Pattern Recognition 46 (2013) 1288–1300

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Optimal classifiers with minimum expected error within a Bayesian framework — Part II: Properties and performance analysis

Lori A. Dalton [a,*], Edward R. Dougherty [b,c]

[a] The Ohio State University, Department of Electrical and Computer Engineering, 205 Dreese Laboratory, 2015 Neil Avenue, Columbus, OH 43210, USA
[b] Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA
[c] Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ, USA

ARTICLE INFO

Article history:
Received 1 August 2012
Received in revised form
20 September 2012
Accepted 21 October 2012
Available online 2 November 2012

Keywords:
Bayesian estimation
Classification
Error estimation
Genomics
Minimum mean-square estimation
Small samples

ABSTRACT

In part I of this two-part study, we introduced a new optimal Bayesian classification methodology that utilizes the same modeling framework proposed in Bayesian minimum-mean-square error (MMSE) error estimation. Optimal Bayesian classification thus completes a Bayesian theory of classification, where both the classifier error and our estimate of the error may be simultaneously optimized and studied probabilistically within the assumed model. Having developed optimal Bayesian classifiers in discrete and Gaussian models in part I, here we explore properties of optimal Bayesian classifiers, in particular, invariance to invertible transformations, convergence to the Bayes classifier, and a connection to Bayesian robust classifiers. We also explicitly derive optimal Bayesian classifiers with non-informative priors, and explore relationships to linear and quadratic discriminant analysis (LDA and QDA), which may be viewed as plug-in rules under Gaussian modeling assumptions. Finally, we present several simulations addressing the robustness of optimal Bayesian classifiers to false modeling assumptions. Companion website: http://gsp.tamu.edu/Publications/supplementary/dalton12a.

1. Introduction

In the first part of this two-part study [1], we defined an *optimal Bayesian classifier* to be a classifier that minimizes the probability of misclassifying a future point relative to the assumed model conditioned on the observed sample, or equivalently minimizes the Bayesian error estimate. The problem of optimal Bayesian classification over an uncertainty class of feature-label distributions arises naturally from two related sources: the need for accurate classification and the need for accurate error estimation. With small samples, the latter is only possible with application of prior knowledge in conjunction with the sample data. Given prior knowledge, it behooves us to find an optimal error estimator and classifier relative to the prior knowledge. Having found optimal Bayesian error estimators in [2,3], found analytic representation of the MSE of these error estimates in [4,5], and found expressions for optimal Bayesian classifiers in terms of the effective class-conditional densities in [1], here, in part II we examine basic properties of optimal Bayesian classifiers.

We study invariance to invertible transformations in discrete and continuous models, convergence to the Bayes classifier, and a connection to robust classification. The latter is a classical filtering problem [6,7], where in the context of classification one wishes to find an optimal classifier over a parameterized uncertainty class of feature-label distributions absent new data [8]. Heretofore, the robust classification problem had only been solved in a suboptimal manner and now the optimal robust classifier falls out from the theory of optimal Bayesian classification. We also explicitly derive optimal Bayesian classifiers using non-informative priors and, using Gaussian modeling assumptions, compare these to plug-in classification rules, such as linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), which are optimal in fixed Gaussian models with common covariance matrix and different covariance matrices, respectively. Finally, we present several simulations addressing the robustness of optimal Bayesian classifiers to false modeling assumptions. Having some robustness to incorrect modeling assumptions is always important in practice because, even if one utilizes statistical techniques, such as hypothesis tests, for model checking, these can at best, even for very small p values, lead to not rejecting the assumed model.

For the sake of completeness, we begin by stating some key definitions and propositions from Part I [1]. An *optimal Bayesian classifier* is any classifier, ψ_{OBC}, satisfying

$$\mathrm{E}_{\pi^*}[\varepsilon(\theta,\psi_{\mathrm{OBC}})] \le \mathrm{E}_{\pi^*}[\varepsilon(\theta,\psi)], \tag{1}$$

for all $\psi \in \mathcal{C}$, where $\varepsilon(\theta,\psi)$ is the true error of classifier ψ under a

* Corresponding author. Tel.: +1 614 292 4594; fax: +1 614 292 7596.
E-mail addresses: dalton@ece.osu.edu (L.A. Dalton), edward@ece.tamu.edu (E.R. Dougherty).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.10.019

feature-label distribution parameterized by $\theta \in \mathbf{\Theta}$ and $\mathcal{C}$ is an arbitrary family of classifiers. In (1), the expectations are taken relative to a posterior distribution, $\pi^*(\theta)$, on the parameters that is updated from a prior, $\pi(\theta)$, after observing a sample, S_n, of size n. An optimal Bayesian classifier minimizes the Bayesian error estimate, $\widehat{\varepsilon}(\psi,S_n) = \mathrm{E}_{\pi^*}[\varepsilon(\theta,\psi)]$. For a binary classification problem, the Bayesian framework defines $\theta = [c,\theta_0,\theta_1]$, where c is the *a priori* probability that a future point comes from class 0 and θ_0 and θ_1 parameterize the class-0 and class-1-conditional distributions, respectively. For a fixed class, $y \in \{0,1\}$, we let $f_{\theta_y}(\mathbf{x}|y)$ be the class-conditional density parameterized by θ_y and denote the marginal posterior of θ_y by $\pi^*(\theta_y)$. If $\mathrm{E}_{\pi^*}[c] = 0$, then the optimal Bayesian classifier is a constant and always assigns class 1; if $\mathrm{E}_{\pi^*}[c] = 1$ then it always assigns class 0. Hence, we typically assume that $0 < \mathrm{E}_{\pi^*}[c] < 1$. Two important theorems from Part I follow.

Theorem 1 (*Evaluating Bayesian error estimators*). *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 measurable sets R_0 and R_1 partition the sample space. Then*

$$\widehat{\varepsilon}(\psi,S_n) = \mathrm{E}_{\pi^*}[c]\int_{R_1} f(\mathbf{x}|0)\,d\mathbf{x} + (1-\mathrm{E}_{\pi^*}[c])\int_{R_0} f(\mathbf{x}|1)\,d\mathbf{x}, \tag{2}$$

where $\mathbf{I}_E$ is an indicator function equal to one if E is true and zero otherwise, and

$$f(\mathbf{x}|y) = \int_{\mathbf{\Theta}_y} f_{\theta_y}(\mathbf{x}|y)\pi^*(\theta_y)\,d\theta_y, \tag{3}$$

is known as the effective class-conditional density.

Theorem 2 (*Optimal Bayesian classification*). *An optimal Bayesian classifier, ψ_{OBC}, satisfying (1) for all $\psi \in \mathcal{C}$, the set of all classifiers with measurable decision regions, exists and is given pointwise by*

$$\psi_{\mathrm{OBC}}(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathrm{E}_{\pi^*}[c]f(\mathbf{x}|0) \geq (1-\mathrm{E}_{\pi^*}[c])f(\mathbf{x}|1), \\ 1 & \text{otherwise.} \end{cases} \tag{4}$$

2. Transformations of the feature space

Consider an invertible transformation, $t : \mathcal{X} \to \overline{\mathcal{X}}$, mapping from some original feature space, $\mathcal{X}$, to a new space, $\overline{\mathcal{X}}$ (in the continuous case we also assume that the inverse map is continuously differentiable). The following theorem shows that the optimal Bayesian classifier in the transformed space can be found by transforming the optimal Bayesian classifier in the original feature space pointwise, and that both classifiers have the same expected true error.

The advantages of this fundamental property are at least twofold. First, the data can be losslessly preprocessed without affecting optimal classifier design or the expected true error,-which is not true in general,for example with LDA classification under a non-linear transformation. Second, it is possible to solve or interpret optimal classification and error estimation problems by transforming to a more manageable space, similar to the "kernel trick" used to map features to a high-dimensional feature space having a meaningful linear classifier. We denote equivalent constants and functions in the transformed space with an overline, for example we write a point $\mathbf{x}$ in the transformed space as $\overline{\mathbf{x}}$.

Theorem 3 (*Invariance to invertible transformations*). *Consider a Bayesian model with posterior $\pi^*(\theta)$ in either a discrete or Euclidean feature space, $\mathcal{X}$. Suppose ψ_{OBC} is an optimal Bayesian classifier satisfying (1) for all $\psi \in \mathcal{C}$, where $\mathcal{C}$ is a family of classifiers (not necessarily all classifiers) with measurable decision regions. Moreover, suppose that the original sample space is transformed by an invertible mapping t, and that in the continuous case t^{-1} is continuously differentiable with an almost everywhere full rank Jacobian. Then the optimal classifier in the transformed space among $\overline{\mathcal{C}} = \{\overline{\psi} \,|\, \overline{\psi} = \psi \circ t^{-1} \text{ for some } \psi \in \mathcal{C}\}$ is $\overline{\psi}_{\mathrm{OBC}}(\overline{\mathbf{x}}) = \psi_{\mathrm{OBC}}(t^{-1}(\overline{\mathbf{x}}))$ and both classifiers possess the same Bayesian error estimate, $\mathrm{E}_{\pi^*}[\varepsilon(\theta,\psi_{\mathrm{OBC}})]$.*

Proof. For a fixed class $y \in \{0,1\}$, in the continuous case the class-conditional density parameterized by θ_y in the transformed space is $\overline{f}_{\theta_y}(\overline{\mathbf{x}}|y) = f_{\theta_y}(t^{-1}(\overline{\mathbf{x}})|y)|\det(J(\overline{\mathbf{x}}))|$, where $J(\overline{\mathbf{x}})$ is the Jacobian of t^{-1} evaluated at $\overline{\mathbf{x}}$. In the discrete case, $\overline{f}_{\theta_y}(\overline{\mathbf{x}}|y) = f_{\theta_y}(t^{-1}(\overline{\mathbf{x}})|y)$ and to unify the two cases we say $|\det(J(\overline{\mathbf{x}}))| = 1$. Although each class-conditional density in our model uncertainty class will change with the transformation, each may still be indexed by the same parameter, θ_y, and hence the same prior and posterior may be used in both spaces. The effective class-conditional density is thus given by

$$\begin{aligned} \overline{f}(\overline{\mathbf{x}}|y) &= \int_{\mathbf{\Theta}_y} \overline{f}_{\theta_y}(\overline{\mathbf{x}}|y)\pi^*(\theta_y)\,d\theta_y \\ &= \int_{\mathbf{\Theta}_y} f_{\theta_y}(t^{-1}(\overline{\mathbf{x}})|y)|\det(J(\overline{\mathbf{x}}))|\pi^*(\theta_y)\,d\theta_y \\ &= f(t^{-1}(\overline{\mathbf{x}})|y)|\det(J(\overline{\mathbf{x}}))|. \end{aligned} \tag{5}$$

Let $\psi \in \mathcal{C}$ be an arbitrary fixed classifier given by $\psi(\mathbf{x}) = \mathbf{I}_{\mathbf{x} \in R_1}$, where R_1 is a measurable set in the original sample space. Then $\overline{\psi}(\overline{\mathbf{x}}) = \mathbf{I}_{\overline{\mathbf{x}} \in \overline{R}_1}$ is the equivalent classifier in the transformed space ($\overline{\psi}(\overline{\mathbf{x}}) = \psi(t^{-1}(\overline{\mathbf{x}}))$), where $\overline{R}_1 = \{t(\mathbf{x}) | \mathbf{x} \in R_1\}$. Noting that $\mathrm{E}_{\pi^*}[c]$ remains unchanged, and by Theorem 1 the expected true error of $\overline{\psi}$ is given by

$$\begin{aligned} \mathrm{E}_{\pi^*}[\varepsilon(\theta,\overline{\psi})] &= \mathrm{E}_{\pi^*}[c]\int_{\overline{R}_1} \overline{f}(\overline{\mathbf{x}}|0)\,d\overline{\mathbf{x}} + (1-\mathrm{E}_{\pi^*}[c])\int_{\overline{\mathcal{X}}-\overline{R}_1} \overline{f}(\overline{\mathbf{x}}|1)\,d\overline{\mathbf{x}} \\ &= \mathrm{E}_{\pi^*}[c]\int_{\overline{R}_1} f(t^{-1}(\overline{\mathbf{x}})|0)|\det(J(\overline{\mathbf{x}}))|\,d\overline{\mathbf{x}} \\ &\quad + (1-\mathrm{E}_{\pi^*}[c])\int_{\overline{\mathcal{X}}-\overline{R}_1} f(t^{-1}(\overline{\mathbf{x}})|1)|\det(J(\overline{\mathbf{x}}))|\,d\overline{\mathbf{x}} \\ &= \mathrm{E}_{\pi^*}[c]\int_{R_1} f(\mathbf{x}|0)\,d\mathbf{x} + (1-\mathrm{E}_{\pi^*}[c])\int_{\mathcal{X}-R_1} f(\mathbf{x}|1)\,d\mathbf{x} \\ &= \mathrm{E}_{\pi^*}[\varepsilon(\theta,\psi)], \end{aligned} \tag{6}$$

where the integrals in the second to last line have applied the substitution $\mathbf{x} = t^{-1}(\overline{\mathbf{x}})$. If ψ_{OBC} is an optimal Bayesian classifier in the original space and $\overline{\psi}_{\mathrm{OBC}}$ is the equivalent classifier in the transformed space, then $\overline{\psi}_{\mathrm{OBC}}$ also minimizes expected true error and thus is an optimal Bayesian classifier in the transformed space. □

3. Convergence to the Bayes classifier

A key property of a classification rule is consistency: does the classifier converges to a Bayes classifier as $n \to \infty$? In contrast to the Bayesian modeling framework, analysis in this section uses frequentist asymptotics, which concern behavior with respect to a fixed parameter and its sampling distribution. In particular, the next theorem shows that consistency holds for optimal Bayesian classification, as long as the true distribution is contained in the parameterized family with mild conditions on the prior.

As discussed in [5], we expect the posterior of θ to converge in some sense to the true value of θ. We review the formalities here, and as such we require measure theory and a few definitions. First, if λ_n and λ are probability measures on a measure space $\mathbf{\Theta}$ with Borel σ-algebra $\mathcal{B}_{\mathbf{\Theta}}$, then $\lambda_n \to \lambda$ weak* if and only if $\int f\,d\lambda_n \to \int f\,d\lambda$ for all bounded continuous functions f on $\mathbf{\Theta}$. Second, in a general Bayesian estimation problem where $\overline{\theta}$ is the unknown true parameter in a parameter space, $\mathbf{\Theta}$, and the feature space, $\mathcal{X}$, is a complete separable metric space, let $F_{\overline{\theta}}$ be the probability measure on $\mathcal{X}$ corresponding to the true distribution parameterized by $\overline{\theta}$. Further assume an independent and identically distributed (i.i.d.) sampling process, and let the infinite