OPTIMAL BAYESIAN CLASSIFICATION

Lori A. Dalton

Edward R. Dougherty

SPIE PRESS
Bellingham, Washington USA

Library of Congress Cataloging-in-Publication Data

Names: Dalton, Lori A., author. | Dougherty, Edward R., author.
Title: Optimal Bayesian classification / Lori A. Dalton and Edward R. Dougherty.
Description: Bellingham, Washington : SPIE Press, [2019] | Includes bibliographical references and index.
Identifiers: LCCN 2019028901 (print) | LCCN 2019028902 (ebook) | ISBN 9781510630697 (paperback) | ISBN 9781510630710 (pdf) | ISBN 9781510630703 (epub) | ISBN 9781510630727 (kindle edition)
Subjects: LCSH: Bayesian statistical decision theory. | Statistical decision.
Classification: LCC QA279.5 D355 2020 (print) | LCC QA279.5 (ebook) | DDC 519.5/42–dc23
LC record available at https://lccn.loc.gov/2019028901
LC ebook record available at https://lccn.loc.gov/2019028902

Published by

SPIE
P.O. Box 10
Bellingham, Washington 98227-0010 USA
Phone: +1 360.676.3290
Fax: +1 360.647.1445
Email: books@spie.org
Web: http://spie.org

The content of this book reflects the work and thought of the author. Every effort has been made to publish reliable and accurate information herein, but the publisher is not responsible for the validity of the information or for any outcomes resulting from reliance thereon.

Cover photographs: Matt Anderson Photography/Moment via Getty Images and Jackie Niam/iStock via Getty Images.

Printed in the United States of America.
First Printing.
For updates to this book, visit http://spie.org and type “PM310” in the search field.

Contents

Preface

The most basic problem of engineering is the design of optimal (or close-to-optimal) operators. The design of optimal operators takes different forms depending on the random process constituting the scientific model and the operator class of interest. The operators might be filters, controllers, or classifiers, each having numerous domains of application. The underlying random process might be a random signal/image for filtering, a Markov process for control, or a feature-label distribution for classification. Here we are interested in classification, and an optimal operator is a Bayes classifier, which is a classifier minimizing the classification error.

With sufficient knowledge we can construct the feature-label distribution and thereby find a Bayes classifier. Rarely, and in practice virtually never, do we possess such knowledge. On the other hand, if we had unlimited data, we could accurately estimate the feature-label distribution and obtain a Bayes classifier. Rarely do we possess sufficient data. Therefore, we must use whatever knowledge and data are available to design a classifier whose performance is hopefully close to that of a Bayes classifier.

Classification theory has historically developed mainly on the side of data, the classical case being *linear discriminant analysis (LDA)*, where a Bayes classifier is deduced from a Gaussian model and the parameters of the classifier are estimated from data via maximum-likelihood estimation. The idea is that we have knowledge that the true feature-label distribution is Gaussian (or close to Gaussian) and the data can fill in the parameters, in this case, the mean vectors and common co-variance matrix. Much contemporary work takes an even less knowledge-driven approach by assuming some very general classifier form such as a neural network and estimating the network parameters by fitting the network to the data in some manner. The more general the classifier form, the more parameters to determine, and the more data needed. Moreover, there is growing danger of overfitting the classifier form to the data as the classifier structure becomes more complex. Lack of knowledge presents us with model uncertainty, and hypothesizing a classifier form and then estimating the parameters is an ad hoc way of dealing with that uncertainty. It is ad hoc because the designer postulates a classification rule based on some heuristics and then applies the rule to the data.

This book takes a Bayesian approach to modeling the feature-label distribution and designs an optimal classifier relative to a posterior distribution governing an uncertainty class of feature-label distributions. In this way it takes full advantage of knowledge regarding the underlying system and the available data. Its origins lie in the need to estimate classifier error when there is insufficient data to hold out test data, in which case an optimal error estimate can be obtained relative to the uncertainty class. A natural next step is to forgo classical ad hoc classifier design and simply find an optimal classifier relative to the posterior distribution over the uncertainty class—this being an *optimal Bayesian classifier.*

A critical point is that, in general, for optimal operator design, the prior distribution is not on the parameters of the operator (controller, filter, classifier), but on the unknown parameters of the scientific model, which for classification is the feature-label distribution. If the model were known with certainty, then one would optimize with respect to the known model; if the model is uncertain, then the optimization is naturally extended to include model uncertainty and the prior distribution on that uncertainty. Model uncertainty induces uncertainty on the operator parameters, and the distribution of the latter uncertainty follows from the prior distribution on the model. If one places the prior directly on the operator parameters while ignoring model uncertainty, then there is a *scientific gap*, meaning that the relation between scientific knowledge and operator design is broken.

The first chapter reviews the basics of classification and error estimation. It addresses the issue that confronts much of contemporary science and engineering: How do we characterize validity when data are insufficient for the complexity of the problem? In particular, what can be said regarding the accuracy of an error estimate? This is the most fundamental question for classification since the error estimate characterizes the predictive capacity of a classifier.

Chapter 2 develops the theory of optimal Bayesian error estimation: What is the best estimate of classifier error given our knowledge and the data? It introduces what is perhaps the most important concept in the book: effective class-conditional densities. Optimal classifier design and error estimation for a particular feature-label distribution are based on the class-conditional densities. In the context of an uncertainty class of class-conditional densities, the key role is played by the effective class-conditional densities, which are the expected densities relative to the posterior distribution. The Bayesian *minimum-mean-square error (MMSE)* theory is developed for the discrete multinomial model and several Gaussian models. Sufficient conditions for error-estimation consistency are provided. The chapter closes with a discussion of optimal Bayesian ROC estimation.

Chapter 3 addresses error-estimation accuracy. In the typical ad hoc classification paradigm, there is no way to address the accuracy of a particular error estimate. We can only quantify the *mean-square error (MSE)* of error estimation relative to the sampling distribution. With Bayesian MMSE error estimation, we can compute the MSE of the error estimate conditioned on the actual sample relative to the uncertainty class. The sample-conditioned MSE is studied in the discrete and Gaussian models, and its consistency is established. Because a running MSE calculation can be performed as new sample points are collected, one can do censored sampling: stop sampling when the error estimate and MSE of the error estimate are sufficiently small. Section 3.9 provides double-asymptotic approximations of the first and second moments of the Bayesian MMSE error estimate relative to the sampling distribution and the uncertainty class, thereby providing asymptotic approximation to the MSE, or, its square root, the *root-mean-square (RMS) error*. Double asymptotic convergence means that both the sample size and the dimension of the space increase to infinity at a fixed rate between the two. Even though we omit many theoretical details (referring instead to the literature), this section is rather long, on account of double asymptotics, and contains many complicated equations. Nevertheless, it provides an instructive analysis of the relationship between the conditional and unconditional RMS. Regarding the chapter as a whole, it is not a logical prerequisite for succeeding chapters and can be skipped by those wishing to move directly to classification.

Chapter 4 defines an optimal Bayesian classifier as one possessing minimum expected error relative to the uncertainty class, this expectation agreeing with the Bayesian MMSE error estimate. Optimal Bayesian classifiers are developed for a discrete model and several Gaussian models, and convergence to a Bayes classifier for the true feature-label distribution is studied. The robustness of assumptions on the prior distribution is discussed. The chapter has a section on *intrinsically Bayesian robust* classification, which is equivalent to optimal Bayesian classification with a null dataset. It next has a section showing how missing values in the data are incorporated into the overall optimization without having to implement an intermediate imputation step, which would cause a loss of optimality. The chapter closes with two sections in which sampling is not random. Section 4.10 considers optimal sampling, and Section 4.11 examines the effect of dependent sampling.

Chapter 5 extends the theory to multi-class classification via optimal Bayesian risk classification. It includes evaluation of the sample-conditioned MSE of risk estimation and evaluation of the posterior mixed moments for both the discrete and Gaussian models.

Chapter 6 extends the multi-class theory to transfer learning. Here, there are data from a different (source) feature-label distribution, and one wishes to use this data together with whatever data are available from the (target) feature-label distribution of interest. The source and target are linked via a joint prior distribution, and an optimal Bayesian transfer learning classifier is derived for the posterior distribution in the target domain. Both Gaussian and negative binomial distributions are considered.

The final chapter addresses the fundamental problem of prior construction: How do we transform prior knowledge into a prior distribution? The first two sections address special cases: using data from discarded features, and using knowledge from a partially known physical system. The heart of the chapter is the development of a general method for transforming scientific knowledge into a prior distribution by performing an information-theoretic optimization over a class of potential priors with the optimization constrained by a set of conditional probability statements characterizing our scientific knowledge.

In a sense, this book is the last of a trilogy. *The Evolution of Scientific Knowledge: From Certainty to Uncertainty* (Dougherty, 2016) traces the epistemology of modern science from its deterministic beginnings in the Seventeenth century up through the inherent stochasticity of quantum theory in the first half of the Twentieth century, and then to the uncertainty in scientific models that has become commonplace in the latter part of the Twentieth century. This uncertainty leads to an inability to validate physical models, thereby limiting the scope of valid science. The last chapter of the book presents, from a philosophical perspective, the structure of operator design in the context of model uncertainty. *Optimal Signal Processing Under Uncertainty* (Dougherty, 2018) develops the mathematical theory articulated in that last chapter, applying it to filtering, control, classification, clustering, and experimental design. In this book, we extensively develop the classification theory summarized in that book.

Edward R. Dougherty
Lori A. Dalton
December 2019

Acknowledgments

The material in this book has been developed over a number of years and involves the contributions of numerous students and colleagues, whom we would like to acknowledge: Mohammadmahdi Rezaei Yousefi, Amin Zollanvari, Mohammad Shahrokh Esfahani, Shahin Boluki, Alireza Karbalayghareh, Siamak Zamani Dadaneh, Roozbeh Dehghannasiri, Xiaoning Qian, Byung-Jun Yoon, and Ulisses Braga-Neto. We also extend our appreciation to Dara Burrows for her careful editing of the book.

Chapter 1
Classification and Error Estimation

A classifier operates on a vector of features, which are random variables, and outputs a decision as to which class the feature vector belongs. We consider binary classification, meaning that the decision is between two classes. Typically, a classifier is designed and its error estimated from sample data. Two basic questions arise. First, given a set of features, how does one design a classifier from the sample data that provides good classification over the general population? Second, how does one estimate the error of a designed classifier from the data? This book examines both classifier design and error estimation when one is given prior knowledge regarding the population. The first chapter provides basic background knowledge.

1.1 Classifiers

Classification involves a *feature vector* $\mathbf{X} = [X_1, X_2, \ldots, X_D]^T$ composed of random variables (*features*) on a D-dimensional Euclidean space $\mathcal{X} = \mathbb{R}^D$, a binary random variable (*label*) Y, and a *classifier* $\psi : \mathbb{R}^D \rightarrow \{0, 1\}$ to predict $Y \in \{0, 1\}$. We assume a joint *feature-label distribution* $f_{\mathbf{X},Y}(\mathbf{x}, y)$ for the random vector–label pair $(\mathbf{X}, Y)$. The feature-label distribution characterizes the classification problem and may be a generalized function. $f_{\mathbf{X}|Y}(\mathbf{x}|0)$ and $f_{\mathbf{X}|Y}(\mathbf{x}|1)$, the distributions of $\mathbf{X}$ given $Y = 0$ and $Y = 1$, respectively, are known as the *class-conditional distributions*. The *classification error* is relative to the feature-label distribution and equals the probability of incorrect classification $\Pr(\psi(\mathbf{X}) \neq Y)$. The error also equals the expected (mean) absolute difference between the label and the classification $\mathrm{E}[|Y - \psi(\mathbf{X})|]$. An optimal classifier is one having minimal error among all classifiers $\psi : \mathbb{R}^D \rightarrow \{0, 1\}$. An optimal classifier ψ_{Bayes} is called a *Bayes classifier* and its error $\varepsilon_{\mathrm{Bayes}}$ is called the *Bayes error*. The Bayes error is intrinsic to the feature-label distribution; however, there may be more than one Bayes classifier.

The error of an arbitrary classifier ψ can be expressed as

$$\begin{aligned}\varepsilon &= \int_{\mathcal{X}} \Pr(\psi(\mathbf{X}) \neq Y | \mathbf{X} = \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\ &= \int_{\psi(\mathbf{x})=0} \eta(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} + \int_{\psi(\mathbf{x})=1} [1 - \eta(\mathbf{x})] f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x},\end{aligned} \tag{1.1}$$

where $f_{\mathbf{X}}(\mathbf{x})$ is the marginal density of $\mathbf{X}$, and $\eta(\mathbf{x}) = \Pr(Y = 1|\mathbf{X} = \mathbf{x})$ is the posterior probability that $Y = 1$. The posterior distribution of Y is proportional to the product of the prior distribution of Y and the class-conditional density for $\mathbf{x}$ via Bayes' theorem:

$$\Pr(Y = y|\mathbf{X} = \mathbf{x}) = \frac{\Pr(Y = y) f_{\mathbf{X}|Y}(\mathbf{x}|y)}{f_{\mathbf{X}}(\mathbf{x})}. \tag{1.2}$$

Hence, the error is decomposed as

$$\varepsilon = c\varepsilon_0 + (1 - c)\varepsilon_1, \tag{1.3}$$

where $c = \Pr(Y = 0)$ is the *a priori* probability of class 0,

$$\begin{aligned}\varepsilon_0 &= \Pr(\psi(\mathbf{X}) = 1 | Y = 0) \\ &= \int_{\psi(\mathbf{x})=1} f_{\mathbf{X}|Y}(\mathbf{x}|0) d\mathbf{x}\end{aligned} \tag{1.4}$$

is the probability of an element from class 0 being wrongly classified (the error contributed by class 0), and, similarly, $\varepsilon_1 = \Pr(\psi(\mathbf{X}) = 0| Y = 1)$.

Since $0 \leq \eta(\mathbf{x}) \leq 1$, the right-hand side of Eq. 1.1 is minimized by the classifier

$$\begin{aligned}\psi_{\text{Bayes}}(\mathbf{x}) &= \begin{cases} 0 & \text{if } \eta(\mathbf{x}) \leq 0.5, \\ 1 & \text{otherwise,} \end{cases} \\ &= \begin{cases} 0 & \text{if } c f_{\mathbf{X}|Y}(\mathbf{x}|0) \geq (1 - c) f_{\mathbf{X}|Y}(\mathbf{x}|1), \\ 1 & \text{otherwise.} \end{cases}\end{aligned} \tag{1.5}$$

Hence, the Bayes classifier $\psi_{\text{Bayes}}(\mathbf{x})$ is defined to be 0 or 1 according to whether Y is more likely to be 0 or 1 given $\mathbf{x}$ (ties may be broken arbitrarily). It follows from Eqs. 1.1 and 1.5 that the Bayes error is given by

$$\varepsilon_{\text{Bayes}} = \int_{\eta(\mathbf{x})\leq 0.5} \eta(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} + \int_{\eta(\mathbf{x})>0.5} [1 - \eta(\mathbf{x})] f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}. \tag{1.6}$$

In practice, the feature-label distribution is typically unknown, and a classifier must be designed from sample data. We assume a dataset consisting of n points, and $\mathrm{E}[\varepsilon_n] > 0.5 - \tau$ (Devroye, 1982). Moreover, even if a classifier is universally consistent, the rate at which $\mathrm{E}[\Delta_n] \to 0$ is critical to application. If we desire a classifier whose expected error is within some tolerance of the Bayes error, consistency is not sufficient. Rather, we would like a statement of the following form: for any $\tau > 0$, there exists $n(\tau)$ such that, for $n > n(\tau)$, $\mathrm{E}[\Delta_n] < \tau$ for any distribution of $(\mathbf{X}, Y)$. Unfortunately, even if a classification rule is universally consistent, the design error converges to 0 arbitrarily slowly relative to all possible distributions. To wit, if $\{a_n\}$ is a decreasing sequence such that $1/16 \geq a_1 \geq a_2 \geq \cdots > 0$, then for any sequence of designed classifiers ψ_n, there exists a distribution of $(\mathbf{X}, Y)$ such that $\varepsilon_{\text{Bayes}} = 0$ and $\mathrm{E}[\varepsilon_n] > a_n$ (Devroye, 1982).

More generally, consistency is of little consequence for small-sample classifier design, which is a key focus of the current text.

1.2 Constrained Classifiers

A common problem with small-sample design is that $\mathrm{E}[\Delta_n]$ tends to be large. A classification rule may yield a classifier that performs well on the sample data. However, if the small sample does not represent the distribution sufficiently, then the designed classifier will not perform well on the distribution. Constraining classifier design means restricting the functions from which a classifier can be chosen to a class $\mathcal{C}$. Constraining the classifier can reduce the expected design error, but at the cost of increasing the error of the best possible classifier. Since optimization in $\mathcal{C}$ is over a subclass of classifiers, the error $\varepsilon_{\mathcal{C}}$ of an optimal classifier $\psi_{\mathcal{C}} \in \mathcal{C}$ will typically exceed the Bayes error, unless a Bayes classifier happens to be in $\mathcal{C}$. We call $\Delta_{\mathcal{C}} = \varepsilon_{\mathcal{C}} - \varepsilon_{\text{Bayes}}$ the *cost of constraint*. A classification rule yields a classifier $\psi_{n,\mathcal{C}} \in \mathcal{C}$ with error $\varepsilon_{n,\mathcal{C}}$, where $\varepsilon_{n,\mathcal{C}} \geq \varepsilon_{\mathcal{C}} \geq \varepsilon_{\text{Bayes}}$. Design error for constrained classification is $\Delta_{n,\mathcal{C}} = \varepsilon_{n,\mathcal{C}} - \varepsilon_{\mathcal{C}}$. The expected error of the designed classifier from $\mathcal{C}$ can be decomposed as

$$\mathrm{E}[\varepsilon_{n,\mathcal{C}}] = \varepsilon_{\text{Bayes}} + \Delta_{\mathcal{C}} + \mathrm{E}[\Delta_{n,\mathcal{C}}]. \tag{1.13}$$

The constraint is beneficial if and only if the cost of constraint is less than the decrease in expected design cost. The dilemma is that strong constraint reduces $\mathrm{E}[\Delta_{n,\mathcal{C}}]$ at the cost of increasing $\varepsilon_{\mathcal{C}}$.

Historically, *discriminant functions* have played an important role in classification. Keeping in mind that the logarithm is a strictly increasing function, if we define the discriminant function by

$$d_y(\mathbf{x}) = \ln f_{\mathbf{X}|Y}(\mathbf{x}|y) + \ln \Pr(Y = y), \tag{1.14}$$

then the misclassification error is minimized by $\psi_{\text{Bayes}}(\mathbf{x}) = y$ if $d_y(\mathbf{x}) \geq d_j(\mathbf{x})$ for $j = 0$, 1, or equivalently, $\psi_{\text{Bayes}}(\mathbf{x}) = 0$ if and only if

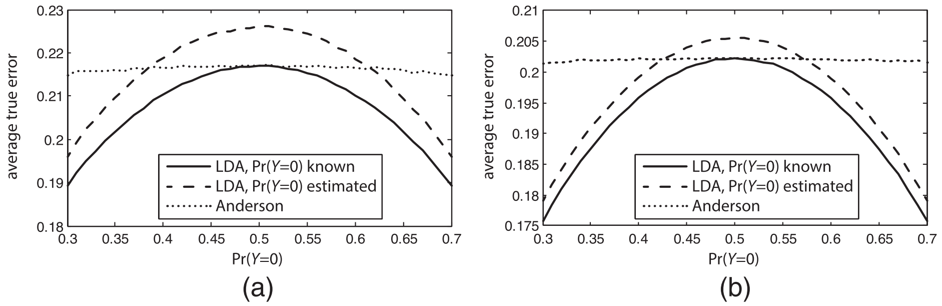

Figure 1.1 Average true error of LDA under Gaussian classes with respect to c: (a) $n = 20$; (b) $n = 50$. [Reprinted from (Esfahani and Dougherty, 2013).]

c estimated by n_0/n (default LDA), and c substituted by 0.5 (Anderson W statistic). The advantage of knowing c is evident, and this advantage is nonnegligible for small n. Estimating c is sufficiently difficult with small samples that in the case of $n = 20$, for $0.38 \leq c \leq 0.62$, the Anderson W statistic outperforms the default LDA.

The true error for any classifier ψ with linear discriminant $g(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$ on Gaussian distributions with known means $\boldsymbol{\mu}_y$ and covariances $\boldsymbol{\Sigma}_y$ ($\boldsymbol{\Sigma}_0 \neq \boldsymbol{\Sigma}_1$ allowable) is given in closed form by placing

$$\varepsilon_y = \Phi\left(\frac{(-1)^y g(\boldsymbol{\mu}_y)}{\sqrt{\mathbf{a}^T\boldsymbol{\Sigma}_y\mathbf{a}}}\right) \tag{1.32}$$

in Eq. 1.3 for $y = 0$, 1, where Φ is the standard normal *cumulative distribution function (CDF)* (Sitgreaves, 1961). Hence, the Bayes error when $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$ is given by

$$\varepsilon_{\text{Bayes}} = c\varepsilon_{\text{Bayes},0} + (1 - c)\varepsilon_{\text{Bayes},1}, \tag{1.33}$$

where

$$\varepsilon_{\text{Bayes},y} = \Phi\left(\frac{(-1)^y g_{\text{Bayes}}(\boldsymbol{\mu}_y)}{\sqrt{\mathbf{a}_{\text{Bayes}}^T\boldsymbol{\Sigma}\mathbf{a}_{\text{Bayes}}}}\right). \tag{1.34}$$

Generally speaking, the more complex a classifier class $\mathcal{C}$, the smaller the constraint cost and the greater the design cost. By this we mean that the more finely the functions in $\mathcal{C}$ partition the feature space $\mathbb{R}^D$, the better functions within it can approximate a Bayes classifier, and, concomitantly, the more they can overfit the data. This notion can be illustrated via a celebrated theorem that provides bounds for $\text{E}[\Delta_{n,\mathcal{C}}]$. It concerns the *empirical error* between the true and cross-validation estimated errors has been mathematically demonstrated in some basic models (Braga-Neto and Dougherty, 2005).

1.4 Random Versus Separate Sampling

Thus far, we have assumed *random sampling*, under which the dataset $\mathcal{S}_n$ is drawn independently from a fixed distribution of feature–label pairs $(\mathbf{X}, Y)$. In particular, this means that if a sample of size n is drawn for a binary classification problem, then the number of sample points in classes 0 and 1, n_0 and n_1, respectively, are binomial random variables such that $n_0 + n_1 = n$. An immediate consequence of the random-sampling assumption is that the prior probability $c = \Pr(Y = 0)$ can be consistently estimated by the sampling ratio $\hat{c} = n_0/n$, namely, $n_0/n \to c$ in probability.

While random sampling is almost invariably assumed (often tacitly) in classification theory, it is quite common in real-world situations for sampling not to be random. Specifically, with *separate sampling*, the class ratios n_0/n and n_1/n are chosen prior to sampling. Here, $\mathcal{S}_n = \mathcal{S}_{n_0} \cup \mathcal{S}_{n_1}$, where the sample points in $\mathcal{S}_{n_0}$ and $\mathcal{S}_{n_1}$ are selected randomly from class 0 and class 1, respectively, but, given n, the individual class counts n_0 and n_1 are not random. In this case, n_0/n is not a meaningful estimate of c. When c is not known, both QDA and LDA involve the estimate $\hat{c} = n_0/n$ by default. Hence, in the case of separate sampling when c is unknown, they are problematic. More generally, most classification rules make no explicit mention of c; however, their behavior depends on the sampling ratio, and their expected performances can be significantly degraded by separate sampling. In the special case when c is known and $n_0/n \approx c$ for separate sampling, the sampling is said to be *stratified*.

Example 1.1. Consider a model composed of multivariate Gaussian distributions with a block covariance structure (Esfahani and Dougherty, 2013). The model has several parameters that can generate various covariance matrices. For example, a 3-block covariance matrix with block size 5 has the structure

$$\boldsymbol{\Sigma}_y = \begin{bmatrix} \mathbf{B}_y & \mathbf{0}_{5\times 5} & \mathbf{0}_{5\times 5} \\ \mathbf{0}_{5\times 5} & \mathbf{B}_y & \mathbf{0}_{5\times 5} \\ \mathbf{0}_{5\times 5} & \mathbf{0}_{5\times 5} & \mathbf{B}_y \end{bmatrix}, \tag{1.45}$$

where $\mathbf{0}_{k\times l}$ is a matrix of size $k \times l$ with all elements being 0, and

$$\mathbf{B}_y = \sigma_y^2 \begin{bmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{bmatrix}, \tag{1.46}$$

(equivalently, the prior knowledge must be sufficiently great) that an acceptable bound can be achieved with an acceptable sample size.

1.5.2 Error RMS in the Gaussian model

To illustrate the error RMS problem, we first consider RMS for LDA in the one-dimensional heteroscedastic model (means μ_0 and μ_1, and variances σ_0^2 and σ_1^2) with separate sampling using fixed class sample sizes n_0 and n_1, and resubstitution error estimation. Here, LDA is given by $\psi_{\mathrm{LDA}}(\mathbf{x}) = 0$ if and only if $W(\mathbf{x}) < 0$, where $W(\mathbf{x})$ is the Anderson W statistic in Eq. 1.29; relative to our default definition, this variant of LDA assumes that $\hat{c} = 0.5$ and assigns the decision boundary to class 1 instead of class 0. The true error and resubstitution estimate for class y are denoted by ε_n^y and $\hat{\varepsilon}_{\mathrm{resub}}^y$, respectively. From here forward we will typically denote the class associated with an error or error estimate using superscripts to avoid cluttered notation. The MSE is

$$\mathrm{MSE}(\hat{\varepsilon}_n) = \mathrm{E}\left[\left|c\varepsilon_n^0 + (1-c)\varepsilon_n^1 - \left[\frac{n_0}{n}\hat{\varepsilon}_{\mathrm{resub}}^0 + \frac{n_1}{n}\hat{\varepsilon}_{\mathrm{resub}}^1\right]\right|^2\right]. \tag{1.50}$$

To evaluate the MSE we need expressions for the various second-order moments involved. These are derived in (Zollanvari et al., 2012) and are provided in (Braga-Neto and Dougherty, 2015). To illustrate the complexity of the problem, even in this very simple setting, we state the forms of the required second-order moments for the RMS. In all cases, the expressions of the form $\mathbf{Z} < 0$ and $\mathbf{Z} \geq 0$ mean that all components of $\mathbf{Z}$ are nonpositive and nonnegative, respectively.

The second-order moments for ε_n^y are

$$\mathrm{E}\left[(\varepsilon_n^0)^2\right] = \Pr(\mathbf{Z}^{\mathrm{I}} < 0) + \Pr(\mathbf{Z}^{\mathrm{I}} \geq 0), \tag{1.51}$$

$$\mathrm{E}\left[\varepsilon_n^0\varepsilon_n^1\right] = \Pr(\mathbf{Z}^{\mathrm{II}} < 0) + \Pr(\mathbf{Z}^{\mathrm{II}} \geq 0), \tag{1.52}$$

$$\mathrm{E}\left[(\varepsilon_n^1)^2\right] = \Pr(\mathbf{Z}^{\mathrm{III}} < 0) + \Pr(\mathbf{Z}^{\mathrm{III}} \geq 0), \tag{1.53}$$

where $\mathbf{Z}^j$, for $j = \mathrm{I}, \mathrm{II}, \mathrm{III}$, are 3-variate Gaussian random vectors with means

$$\boldsymbol{\mu}_{\mathbf{Z}^{\mathrm{I}}} = \boldsymbol{\mu}_{\mathbf{Z}^{\mathrm{II}}} = \boldsymbol{\mu}_{\mathbf{Z}^{\mathrm{III}}} = \begin{bmatrix} \frac{\mu_0-\mu_1}{2} \\ \mu_1 - \mu_0 \\ \frac{\mu_0-\mu_1}{2} \end{bmatrix}, \tag{1.54}$$

and covariance matrices

$$\boldsymbol{\Sigma}_{\mathbf{Z}^{\mathrm{I}}} = \begin{bmatrix} s+\sigma_0^2 & 2d & s \\ 2d & 4s & 2d \\ s & 2d & s+\sigma_0^2 \end{bmatrix}, \tag{1.55}$$

Chapter 2
Optimal Bayesian Error Estimation

Given that a distributional model is needed to achieve useful performance bounds for classifier error estimation when using the training data, a natural course of action is to define a prior distribution over the uncertainty class of feature-label distributions and then find an optimal *minimum-mean-square-error (MMSE)* error estimator relative to the uncertainty class (Dalton and Dougherty, 2011b).

2.1 The Bayesian MMSE Error Estimator

Consider finding a MMSE estimator (filter) of a nonnegative function $g(X,\ Y)$ of two random variables based on observing Y; that is, minimize $\mathrm{E}_{X,Y}[|g(X,\ Y) - h(Y)|^2]$ over all Borel measurable functions h. The optimal estimator,

$$\hat{g} = \arg\min_h \mathrm{E}_{X,Y}[|g(X,\ Y) - h(Y)|^2], \tag{2.1}$$

is given by the conditional expectation

$$\hat{g}(Y) = \mathrm{E}_X[g(X,\ Y)|\ Y]. \tag{2.2}$$

Moreover, $\hat{g}(Y)$ is an unbiased estimator over the distribution $f(x,\ y)$ of $(X,\ Y)$:

$$\mathrm{E}_Y[\hat{g}(Y)] = \mathrm{E}_{X,Y}[g(X,\ Y)]. \tag{2.3}$$

The fact that $\hat{g}(Y)$ is an unbiased MMSE estimator of $g(X,\ Y)$ over $f(x,\ y)$ does not tell us how well $\hat{g}(Y)$ estimates $g(\bar{x},\ Y)$ for some specific value $X = \bar{x}$. This has to do with the expected difference

Further, by Bayes' theorem,

$$\begin{aligned}\pi^*(\boldsymbol{\theta}_y) &= f(\boldsymbol{\theta}_y|\{\mathbf{x}_i^y\}_1^{n_y}) \\ &\propto \pi(\boldsymbol{\theta}_y)f(\{\mathbf{x}_i^y\}_1^{n_y}|\boldsymbol{\theta}_y) \\ &= \pi(\boldsymbol{\theta}_y)\prod_{i=1}^{n_y} f_{\boldsymbol{\theta}_y}(\mathbf{x}_i^y|y).\end{aligned} \tag{2.17}$$

We assume that sample points are independent throughout, with the exception of Sections 4.10 and 4.11. As is common in Bayesian analysis, we often characterize a posterior as being proportional to a prior times a likelihood; the normalization constant will still be accounted for throughout our analysis.

Although we call $\pi(\boldsymbol{\theta}_y)$ the "prior probabilities," they are not required to be valid density functions. The priors are *proper* if the integral of $\pi(\boldsymbol{\theta}_y)$ is finite, and they are *improper* if the integral of $\pi(\boldsymbol{\theta}_y)$ is infinite, i.e., if $\pi(\boldsymbol{\theta}_y)$ induces a σ-finite measure but not a finite probability measure. The flat prior over an infinite support is necessarily an improper prior. When improper priors are used, Bayes' theorem does not apply. However, as long as the product of the prior and likelihood function is integrable, we define the posterior to be their normalized product, e.g., we take Eq. 2.17 (following normalization) as the definition of the posterior in the mutually independent case.

For the class prior probabilities, under random sampling, we only need to consider the size of each class:

$$\pi^*(c) = f(c|n_0) \propto \pi(c)f(n_0|c) \propto \pi(c)c^{n_0}(1-c)^{n_1}, \tag{2.18}$$

where we have taken advantage of the fact that n_0 has a binomial(n, c) distribution given c. We consider three models for the prior distributions of the *a priori* class probabilities: beta, uniform, and known.

Suppose that the prior distribution for c follows a beta(α_0, α_1) distribution,

$$\pi(c) = \frac{c^{\alpha_0-1}(1-c)^{\alpha_1-1}}{B(\alpha_0, \alpha_1)}, \tag{2.19}$$

where

$$B(\alpha_0, \alpha_1) = \frac{\Gamma(\alpha_0)\Gamma(\alpha_1)}{\Gamma(\alpha_0+\alpha_1)} \tag{2.20}$$

is the beta function, and

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx \tag{2.21}$$

2.2 Evaluation of the Bayesian MMSE Error Estimator

Direct evaluation of the Bayesian MMSE error estimator is accomplished by deriving $\mathrm{E}_{\pi^*}[\varepsilon_n^y]$ for each class using Eq. 2.14, finding $\mathrm{E}_{\pi^*}[c]$ according to the prior model for c, and referring to Eq. 2.13 for the complete Bayesian MMSE error estimator. This can be tedious owing to the need to evaluate a challenging integral. Fortunately, the matter can be greatly simplified. Relative to the uncertainty class $\boldsymbol{\Theta}$ and the posterior distribution $\pi^*(\boldsymbol{\theta}_y)$ for $y = 0, 1$, we define the *effective class-conditional density* as

$$f_{\boldsymbol{\Theta}}(\mathbf{x}|y) = \int_{\boldsymbol{\Theta}_y} f_{\boldsymbol{\theta}_y}(\mathbf{x}|y)\pi^*(\boldsymbol{\theta}_y)d\boldsymbol{\theta}_y. \tag{2.26}$$

The next theorem shows that the effective class-conditional density, which is the average of the state-specific conditional densities relative to the posterior distribution, can be used to more easily obtain the Bayesian MMSE error estimator than by going through the route of direct evaluation.

Theorem 2.1 (Dalton and Dougherty, 2013a). *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 feature space. Then the Bayesian MMSE error estimator can be found by*

$$\begin{aligned}\hat{\varepsilon}_n(\mathcal{S}_n, \psi) &= \mathrm{E}_{\pi^*}[c]\int_{R_1} f_{\boldsymbol{\Theta}}(\mathbf{x}|0)d\mathbf{x} + (1 - \mathrm{E}_{\pi^*}[c])\int_{R_0} f_{\boldsymbol{\Theta}}(\mathbf{x}|1)d\mathbf{x} \\ &= \int_{\mathcal{X}} [\mathrm{E}_{\pi^*}[c] f_{\boldsymbol{\Theta}}(\mathbf{x}|0)\mathrm{I}_{\mathbf{x}\in R_1} + (1 - \mathrm{E}_{\pi^*}[c]) f_{\boldsymbol{\Theta}}(\mathbf{x}|1)\mathrm{I}_{\mathbf{x}\in R_0}]d\mathbf{x}.\end{aligned} \tag{2.27}$$

Proof. For a fixed distribution $\boldsymbol{\theta}_y$ and classifier ψ, the true error contributed by class $y \in \{0, 1\}$ may be written as

$$\varepsilon_n^y(\boldsymbol{\theta}_y, \psi) = \int_{R_{1-y}} f_{\boldsymbol{\theta}_y}(\mathbf{x}|y)d\mathbf{x}. \tag{2.28}$$

Averaging over the posterior yields

$$\begin{aligned}\mathrm{E}_{\pi^*}[\varepsilon_n^y(\boldsymbol{\theta}_y, \psi)] &= \int_{\boldsymbol{\Theta}_y} \varepsilon_n^y(\boldsymbol{\theta}_y, \psi)\pi^*(\boldsymbol{\theta}_y)d\boldsymbol{\theta}_y \\ &= \int_{\boldsymbol{\Theta}_y}\int_{R_{1-y}} f_{\boldsymbol{\theta}_y}(\mathbf{x}|y)d\mathbf{x}\pi^*(\boldsymbol{\theta}_y)d\boldsymbol{\theta}_y \\ &= \int_{R_{1-y}}\int_{\boldsymbol{\Theta}_y} f_{\boldsymbol{\theta}_y}(\mathbf{x}|y)\pi^*(\boldsymbol{\theta}_y)d\boldsymbol{\theta}_y d\mathbf{x} \\ &= \int_{R_{1-y}} f_{\boldsymbol{\Theta}}(\mathbf{x}|y)d\mathbf{x},\end{aligned} \tag{2.29}$$

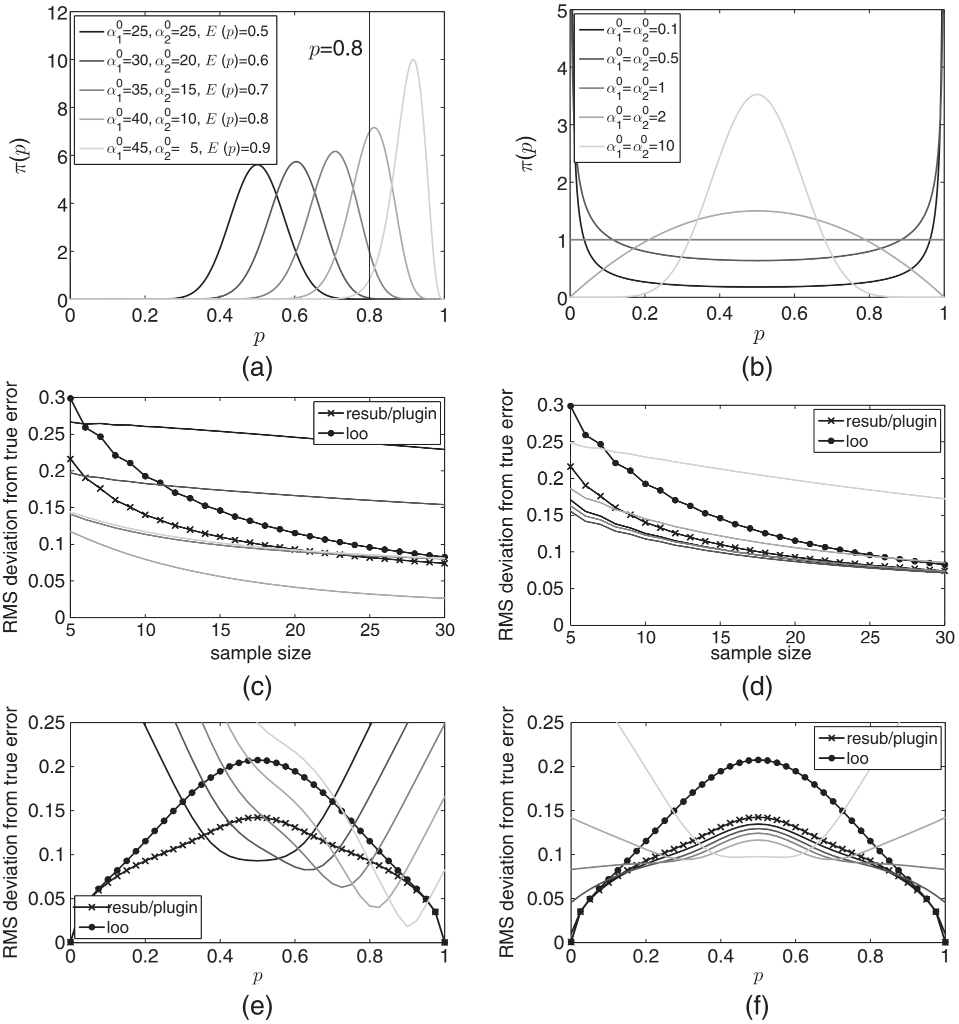

Figure 2.2 RMS deviation from true error for discrete classification ($b = 2$, $c = 0.5$): (a) using low variance priors; (b) using priors centered at $p = 0.5$; (c) versus sample size (low variance priors, $p = 0.8$); (d) versus sample size (centered priors, $p = 0.8$); (e) versus p (low variance priors, $n = 20$); (f) versus p (centered priors, $n = 20$). Lines without markers represent Bayesian MMSE error estimators with different beta priors, which are labeled and shown in the graph at the top of the corresponding column. [Reprinted from (Dalton and Dougherty, 2011b).]

The third row of Fig. 2.2 shows performance graphs with sample size $n = 20$, as a function of p. These illustrate how each prior performs as the true distributions vary. In all cases, performance is best in the ranges of p and q that are well represented in the prior distributions, but outside this range results can be poor. This is best seen in Fig. 2.2(e), where the RMS curves move to the right as the priors move right. High-information priors offer better performance if they are within the targeted range of parameters, but

Theorem 2.9 (Dalton and Dougherty, 2013a). *Assuming that $\nu^* > 0$, $\kappa^* > D - 1$, and $\mathbf{S}^*$ is symmetric positive definite, the effective class-conditional density for the general covariance model is a multivariate t-distribution with $k = \kappa^* - D + 1$ degrees of freedom, location vector $\mathbf{m}^*$, and scale matrix $[(\nu^* + 1)/((\kappa^* - D + 1)\nu^*)]\mathbf{S}^*$. That is,*

$$\begin{aligned} f_{\Theta}(\mathbf{x}|y) &= \frac{\Gamma(\frac{k+D}{2})}{\Gamma(\frac{k}{2})k^{\frac{D}{2}}\pi^{\frac{D}{2}}|\frac{\nu^*+1}{(\kappa^*-D+1)\nu^*}\mathbf{S}^*|^{\frac{1}{2}}} \\ &\quad \times \left[1 + \frac{1}{k}(\mathbf{x}-\mathbf{m}^*)^T\left(\frac{\nu^*+1}{(\kappa^*-D+1)\nu^*}\mathbf{S}^*\right)^{-1}(\mathbf{x}-\mathbf{m}^*)\right]^{-\frac{k+D}{2}}. \end{aligned} \tag{2.120}$$

This distribution is proper, the mean exists and is $\mathbf{m}^*$ as long as $\kappa^* > D$, and the covariance exists and is $[(\nu^* + 1)/((\kappa^* - D - 1)\nu^*)]\mathbf{S}^*$ as long as $\kappa^* > D + 1$.

Proof. By definition,

$$\begin{aligned} f_{\Theta}(\mathbf{x}|y) &= \int_{\mathcal{L}_y}\int_{\mathbb{R}^D} f_{\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y}(\mathbf{x})\pi^*(\boldsymbol{\mu}_y|\boldsymbol{\Sigma}_y)\pi^*(\boldsymbol{\Sigma}_y)d\boldsymbol{\mu}_y d\boldsymbol{\Sigma}_y \\ &= \int_{\mathcal{L}_y} f_{\mathbf{m}^*,\,\frac{\nu^*+1}{\nu^*}\boldsymbol{\Sigma}_y}(\mathbf{x})\pi^*(\boldsymbol{\Sigma}_y)d\boldsymbol{\Sigma}_y, \end{aligned} \tag{2.121}$$

where in the last line we have used Theorem 2.6 for the fixed covariance model. Continuing,

$$\begin{aligned} f_{\Theta}(\mathbf{x}|y) &= \int_{\mathcal{L}_y} \frac{(\nu^*)^{\frac{D}{2}}}{(\nu^*+1)^{\frac{D}{2}}(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}_y|^{\frac{1}{2}}} \\ &\qquad \times \exp\left(-\frac{\nu^*}{2(\nu^*+1)}(\mathbf{x}-\mathbf{m}^*)^T\boldsymbol{\Sigma}_y^{-1}(\mathbf{x}-\mathbf{m}^*)\right) \\ &\qquad \times \frac{|\mathbf{S}^*|^{\frac{\kappa^*}{2}}}{2^{\frac{\kappa^* D}{2}}\Gamma_D(\frac{\kappa^*}{2})}|\boldsymbol{\Sigma}_y|^{-\frac{\kappa^*+D+1}{2}}\mathrm{etr}\left(-\frac{1}{2}\mathbf{S}^*\boldsymbol{\Sigma}_y^{-1}\right)d\boldsymbol{\Sigma}_y \\ &= \int_{\mathcal{L}_y} \frac{(\nu^*)^{\frac{D}{2}}}{(\nu^*+1)^{\frac{D}{2}}(2\pi)^{\frac{D}{2}}}\cdot\frac{|\mathbf{S}^*|^{\frac{\kappa^*}{2}}}{2^{\frac{\kappa^* D}{2}}\Gamma_D(\frac{\kappa^*}{2})}|\boldsymbol{\Sigma}_y|^{-\frac{\kappa^*+D+2}{2}} \\ &\qquad \times \mathrm{etr}\left(-\frac{1}{2}\left[\mathbf{S}^* + \frac{\nu^*}{\nu^*+1}(\mathbf{x}-\mathbf{m}^*)(\mathbf{x}-\mathbf{m}^*)^T\right]\boldsymbol{\Sigma}_y^{-1}\right)d\boldsymbol{\Sigma}_y. \end{aligned} \tag{2.122}$$

The integrand is essentially an inverse-Wishart distribution and, therefore,

Chapter 3
Sample-Conditioned MSE of Error Estimation

There are two sources of randomness in the Bayesian model. The first is the sample, which randomizes the designed classifier and its true error. Most results on error estimator performance are averaged over random samples, which demonstrates performance relative to a fixed feature-label distribution. The second source of randomness is uncertainty in the underlying feature-label distribution. The Bayesian MMSE error estimator addresses the second source of randomness. This gives rise to a practical expected measure of performance given a fixed sample and classifier.

Up until this point, unless otherwise stated, we have assumed that c and $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ are independent to simplify the Bayesian MMSE error estimator. However, our interest is now in evaluating the MSE of an error estimator itself, which generally depends on the variances and correlations between the errors contributed by both classes. For the sake of simplicity, throughout this chapter we avoid the need to evaluate correlations by making the stronger assumption that c, $\boldsymbol{\theta}_0$, and $\boldsymbol{\theta}_1$ are all mutually independent. See Chapter 5 and (Dalton and Yousefi, 2015) for derivations in the general case.

3.1 Conditional MSE of Error Estimators

For a fixed sample $\mathcal{S}_n$, the sample-conditioned MSE of an arbitrary error estimator $\hat{\varepsilon}_\bullet$ is defined to be

$$\mathrm{MSE}(\hat{\varepsilon}_\bullet(\mathcal{S}_n, \psi)|\mathcal{S}_n) = \mathrm{E}_{\pi^*}[(\varepsilon_n(\boldsymbol{\theta}, \psi) - \hat{\varepsilon}_\bullet(\mathcal{S}_n, \psi))^2]. \tag{3.1}$$

This is precisely the objective function optimized by the Bayesian MMSE error estimator. Also define the conditional MSE for the Bayesian MMSE error estimate for each class:

$$\mathrm{MSE}(\hat{\varepsilon}_n^y(\mathcal{S}_n, \psi)|\mathcal{S}_n) = \mathrm{E}_{\pi^*}[(\varepsilon_n^y(\boldsymbol{\theta}_y, \psi) - \hat{\varepsilon}_n^y(\mathcal{S}_n, \psi))^2]. \tag{3.2}$$

the problem. In particular, the expressions for $\hat{\varepsilon}_n^0$, $\hat{\varepsilon}_n^1$, $\mathrm{E}_{\pi^*}[(\varepsilon_n^0(\boldsymbol{\theta}_0))^2]$, and $\mathrm{E}_{\pi^*}[(\varepsilon_n^1(\boldsymbol{\theta}_1))^2]$ can take on only a finite set of values, which is especially small for a small number of bins or sample points. In both parts of Fig. 3.1 (as well as in other unshown plots for different values of b and n), the density of the conditional RMS for the Bayesian MMSE error estimator is much tighter than that of leave-one-out. For example, in Fig. 3.1(b) the conditional RMS of the Bayesian MMSE error estimator tends to be very close to 0.05, whereas the leave-one-out error estimator has a long tail with substantial mass between 0.05 and 0.2. Furthermore, the conditional RMS for the Bayesian MMSE error estimator is concentrated on lower values of RMS, so much so that in all cases the unconditional RMS of the Bayesian MMSE error estimator is less than half that of the leave-one-out error estimator.

Without any kind of modeling assumptions, distribution-free bounds on the unconditional RMS are too loose to be useful. In fact, the bound from Eq. 1.48 is greater than 0.85 in both subplots of Fig. 3.1. On the other hand, a Bayesian framework facilitates exact expressions for the RMS conditioned on the sample for both the Bayesian MMSE error estimator and any other error estimation rule.

3.4 Gaussian Model

We next consider the Gaussian model. For linear classifiers, we have closed-form Bayesian MMSE error estimators for four models: fixed covariance, scaled identity covariance, diagonal covariance, and general covariance. However, note that independence between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ implies that the following results are not applicable in homoscedastic models. To avoid cluttered notation, in this section we denote hyperparameters without subscripts.

3.4.1 Effective joint class-conditional densities

Known Covariance

The following theorem provides a closed form for the effective joint density in the known covariance model.

Theorem 3.6 (Dalton and Yousefi, 2015). *If $\nu^* > 0$ and $\boldsymbol{\Sigma}_y$ is a fixed symmetric positive definite matrix, then*

$$f_{\boldsymbol{\Theta}}(\mathbf{x}, \mathbf{z}|y) \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m}^* \\ \mathbf{m}^* \end{bmatrix}, \begin{bmatrix} \frac{\nu^*+1}{\nu^*}\boldsymbol{\Sigma} & \frac{1}{\nu^*}\boldsymbol{\Sigma}_y \\ \frac{1}{\nu^*}\boldsymbol{\Sigma}_y & \frac{\nu^*+1}{\nu^*}\boldsymbol{\Sigma}_y \end{bmatrix} \right). \tag{3.29}$$

and

$$F_1\left(\frac{1}{2};\ \frac{1}{2},\ 1;\ \frac{3}{2};\ w,\ z\right) = \frac{1}{\sqrt{w-z}}\tan^{-1}\left(\sqrt{\frac{w-z}{1-w}}\right). \tag{3.139}$$

Further simplification gives the result in the statement of the lemma. ■

3.5 Average Performance in the Gaussian Model

In this section we examine performance in Gaussian models under proper priors with fixed sample size, illustrating that different samples condition RMS performance to different extents and that models using more informative priors have better RMS performance.

Consider an independent general covariance Gaussian model with known and fixed $c = 0.5$. Let $\nu_0 = 6D$, $\nu_1 = 3D$, $\mathbf{m}_0 = \mathbf{0}_D$, $\mathbf{m}_1 = -0.1719\cdot\mathbf{1}_D$, $\kappa_y = 3D$, and $\mathbf{S}_y = 0.03(\kappa_y - D - 1)\mathbf{I}_D$. This is a proper prior, where $\mathbf{m}_1$ has been calibrated to give an expected true error of 0.25 with $D = 1$.

Following the procedure in Fig. 2.5, in step 1, $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$, $\boldsymbol{\mu}_1$, and $\boldsymbol{\Sigma}_1$ are generated according to the specified priors. This is done by generating a random covariance according to the inverse-Wishart distribution $\pi(\boldsymbol{\Sigma}_y)$ using methods in (Johnson, 1987). Conditioned on the covariance, we generate a random mean from the Gaussian distribution $\pi(\boldsymbol{\mu}_y|\boldsymbol{\Sigma}_y) \sim \mathcal{N}(\mathbf{m}_y, \boldsymbol{\Sigma}_y/\nu_y)$, resulting in a normal-inverse-Wishart distributed mean and covariance pair. The parameters for class 0 are generated independently from those of class 1.

In step 2A we generate a random sample of size n, in step 2B the prior is updated to a posterior, and in step 2C we train an LDA classifier. In step 2D, the true error of the classifier is computed exactly, the training data are used to evaluate the 5-fold cross-validation error estimator, a Bayesian MMSE error estimator is found exactly using the posterior, and the theoretical sample-conditioned RMS for the Bayesian MMSE error estimator is computed exactly. The sampling procedure is repeated $t = 1000$ times for each fixed feature-label distribution, with $T = 10{,}000$ feature-label distributions, for a total of 10,000,000 samples.

Table 3.1 shows the accuracy of the analytical formulas for conditional RMS using $n = 60$ with different feature sizes ($D = 1$, 2, and 5). There

Table 3.1 Average true error, semi-analytical RMS, and absolute difference between the semi-analytical and empirical RMS for the Bayesian MMSE error estimator.

Simulation settings	Average true error	Semi-analytical RMS	Absolute difference between RMSs
$n = 60,\ D = 1$	0.2474	0.0377	5.208×10^{-6}
$n = 60,\ D = 2$	0.1999	0.0358	3.110×10^{-5}
$n = 60,\ D = 5$	0.1156	0.0262	8.971×10^{-5}

Chapter 5

Optimal Bayesian Risk-based Multi-class Classification

In this chapter we consider classification under multiple classes and allow for different types of error to be associated with different levels of risk or loss. A few classical classification algorithms naturally permit multiple classes and arbitrary loss functions; for example, a plug-in rule takes the functional form for an optimal Bayes decision rule under a given modeling assumption and substitutes sample estimates of model parameters in place of the true parameters. This can be done with LDA and QDA for multiple classes with arbitrary loss functions, which essentially assume that the underlying class-conditional densities are Gaussian with equal or unequal covariances, respectively. Most training-data error estimation methods, for instance, cross-validation, can also be generalized to handle multiple classes and arbitrary loss functions. However, it is expected that the same difficulties encountered under binary classes with simple zero-one loss functions (where the expected risk reduces to the probability of misclassification) will carry over to the more general setting, as they have in ROC curve estimation (Hanczar et al., 2010).

Support vector machines are inherently binary but can be adapted to incorporate penalties that influence risk by implementing slack terms or applying a shrinkage or robustifying objective function (Xu et al., 2009a,b). It is also common to construct multi-class classifiers from binary classifiers using the popular *one-versus-all* or *all-versus-all* strategies (Bishop, 2006). The former method builds several binary classifiers by discriminating one class, in turn, against all others, and at a given test point reports the class corresponding to the highest classification score. The latter discriminates between each combination of pairs of classes and reports a majority vote. However, it is unclear how one may assess the precise effect of these adaptations on the expected risk.

Here we generalize the Bayesian MMSE error estimator, sample-conditioned MSE, and OBC to treat multiple classes with arbitrary loss functions. We will present the analogous concepts of the Bayesian risk estimator, sample-conditioned MSE for risk estimators, and optimal Bayesian risk classifier. We will show that

5.7 Evaluation of Posterior Mixed Moments: Gaussian Models

We now consider Gaussian models with mean vectors $\boldsymbol{\mu}_y$ and covariance matrices $\boldsymbol{\Sigma}_y$. All posteriors on parameters, effective densities, and effective joint densities found in this section apply under arbitrary multi-class classifiers. We also find analytic forms for posterior mixed moments, the BRE, and the conditional MSE under binary linear classifiers ψ of the form

$$\psi(\mathbf{x}) = \begin{cases} 0 & \text{if } g(\mathbf{x}) \leq 0, \\ 1 & \text{otherwise,} \end{cases} \tag{5.53}$$

where $g(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$ for some vector $\mathbf{a}$ and scalar b. Under arbitrary classifiers, the BRE and conditional MSE may be approximated using techniques described in Section 5.5.

5.7.1 Known covariance

Assume that $\boldsymbol{\Sigma}_y$ is known and is a valid invertible covariance matrix. Then $\mathbf{T}_y = \boldsymbol{\mu}_y$ with parameter space $\boldsymbol{\Theta}_y = \mathbb{R}^D$. We assume that the $\boldsymbol{\mu}_y$ are mutually independent and use the prior in Eq. 2.58:

$$\pi(\boldsymbol{\mu}_y) \propto |\boldsymbol{\Sigma}_y|^{-\frac{1}{2}} \exp\left(-\frac{\nu_y}{2}(\boldsymbol{\mu}_y - \mathbf{m}_y)^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{\mu}_y - \mathbf{m}_y)\right), \tag{5.54}$$

where $\nu_y \in \mathbb{R}$ and $\mathbf{m}_y \in \mathbb{R}^D$. The posterior is of the same form as the prior, with updated hyperparameters ν_y^* and $\mathbf{m}_y^*$ given in Eqs. 2.65 and 2.66, respectively. We require that $\nu_y^* > 0$ for a proper posterior. The effective density is given in Eq. 2.96 of Theorem 2.6.

To find the BRE we require $\hat{\varepsilon}_n^{i,y}(\psi, \mathcal{S}_n)$. Under linear classifiers, this is essentially given by Theorem 2.10, except that the exponent of -1 is altered to obtain

$$\hat{\varepsilon}_n^{i,y}(\psi, \mathcal{S}_n) = \Phi\left(\frac{(-1)^{i+1}g(\mathbf{m}_y^*)}{\sqrt{\mathbf{a}^T\boldsymbol{\Sigma}_y\mathbf{a}}}\sqrt{\frac{\nu_y^*}{\nu_y^*+1}}\right). \tag{5.55}$$

To find the MSE under linear classification, note that $f_{\boldsymbol{\Theta}}(\mathbf{w}|\mathbf{x}, y, z)$ is of the same form as $f_{\boldsymbol{\Theta}}(\mathbf{x}|y)$ with posterior hyperparameters updated with $(\mathbf{x}, y)$ as a new sample point. Hence, for $y = z$,

$$f_{\boldsymbol{\Theta}}(\mathbf{w}|\mathbf{x}, y, y) \sim \mathcal{N}\left(\mathbf{m}_y^* + \frac{\mathbf{x} - \mathbf{m}_y^*}{\nu_y^*+1}, \frac{\nu_y^*+2}{\nu_y^*+1}\boldsymbol{\Sigma}_y\right), \tag{5.56}$$

Chapter 7
Construction of Prior Distributions

Up to this point we have ignored the issue of prior construction, assuming that the characterization of uncertainty is known. For optimal Bayesian classification, the problem consists of transforming scientific knowledge into a probability distribution governing uncertainty in the feature-label distribution. Regarding prior construction in general, in 1968, E. T. Jaynes remarked, "Bayesian methods, for all their advantages, will not be entirely satisfactory until we face the problem of finding the prior probability squarely" (Jaynes, 1968). Twelve years later, he added, "There must exist a general formal theory of determination of priors by logical analysis of prior information—and that to develop it is today the top priority research problem of Bayesian theory" (Jaynes, 1980).

Historically, prior construction has tended to utilize general methodologies not targeting any specific type of prior information and has usually been treated independently (even subjectively) of real available prior knowledge and sample data. Subsequent to the introduction of the Jeffreys rule prior (Jeffreys, 1946), objective-based methods were proposed, two early ones being (Kashyap, 1971) and (Bernardo, 1979). There appeared a series of information-theoretic and statistical approaches: non-informative priors for integers (Rissanen, 1983), entropic priors (Rodriguez, 1991), *maximal data information priors (MDIP)* (Zellner, 1995), reference (non-informative) priors obtained through maximization of the missing information (Berger and Bernardo, 1992), and least-informative priors (Spall and Hill, 1990) [see also (Bernardo, 1979; Kass and Wasserman, 1996; Berger et al., 2012)]. The principle of maximum entropy can be seen as a method of constructing least-informative priors (Jaynes, 1957, 1968). Except in the Jeffreys rule prior, almost all of the methods are based on optimization: maximizing or minimizing an objective function, usually an information theoretic one. The least-informative prior in (Spall and Hill, 1990) is found among a restricted set of distributions, where the feasible region is a set of convex combinations of