. We denote the *a priori* probability that a point is from class 0 by c and the class-y-conditional distribution, $y \in \{0,1\}$, by $f_{\theta_y}(\mathbf{x}|y)$, where $\mathbf{x} \in \mathcal{X}$ is a feature vector and θ_y is a fixed parameter. The feature-label distribution is completely characterized by $\theta = [c, \theta_0, \theta_1]$.

The misclassification rate of classifier $\psi : \mathcal{X} \rightarrow \{0,1\}$ is the probability of mislabeling a sample point, which is defined relative to the underlying feature-label distribution. For a fixed parameter, θ, and fixed classifier, ψ, the true error is

$$\varepsilon(\theta, \psi) = c\varepsilon^0(\theta_0, \psi) + (1-c)\varepsilon^1(\theta_1, \psi), \tag{2}$$

where ε^y is the probability of mislabeling a class y point, i.e., $\varepsilon^y(\theta_y, \psi) = \int_{\{\mathbf{x} \in \mathcal{X} : \psi(\mathbf{x}) \neq y\}} f_{\theta_y}(\mathbf{x}|y) d\mathbf{x}$. For a fixed θ, the optimal classifier, or *Bayes classifier*, is:

$$\psi_{\text{Bayes}}(\mathbf{x}) = \begin{cases} 0 & \text{if } cf_{\theta_0}(\mathbf{x}|0) \geq (1-c)f_{\theta_1}(\mathbf{x}|1), \\ 1 & \text{otherwise.} \end{cases} \tag{3}$$

In practice the feature-label distribution is unknown, so that we must train a classifier and estimate the error with data. The Bayesian framework assumes a parameterized uncertainty class of feature-label distributions, i.e., we define parameter spaces such that $c \in [0,1]$, $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$. Using either a non-informative approach or expert information, we assign a "prior" distribution to θ. To facilitate analytic representations, we assume that c and $[\theta_0, \theta_1]$ are independent prior to observing the data, and denote the marginal priors by $\pi(c)$, $\pi(\theta_0)$ and $\pi(\theta_1)$. After observing the sample, the priors are updated to posterior densities, $\pi^*(c)$, $\pi^*(\theta_0)$ and $\pi^*(\theta_1)$, where independence is preserved. Throughout, let S_n be a sample of size n drawn from the sample space $\mathcal{X}$, and let n_y denote the number of points in each class $y \in \{0,1\}$.

Given a beta prior on c with hyperparameters α and β, a special case being when $\alpha = \beta = 1$ for uniform c, then for a random sample $\pi^*(c)$ is still beta with hyperparameters $\alpha^* = \alpha + n_0$ and $\beta^* = \beta + n_1$. Furthermore, the posterior probability that a sample point is from class 0 is

$$\widehat{c}(S_n) \equiv E_{\pi^*}[c] = \tfrac{\alpha^*}{\alpha^* + \beta^*} = \tfrac{n_0 + \alpha}{n + \alpha + \beta}, \tag{4}$$

where E_{π^*} is a conditional expectation given the sample.

The posterior distribution for θ_y, $y \in \{0,1\}$, is

$$\pi^*(\theta_y) \propto \pi(\theta_y) \textstyle\prod_{i=1}^{n_y} f_{\theta_y}(\mathbf{x}_i^y|y),$$

where $\mathbf{x}_i^y$ is the i^{th} sample point in class y and the product on the right is the usual "likelihood function." The constant of proportionality is found by normalizing the integral of $\pi^*(\theta_y)$ to 1. When the prior density is proper this follows from Bayes' rule, and if $\pi(\theta_y)$ is improper this is taken as a definition. Two important models for the θ_y have been considered: multinomial distributions with Dirichlet priors on the bin probabilities (henceforth referred to as the discrete model) and Gaussian distributions with normal-inverse-Wishart priors on the mean and covariance pair (the Gaussian model).

As a modeling assumption, priors quantify the uncertainty we have about the distribution before observing the data. We have the option of using non-informative or flat priors, as long as the posterior is normalizeable. Alternatively, informative priors can supplement the classification problem with expert information to make the problem tractable or to improve performance with distributional information when the sample size is small. This is key for problems constrained by a lack of information or expensive information. Whatever prior is used, it has been proven in [7] that under mild regularity conditions the posteriors converge to delta functions on the true parameters for both the discrete and Gaussian models. More informative priors may help the posteriors converge faster, but as long as the prior does not exclude the true distribution as impossible, convergence is assured.

2.1. Bayesian MMSE Classifier Error Estimation

A Bayesian error estimator is a classical MMSE estimator, equivalent to the conditional expectation of the true error given observed measurements. Given a sample, S_n, and any fixed classifier, ψ, the BEE (a function of S_n and ψ) for the true classifier error (a function of θ and ψ) is

$$\begin{aligned} \widehat{\varepsilon}(S_n, \psi) &= \mathrm{E}_{c,\theta_0,\theta_1}[\varepsilon(\theta,\psi)|S_n] \\ &= \widehat{c}(S_n)\widehat{\varepsilon}^0(S_n,\psi) + (1-\widehat{c}(S_n))\widehat{\varepsilon}^1(S_n,\psi), \end{aligned} \tag{5}$$

where we have used the posterior independence between c and $[\theta_0, \theta_1]$. We also define $\widehat{\varepsilon}^y(S_n, \psi) = \mathrm{E}_{\pi^*}[\varepsilon^y(\theta_y, \psi)]$, which may be viewed as the posterior probability of incorrectly labeling a class y point. When the prior probabilities are improper, this is called the generalized BEE. To evaluate the BEE, $\widehat{c}(S_n)$ depends on our prior model for c, see for instance (4), and $\widehat{\varepsilon}^y(S_n, \phi)$ can be found using the following theorem, originally proved in [5].

Theorem 2.1 *Let ψ be a fixed classifier given by $\psi(\mathbf{x}) = 0$ if $\mathbf{x} \in R_0$ and $\psi(\mathbf{x}) = 1$ if $\mathbf{x} \in R_1$, where R_0 and R_1 are measurable sets partitioning the sample space. Then*

$$\begin{aligned} \widehat{\varepsilon}^y(S_n, \psi) &= \textstyle\int_{R_{1-y}} f(\mathbf{x}|y)\, d\mathbf{x}, \\ f(\mathbf{x}|y) &= \textstyle\int_{\Theta_y} f_{\theta_y}(\mathbf{x}|y)\, \pi^*(\theta_y)\, d\theta_y. \end{aligned} \tag{6}$$

The Bayesian error estimator has been solved in closed form for both discrete models under arbitrary classifiers and Gaussian models under linear classifiers. When closed-form solutions are not available, the expected error of a classifier may be found via Monte-Carlo integral approximation by drawing a large synthetic sample from the effective densities $f(\mathbf{x}|y)$ and evaluating the proportion of misclassified points. Classical frequentist consistency also holds for Bayesian error estimators on fixed distributions in the parameterized family owing to the convergence of posteriors in both the discrete and Gaussian models [7]. A number of practical considerations for Bayesian error estimation have also been addressed, including robustness to false modeling assumptions and application to microarray data analysis [8], where priors are calibrated with a method-of-moments approach using features from the microarray dataset that are discarded by feature selection. Performance is often superior to classical error estimation schemes on real gene expression data.

2.2. Optimal Bayesian Classification

An optimal Bayesian classifier (OBC) is defined to be any classifier that minimizes the expected error:

$$\psi_{\mathrm{OBC}} = \arg\inf_{\psi\in\mathcal{C}} \mathrm{E}_{\pi^*}\left[\varepsilon\left(\theta,\psi\right)\right] = \arg\inf_{\psi\in\mathcal{C}} \widehat{\varepsilon}\left(S_n,\psi\right), \qquad (7)$$

where $\mathcal{C}$ is an arbitrary family of classifiers. If $\mathcal{C}$ is the set of all classifiers with measurable decision regions, then the following theorem, originally proved in [5], states that the optimal classifier is equivalent to the Bayes classifier under the effective model.

Theorem 2.2 *An OBC classifier,* ψ_{OBC}*, satisfying* (7)*, where* $\mathcal{C}$ *is the set of all classifiers with measurable decision regions, exists and is given pointwise by*

$$\psi_{\mathrm{OBC}}\left(\mathbf{x}\right) = \begin{cases} 0 & \textit{if } \widehat{c}(S_n) f\left(\mathbf{x}|0\right) \geq (1-\widehat{c}(S_n)) f\left(\mathbf{x}|1\right), \\ 1 & \textit{otherwise}, \end{cases}$$

where $f\left(\mathbf{x}|y\right)$ *is defined in* (6)*.*

We call $f\left(\mathbf{x}|y\right)$ the *effective class-conditional density*, which is a pointwise average of the class-conditional densities $f_{\theta_y}\left(\mathbf{x}|y\right)$ considered in the model and depends on the sample because it depends on π^*. When we substitute $\widehat{c}(S_n)$ for the true *a priori* probability of a class 0 point, c, and $f\left(\mathbf{x}|y\right)$ for the true class-conditional density, $f_{\theta_y}\left(\mathbf{x}|y\right)$, for both $y = 0$ and $y = 1$, then the BEE for any classifier is solved by finding the misclassification error of the classifier under the effective model, and the OBC is solved by finding the Bayes classifier under the effective model.

OBCs have been solved in closed form for both the discrete and Gaussian models [5]. A number of important properties have also been shown, including invariance to invertible transformations, pointwise convergence to the Bayes classifier, and robustness to false modeling assumptions.

2.3. Bayesian Sample-Conditioned MSE

The sample-conditioned MSE of a BEE $\widehat{\varepsilon}$ quantifies the accuracy of $\widehat{\varepsilon}$ as an estimator of ε, conditioned on the actual sample in hand. It is defined as [6],

$$\begin{aligned} \mathrm{MSE}(\widehat{\varepsilon}(S_n,\psi)|S_n) &= \mathrm{E}_{\pi^*}[(\varepsilon\left(\theta,\psi\right) - \widehat{\varepsilon}(S_n,\psi))^2] \\ &= \mathrm{Var}_{\pi^*}\left(\varepsilon(\theta,\psi)\right). \end{aligned} \qquad (8)$$

The BEE is the first moment of the true error and the conditional MSE is the second central moment, both conditioned on the observed sample. Thanks to the posterior independence between c and $[\theta_0,\theta_1]$, we decompose the conditional MSE via the basic variance identity,

$$\begin{aligned} \mathrm{MSE}(\widehat{\varepsilon}(S_n,\psi)|S_n) = (\widehat{\varepsilon}^0(S_n,\psi) - \widehat{\varepsilon}^1(S_n,\psi))^2 \mathrm{Var}_{\pi^*}\left(c\right) \\ + \mathrm{E}_{\pi^*}\left[c^2\right] \mathrm{MSE}(\widehat{\varepsilon}^0(S_n,\psi)|S_n) \\ + \mathrm{E}_{\pi^*}\left[(1-c)^2\right] \mathrm{MSE}(\widehat{\varepsilon}^1(S_n,\psi)|S_n), \end{aligned}$$

where $\mathrm{MSE}(\widehat{\varepsilon}^y(S_n,\psi)|S_n) = \mathrm{Var}_{\pi^*}\left(\varepsilon^y(\theta_y,\psi)\right)$. Moments related to c depend on our prior model for c, but are straightforward to find for a given posterior $\pi^*(c)$. Furthermore,

$$\mathrm{MSE}(\widehat{\varepsilon}^y(S_n,\psi)|S_n) = \mathrm{E}_{\pi^*}\left[\left(\varepsilon^y(\theta_y,\psi)\right)^2\right] - \left(\widehat{\varepsilon}^y(S_n,\psi)\right)^2,$$

so that the conditional MSE reduces to finding $\widehat{\varepsilon}^y$ (a part of the BEE) and $\mathrm{E}_{\pi^*}[(\varepsilon^y(\theta_y,\psi))^2]$ for both classes. The sample-conditioned MSE for a BEE converges to zero almost surely for both discrete models under arbitrary classifiers and Gaussian models under linear classifiers, where closed form expressions for the MSE are also available [7].

The conditional MSE for an arbitrary error estimate, $\widehat{\varepsilon}_{\bullet}(S_n,\psi)$, can also be evaluated from a given sample:

$$\begin{aligned} &\mathrm{MSE}(\widehat{\varepsilon}_{\bullet}(S_n,\psi)|S_n) \\ &= \mathrm{MSE}(\widehat{\varepsilon}(S_n,\psi)|S_n) + (\widehat{\varepsilon}(S_n,\psi) - \widehat{\varepsilon}_{\bullet}(S_n,\psi))^2. \end{aligned} \qquad (9)$$

Note the optimality of Bayesian error estimation.

Consider a typical classification scenario in which we train a classifier from data and use the same data to estimate the error of this classifier. A key question arises: How close is the estimate to the actual error? Whereas in a classical distribution-free approach nothing can be said given a single sample, the Bayesian approach answers this question with the sample-conditioned MSE. The sample conditions uncertainty, and different samples condition it to different extents.

3. THE GAUSSIAN MODEL

Suppose each sample point is a column vector of D features, where the class-y-conditional distribution is Gaussian with parameter $\theta_y = [\mu_y, \Sigma_y]$, μ_y being the mean and Σ_y the covariance. The parameter space of Σ_y consists of all positive definite (valid covariance) matrices. Herein we assume θ_0 and θ_1 are independent prior to observing the data, although a homoscedastic covariance model has also been treated in [5].

We assume a conjugate prior where Σ_y is invertible with probability 1, and for invertible Σ_y we have

$$\pi(\theta_y) = \pi(\mu_y|\Sigma_y)\pi(\Sigma_y), \qquad (10)$$

where given hyperparameters consisting of a constant ν_y, constant κ_y, length D vector $\mathbf{m}_y$ and $D \times D$ matrix S_y,

$$\begin{aligned} \pi(\mu_y|\Sigma_y) \propto |\Sigma_y|^{-\frac{1}{2}} \exp\left(-\tfrac{\nu_y}{2}\left(\mu_y - \mathbf{m}_y\right)^T \Sigma_y^{-1}\left(\mu_y - \mathbf{m}_y\right)\right) \\ \pi(\Sigma_y) \propto |\Sigma_y|^{-\frac{\kappa_y+D+1}{2}} \exp\left(-\tfrac{1}{2}\mathrm{trace}\left(S_y\Sigma_y^{-1}\right)\right). \end{aligned}$$

If $\nu_y > 0$, $\kappa_y > D-1$ and S_y is positive definite, then this is a proper prior [9] where $\pi(\mu_y|\Sigma_y)$ is Gaussian with mean $\mathbf{m}_y$ and covariance Σ_y/ν_y and $\pi(\Sigma_y)$ is an inverse-Wishart distribution such that $\mathrm{E}_\pi[\Sigma_y] = S_y/(\kappa_y - D - 1)$.

It can be shown that the posterior, $\pi^*(\theta_y)$, has the same form as the prior with updated hyperparameters $\nu_y^* = \nu_y + n_y$, $\kappa_y^* = \kappa_y + n_y$, $\mathbf{m}_y^* = \frac{\nu_y \mathbf{m}_y + n_y \widehat{\mu}_y}{\nu_y + n_y}$, and

$$S_y^* = S_y + (n_y - 1)\widehat{\Sigma}_y + \tfrac{\nu_y n_y}{\nu_y + n_y}(\widehat{\mu}_y - \mathbf{m}_y)(\widehat{\mu}_y - \mathbf{m}_y)^T,$$

where $\widehat{\mu}_y$ and $\widehat{\Sigma}_y$ are the usual sample mean and covariance, respectively, of the n_y points in class y. The posteriors are proper if $\nu_y^* > 0$, $\kappa_y^* > D - 1$ and S_y^* is positive definite.

The effective density is a multivariate student's t distribution having location vector $\mathbf{m}_y^*$, scale matrix $\Psi_y = \frac{\nu_y^*+1}{(\kappa_y^* - D + 1)\nu_y^*} S_y^*$ and $k_y = \kappa_y^* - D + 1$ degrees of freedom:

$$f(\mathbf{x}|y) \propto \left(1 + \tfrac{1}{k_y}\left(\mathbf{x} - \mathbf{m}_y^*\right)^T \Psi_y^{-1} \left(\mathbf{x} - \mathbf{m}_y^*\right)\right)^{-\frac{k_y + D}{2}}.$$

As long as π^* is proper, the effective density is also proper.

The OBC classifier can be expressed as $\psi_{\mathrm{OBC}}(\mathbf{x}) = 0$ if $g_{\mathrm{OBC}}(\mathbf{x}) \leq 0$ and $\psi_{\mathrm{OBC}}(\mathbf{x}) = 1$ if $g_{\mathrm{OBC}}(\mathbf{x}) > 0$, where

$$\begin{aligned} g_{\mathrm{OBC}}(\mathbf{x}) = K &\left(1 + \tfrac{1}{k_0}\left(\mathbf{x} - \mathbf{m}_0^*\right)^T \Psi_0^{-1}\left(\mathbf{x} - \mathbf{m}_0^*\right)\right)^{k_0 + D} \\ &- \left(1 + \tfrac{1}{k_1}\left(\mathbf{x} - \mathbf{m}_1^*\right)^T \Psi_1^{-1}\left(\mathbf{x} - \mathbf{m}_1^*\right)\right)^{k_1 + D}, \\ K = &\left(\tfrac{1 - \widehat{c}(S_n)}{\widehat{c}(S_n)}\right)^2 \left(\tfrac{k_0}{k_1}\right)^D \tfrac{|\Psi_0|}{|\Psi_1|} \left(\tfrac{\Gamma(k_0/2)\Gamma((k_1+D)/2)}{\Gamma((k_0+D)/2)\Gamma(k_1/2)}\right)^2. \end{aligned}$$

This classifier has a polynomial decision boundary whenever κ_0 and κ_1 are integers. In particular, although we only consider Gaussian distributions in our model, the OBC is not necessarily linear or even quadratic.

4. EVALUATING THE CONDITIONAL MSE

To evaluate the conditional MSE of the BEE for classifier ψ,

$$\mathrm{E}_{\pi^*}\left[\left(\varepsilon^y(\theta_y, \psi)\right)^2\right] = \textstyle\int_{R_{1-y}} \int_{R_{1-y}} g(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z}, \qquad (11)$$

where $g(\mathbf{x}, \mathbf{z})$ is a valid joint density function:

$$g(\mathbf{x}, \mathbf{z}) = \textstyle\int_{\Theta_y} f_{\theta_y}(\mathbf{x}|y) f_{\theta_y}(\mathbf{z}|y) \pi^*(\theta_y) d\theta_y.$$

For the Gaussian model, after some simplification,

$$\begin{aligned} g(\mathbf{x}, \mathbf{z}) \propto \Big| &\tfrac{2\nu_y^*}{\nu_y^* + 2}\left(\tfrac{\mathbf{x}+\mathbf{z}}{2} - \mathbf{m}_y^*\right)\left(\tfrac{\mathbf{x}+\mathbf{z}}{2} - \mathbf{m}_y^*\right)^T \\ &+ \tfrac{1}{2}\left(\mathbf{x} - \mathbf{z}\right)\left(\mathbf{x} - \mathbf{z}\right)^T + S^*\Big|^{-\frac{\kappa^* + 2}{2}}. \end{aligned}$$

Applying a change of variables, $\mathbf{a} = \sqrt{\frac{\nu_y^*}{2\nu_y^* + 4}}\left(\mathbf{x} + \mathbf{z}\right) - \sqrt{\frac{2\nu_y^*}{\nu_y^* + 2}}\mathbf{m}_y^*$ and $\mathbf{b} = \frac{1}{\sqrt{2}}\left(\mathbf{x} - \mathbf{z}\right)$, we obtain a new density

$$h(\mathbf{a}, \mathbf{b}) \propto \left|\mathbf{a}\mathbf{a}^T + \mathbf{b}\mathbf{b}^T + S^*\right|^{-\frac{\kappa^* + 2}{2}}. \qquad (12)$$

The marginal density of $\mathbf{b}$ is a multivariate student's t distribution with zero mean, $\kappa^* - D + 1$ degrees of freedom and scale matrix $\frac{1}{\kappa^* - D + 1} S^*$. The conditional density of $\mathbf{a}$ given $\mathbf{b}$ is also a multivariate student's t distribution with zero mean, $\kappa^* - D + 2$ degrees of freedom and scale matrix $\frac{1}{\kappa^* - D + 2}\left(\mathbf{b}\mathbf{b}^T + S^*\right)$.

To evaluate the conditional MSE for arbitrary classifiers under a Gaussian model when closed-form solutions are unavailable, we may first draw a large synthetic sample from the marginal of $\mathbf{b}$, draw a single realization $\mathbf{a}$ for each $\mathbf{b}$ using the conditional density, and finally apply the following inverse transformation to each $(\mathbf{a}, \mathbf{b})$ pair:

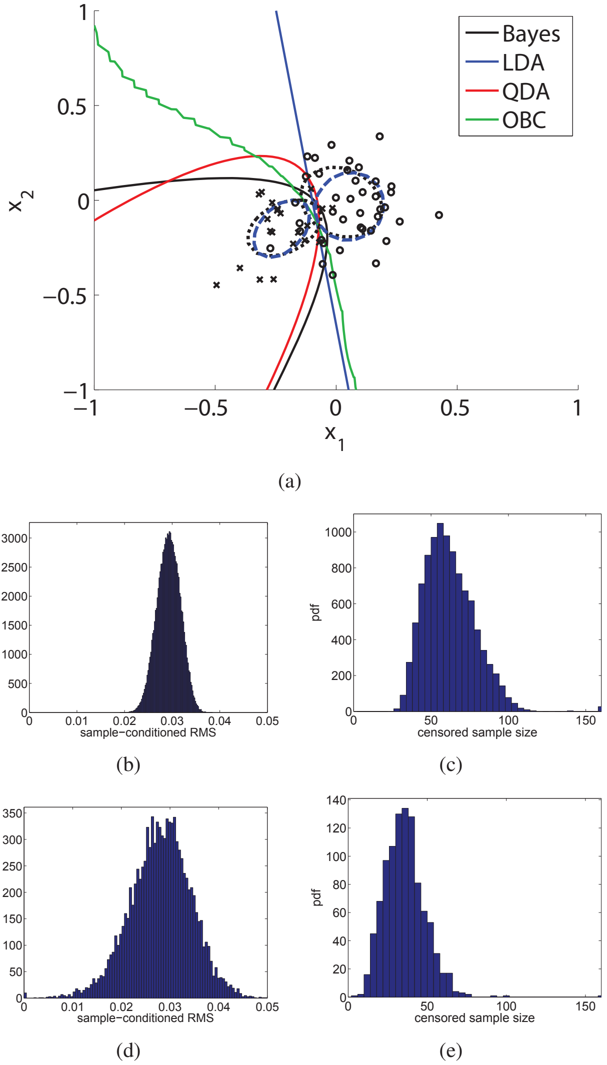
$$\mathbf{x} = \mathbf{m}_y^* + \sqrt{\tfrac{\nu_y^* + 2}{2\nu_y^*}}\mathbf{a} + \tfrac{1}{\sqrt{2}}\mathbf{b}, \quad \mathbf{z} = \mathbf{m}_y^* + \sqrt{\tfrac{\nu_y^* + 2}{2\nu_y^*}}\mathbf{a} - \tfrac{1}{\sqrt{2}}\mathbf{b}.$$

Effectively, this procedure generates a large synthetic sample from the joint density $g\left(\mathbf{x}, \mathbf{z}\right)$. Then, (11) is approximated by evaluating the proportion of pairs for which both $\psi(\mathbf{x}) \neq y$ and $\psi(\mathbf{z}) \neq y$, i.e., both points are misclassified.

5. CENSORED SAMPLING

Given a method to approximate the conditional MSE, it is now possible to apply censored sampling with non-linear classification like the OBC. An example is provided in Fig. 1 for a synthetic Gaussian model with $D = 2$ features, $c = 0.5$ and known "medium information" priors for θ_0 and θ_1 from [7]. In all simulations, distributions are drawn randomly from the prior, with an average Bayes error of 0.158. Part (a) demonstrates a typical distribution (level curves with dotted black lines), sample (class 0 marked with 'o' and 1 with 'x'), estimated densities (level curves with dashed blue lines), and decision boundaries for the Bayes classifier, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and the OBC. For LDA, parts (b) and (c) show a probability density of the conditional RMS for fixed random samples with $n = 60$ (the average true error is 0.170 and the square root of the average MSE is 0.0295) and a probability density of the censored sample size (the average true error is 0.169, the root of the average MSE is 0.029, and the average sample size is 62.1), respectively. In our implementation of censored sampling, each sample is initialized with 2 points in each class, and 4 points with random labels are added in each iteration up to a maximum of 160 points. The stopping criteria are $\mathrm{RMS}(\widehat{\varepsilon}|S_n) \leq 0.0295$ (corresponding to the root of the average MSE with LDA and $n = 60$ fixed), $\mathrm{RMS}(\widehat{\varepsilon}^y|S_n) \leq 0.0295 \times 2$, $\widehat{\varepsilon} \leq 0.3$, and $\widehat{\varepsilon}^y \leq 0.35$, for $y \in \{0, 1\}$. Parts (d) and (e) are analogous for OBC. In (d) the average true error is 0.162 and the root of the average MSE is 0.0288, and in (e) the average true error is 0.165, the root of the average MSE is 0.0258, and the average sample size is 35.8.

The average Bayes error for this model is well below 0.3, so in most cases the conditional RMS determines the censored



(a)

(b)

(c)

(d)

(e)

Fig. 1. Classification with $D = 2$ features, $c = 0.5$ and medium information priors. (a) Example classifiers, $n = 60$; (b) Density of conditional RMS, LDA, $n = 60$; (c) Density of censored sample size, LDA; (d) Density of conditional RMS, OBC, $n = 60$; (e) Density of censored sample size, OBC.

sample size. Observe for LDA that both fixed and censored sampling result in roughly the same unconditional RMS and average sample size. We trade certainty in sample size and uncertainty in performance for uncertainty in sample size and certainty in performance. Applying the same censored sampling criteria with OBCs results in a smaller average error and much smaller sample size on average relative to LDA.

6. CONCLUSION

There are numerous benefits to a Bayesian framework: it can easily integrate expert knowledge or target moderately difficult classification problems with Bayes errors in the mid range [8], it facilitates optimal expected error classification and MMSE error estimation. It also gives rise to the sample-conditioned MSE, a new and practical tool to measure error estimation accuracy with important applications in censored sampling. This work extends the theory further with a practical method to approximate the conditional MSE under Gaussian models when closed form solutions are not available.

7. REFERENCES

[1] B. Hanczar, J. Hua, and E. R. Dougherty, "Decorrelation of the true and estimated classifier errors in high-dimensional settings," *EURASIP J. Bioinf. Sys. Bio.*, vol. 2007, 2007, Article ID 38473, 12 pages.

[2] U. Braga-Neto and E. R. Dougherty, "Exact performance of error estimators for discrete classifiers," *Pattern Recogn.*, vol. 38, no. 11, pp. 1799–1814, 2005.

[3] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York, 1996.

[4] E. R. Dougherty, A. Zollanvari, and U. M. Braga-Neto, "The illusion of distribution-free small-sample classification in genomics," *Curr. Genomics*, vol. 12, no. 5, pp. 333–341, 2011.

[5] L. A. Dalton and E. R. Dougherty, "Optimal classifiers with minimum expected error within a Bayesian framework–Part I: Discrete and Gaussian models," *Pattern Recog.*, vol. 46, no. 5, pp. 1301–1314, 2013.

[6] L. A. Dalton and E. R. Dougherty, "Exact sample conditioned MSE performance of the Bayesian MMSE estimator for classification error–Part I: Representation," *IEEE Trans. Signal Process.*, vol. 60, no. 5, pp. 2575–2587, 2012.

[7] L. A. Dalton and E. R. Dougherty, "Exact sample conditioned MSE performance of the Bayesian MMSE estimator for classification error–Part II: Consistency and performance analysis," *IEEE Trans. Signal Process.*, vol. 60, no. 5, pp. 2588–2603, 2012.

[8] L. A. Dalton and E. R. Dougherty, "Application of the Bayesian MMSE estimator for classification error to gene expression microarray data," *Bioinf.*, vol. 27, no. 13, pp. 1822–1831, 2011.

[9] Morris H. DeGroot, *Optimal Statistical Decisions*, McGraw-Hill, New York, 1970.