Copyright 2022. All Rights Reserved. These notes cannot be copied or distributed in print or electronically without the written consent of Cambridge University Press and the author. The notes are distributed to students attending the course EE566: Adaptation and Learning taught by the author at EPFL during Spring 2022. The notes originate from the text: A. H. Sayed, *Inference and Learning from Data*, Vol. II: *Inference*, Cambridge University Press, 2022.

28 Bayesian Inference

The mean-square-error (MSE) criterion (27.17) is one notable example of the Bayesian approach to statistical inference. In the Bayesian approach, *both* the unknown quantity, $\boldsymbol{x}$, and the observation, $\boldsymbol{y}$, are treated as random variables and an estimator $\widehat{\boldsymbol{x}}$ for $\boldsymbol{x}$ is sought by minimizing the expected value of some other *loss* function denoted by $Q(\boldsymbol{x}, \widehat{\boldsymbol{x}})$. In the previous chapter, we focused exclusively on the quadratic loss $Q(\boldsymbol{x}, \widehat{\boldsymbol{x}}) = (\boldsymbol{x} - \widehat{\boldsymbol{x}})^2$ for scalar $\boldsymbol{x}$. In this chapter, we consider more general loss functions, which will lead to other types of inference solutions such as the mean-absolute error (MAE) and the maximum a-posteriori (MAP) estimators. We will also derive the famed Bayes classifier as a special case when the realizations for $\boldsymbol{x}$ are limited to the discrete values $\boldsymbol{x} \in \{\pm 1\}$.

28.1 BAYESIAN FORMULATION

Consider scalar random variables $\{\boldsymbol{x}, \boldsymbol{y}\}$, where $\boldsymbol{y}$ is observable and the objective is to infer the value of $\boldsymbol{x}$. The estimator for $\boldsymbol{x}$ is denoted by $\widehat{\boldsymbol{x}}$ and is defined as some function of $\boldsymbol{y}$, denoted by $c(\boldsymbol{y})$, to be determined by minimizing an average loss over the joint distribution of $\{\boldsymbol{x}, \boldsymbol{y}\}$. The purpose of the loss function is to measure the discrepancy between $\boldsymbol{x}$ and its estimator. The inference problem is stated as:

$$\boxed{\widehat{\boldsymbol{x}}_Q \triangleq \underset{\widehat{\boldsymbol{x}}=c(\boldsymbol{y})}{\operatorname{argmin}} \mathbb{E}\, Q(\boldsymbol{x}, \widehat{\boldsymbol{x}})} \tag{28.1}$$

where the loss $Q(\cdot, \cdot)$ is non-negative, and the expectation is over the joint pdf $f_{\boldsymbol{x},\boldsymbol{y}}(x, y)$. Similar to what we did in the last chapter, we will continue to employ continuous-time distributions in our presentation with the understanding that the arguments can be easily adjusted for discrete distributions.

Observe that we are attaching a subscript Q to $\widehat{\boldsymbol{x}}_Q$ to highlight its dependence on the choice of loss functions. The cost that appears in (28.1) in the form of an expected loss is also referred to as the *risk* and is denoted by:

$$R(c) \triangleq \mathbb{E}\, Q(\boldsymbol{x}, \widehat{\boldsymbol{x}}), \qquad (\textbf{risk function}) \tag{28.2}$$

Note that the risk depends on $c(\boldsymbol{y})$. Different choices for $c(\boldsymbol{y})$ will generally have

different risk values and the objective is to choose an optimal mapping, denoted by $c^o(\boldsymbol{y})$, with the smallest risk value possible:

$$R(c^o) = \min_{\widehat{\boldsymbol{x}}=c(\boldsymbol{y})} \mathbb{E}\, Q(\boldsymbol{x}, \widehat{\boldsymbol{x}}) \tag{28.3}$$

For later use, it is useful to note that formulation (28.1) admits an equivalent characterization. Using the conditional mean property (27.24), we rewrite the mean loss in the following form by conditioning on the observation $\boldsymbol{y}$:

$$\mathbb{E}\; Q(\boldsymbol{x}, \widehat{\boldsymbol{x}}) \;=\; \mathbb{E}_{\boldsymbol{y}} \left\{ \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}} \Big(Q(\boldsymbol{x}, \widehat{\boldsymbol{x}}|\boldsymbol{y}) \Big) \right\} \tag{28.4}$$

Now, since the loss function assumes nonnegative values, the minimization in (28.1) can be attained by solving instead:

$$\boxed{\widehat{x}_Q(y) \;\triangleq\; \underset{\widehat{x}=c(y)}{\operatorname{argmin}} \left\{ \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}} \Big(Q(\boldsymbol{x}, \widehat{\boldsymbol{x}}|\boldsymbol{y}=y) \Big) \right\}} \tag{28.5}$$

where the expectation of the loss function is now evaluated relative to the conditional pdf, $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$. This conditional pdf is known as the *predictive* distribution because it enables us to "predict" values for $\boldsymbol{x}$ for each individual observation for $\boldsymbol{y}$. The predictor $\widehat{x}_Q$ in (28.5) is a function of y and that is why we are denoting it more explicitly by writing $\widehat{x}_Q(y)$, with an argument y. Using relation (28.4), we then find that the minimal risk value admits the representation:

$$R(c^o) = \mathbb{E}_{\boldsymbol{y}} \Big\{ \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}} \Big(Q(\boldsymbol{x}, \widehat{\boldsymbol{x}}_Q(\boldsymbol{y})) \,|\, \boldsymbol{y}=y \Big) \Big\} \tag{28.6}$$

Formulations (28.1) and (28.5) are also valid when either $\boldsymbol{x}$ or $\boldsymbol{y}$ (or both) are vector-valued. We continue with the scalar case for illustration purposes. Two special cases of Bayesian estimators are evident.

Mean-square-error (MSE) inference

In the mean-square-error case, the loss function is quadratic and chosen as

$$Q(\boldsymbol{x}, \widehat{\boldsymbol{x}}) = (\boldsymbol{x} - \widehat{\boldsymbol{x}})^2 \tag{28.7}$$

In this case, we can solve problem (28.1) explicitly and showed in the last chapter that the estimator is the mean of the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{y}$, i.e.,

$$\widehat{\boldsymbol{x}}_{\text{MSE}} \;=\; \mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y}) \tag{28.8}$$

Note that we are attaching a subscript MSE to distinguish this estimator from other estimators discussed below.

Mean-absolute error (MAE) inference

In this case, the loss function is the absolute error, namely,

$$Q(\boldsymbol{x}, \widehat{\boldsymbol{x}}) \;=\; |\boldsymbol{x} - \widehat{\boldsymbol{x}}| \tag{28.9}$$

We showed in Prob. 27.13 that the corresponding estimator $\widehat{\boldsymbol{x}}$ is given by the *median* (rather than the mean) of the conditional distribution, $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$. That is, the value of $\widehat{x}_{\rm MAE}$ is the point that enforces the equality:

$$\int_{-\infty}^{\widehat{x}_{\rm MAE}} f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)dx \;=\; \int_{\widehat{x}_{\rm MAE}}^{\infty} f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)dx \;=\; 1/2 \tag{28.10}$$

28.2 MAXIMUM A-POSTERIORI INFERENCE

Another popular inference solution is the maximum a-posteriori (MAP) estimator. While the MSE estimator, $\widehat{\boldsymbol{x}}_{\rm MSE}$, selects the value x that corresponds to the mean of the conditional pdf, $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$, the MAP estimator, denoted by $\widehat{\boldsymbol{x}}_{\rm MAP}$, selects the location x that corresponds to the peak of the same pdf:

$$\widehat{x}_{\rm MAP} \;=\; \underset{x\in\mathcal{X}}{\rm argmax}\; f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) \tag{28.11}$$

where the maximization is over the domain of $x \in \mathcal{X}$. MAP estimators need not be unique because $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$ may be a multi-modal distribution.

MAP estimators can be viewed as a limiting case of Bayesian inference if the loss function is set to the $0/1-$loss defined as follows — see Prob. 28.1:

$$Q(\boldsymbol{x}, \widehat{\boldsymbol{x}}) \;\triangleq\; \begin{cases} 1, & \boldsymbol{x} \neq \widehat{\boldsymbol{x}} \\ 0, & \text{otherwise} \end{cases} \tag{28.12}$$

We illustrate this fact by considering the case in which $\boldsymbol{x}$ has a discrete support set. Thus, given an observation $\boldsymbol{y} = y$, and considering the $0/1-$loss (28.12), we have:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\, Q(\boldsymbol{x}, \widehat{\boldsymbol{x}}|\boldsymbol{y} = y) &\overset{(a)}{=} \sum_{x\in\mathcal{X}} \mathbb{P}(\boldsymbol{x} \neq \widehat{\boldsymbol{x}}|\boldsymbol{y} = y) Q(x, \widehat{x}) \\ &\overset{(28.12)}{=} \sum_{x\neq\widehat{x}} \mathbb{P}(\boldsymbol{x} \neq \widehat{\boldsymbol{x}}|\boldsymbol{y} = y) \\ &= 1 - \mathbb{P}(\boldsymbol{x} = \widehat{\boldsymbol{x}}|\boldsymbol{y} = y) \end{aligned} \tag{28.13}$$

where the expectation on the left-hand side of (a) is relative to the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{y} = y$. It follows that

$$\underset{\widehat{x}}{\rm argmin}\; \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\, Q(\boldsymbol{x}, \widehat{\boldsymbol{x}}|\boldsymbol{y} = y) \;=\; \underset{\widehat{x}}{\rm argmax}\; \mathbb{P}(\boldsymbol{x} = \widehat{x}|\boldsymbol{y} = y) \tag{28.14}$$

In other words, and in view of (28.5), the mean of the loss function (28.12) is

minimized when $\widehat{x}$ is selected as the location that maximizes the conditional distribution, $\mathbb{P}(\boldsymbol{x} = x|\boldsymbol{y} = y)$.

For jointly Gaussian-distributed random variables $\{\boldsymbol{x}, \boldsymbol{y}\}$, the MSE and MAP estimators for $\boldsymbol{x}$ will agree with each other. This is because the conditional pdf, $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$, will be Gaussian and the locations of its mean and peak will coincide. This conclusion, however, is not generally true for other distributions.

Example 28.1 (**MSE and MAP estimators**) Assume the conditional pdf of a scalar random variable, $\boldsymbol{x}$, given observations of another scalar random variable, $\boldsymbol{y} > 0$, follows a Rayleigh distribution of the form (3.26), namely,

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) \;=\; \frac{x}{y^2}\, e^{-x^2/2y^2}, \quad x \geq 0 \tag{28.15}$$

Then, we know from (3.27) that the mean and variance of this distribution, denoted by $\mu_{x|y}$ and $\sigma^2_{x|y}$, respectively, are given by

$$\mu_{x|y} = y\sqrt{\frac{\pi}{2}}, \qquad \sigma^2_{x|y} = \left(2 - \frac{\pi}{2}\right) y^2 \tag{28.16}$$

Moreover, the peak location of the Rayleigh distribution (its mode location) and its median are given by

$$\text{mode} \;=\; y, \qquad \text{median} \;=\; y\sqrt{2\ln 2} \tag{28.17}$$

It follows from these expressions that the MSE, MAE, and MAP estimators for $\boldsymbol{x}$ given $\boldsymbol{y}$ are given by

$$\widehat{\boldsymbol{x}}_{\text{MSE}} = \boldsymbol{y}\sqrt{\frac{\pi}{2}}, \qquad \widehat{\boldsymbol{x}}_{\text{MAE}} = \boldsymbol{y}\sqrt{2\ln 2}, \qquad \widehat{\boldsymbol{x}}_{\text{MAP}} = \boldsymbol{y} \tag{28.18}$$

Example 28.2 (**Election poll**) Two candidates $\mathbb{A}$ and $\mathbb{B}$ are running for office in a local district election. The probability of success for candidate $\mathbb{A}$ is p. We survey a fraction of the voters in the district, say, a number of N potential voters, and ask them whether they will be voting for one candidate or the other. We would like to use the result of the survey to estimate p, i.e., the likelihood of success for candidate $\mathbb{A}$.

Let $\boldsymbol{y}$ denote a binomial variable with parameters N and p. The probability of observing y successes in N trials (i.e., the probability of obtaining y positive answers in favor of candidate $\mathbb{A}$ out of N) is given by the expression:

$$\mathbb{P}(\boldsymbol{y} = y) = \binom{N}{y} p^y (1-p)^{N-y}, \quad y = 0, 1, \ldots, N \tag{28.19}$$

The value of the parameter p can be estimated in a number of ways, for example, by using a mean-square-error formulation (as described in Prob. 28.12), or a maximum-likelihood formulation (as discussed in future Prob. 31.8), or a maximum a-posteriori (MAP) formulation. In this example, we focus on the MAP approach.

In Bayesian inference, we treat the quantities we wish to estimate as random variables. For this reason, we will need to model $\boldsymbol{p}$ as a random variable and then determine an expression for the conditional pdf, $f_{\boldsymbol{p}|\boldsymbol{y}}(p|y)$. Once this pdf is computed, its peak location will provide the desired MAP estimate, $\widehat{p}_{\text{MAP}}$.

Treating $\boldsymbol{p}$ as random requires that we specify its distribution, $f_{\boldsymbol{p}}(p)$, also called the

prior. Since the value of $\boldsymbol{p}$ is confined to the interval $[0,1]$, we can select the prior from the family of *Beta distributions*. This family is useful in modeling random variables that are confined to *finite* intervals. The Beta distribution is defined by two positive *shape parameters* (a,b) as follows:

$$f_{\boldsymbol{p}}(p;a,b) = \begin{cases} \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, p^{a-1}(1-p)^{b-1}, & 0 \le p \le 1 \\ 0, & \text{otherwise} \end{cases} \tag{28.20}$$

where $\Gamma(x)$ denotes the Gamma function defined earlier in Prob. 4.3. Different choices for (a,b) result in different behavior for the distribution $f_{\boldsymbol{p}}(p)$. For example, the uniform distribution over the interval $[0,1]$ corresponds to the choice $a=b=1$. In this case, the variable $\boldsymbol{p}$ is equally likely to assume any value within the interval. Other values for a and b will give more likelihood to smaller or larger values in the interval. The top row in Fig. 28.1 plots some typical curves for the Beta distribution.

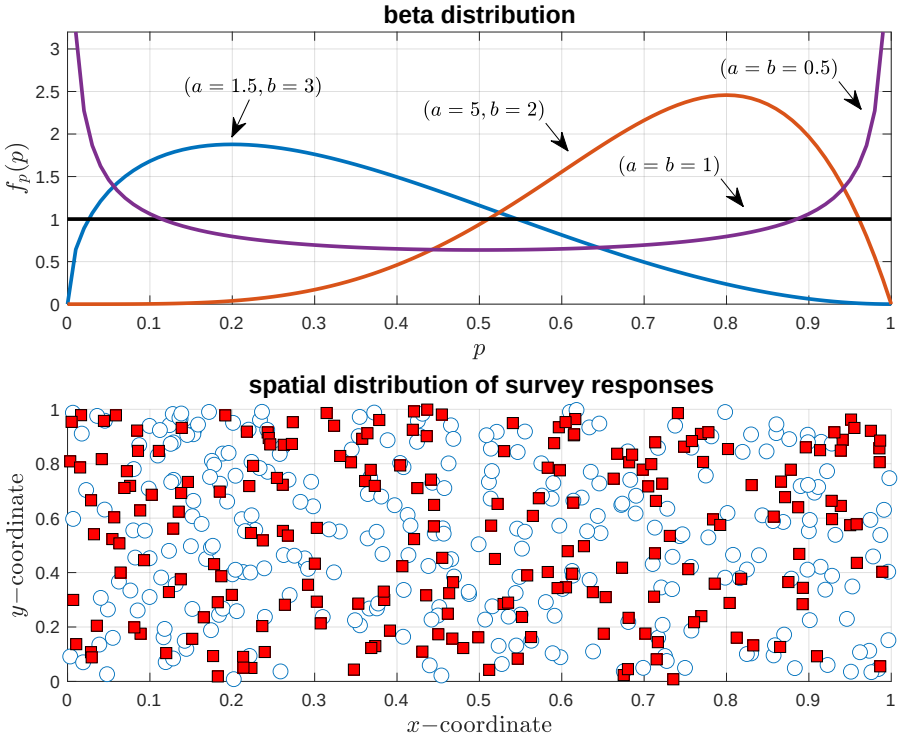

Figure 28.1 (*Top*) Plots of several Beta distributions for different values of the shape parameters (a,b). Observe how $a=b=1$ results in the uniform distribution, while other values for (a,b) give more likelihood to smaller or larger values within the interval $[0,1]$. (*Bottom*) Results of polling $N=500$ likely voters. The colors refer to votes for candidates $\mathbb{A}$ or $\mathbb{B}$.

The mean and variance of the Beta distribution (28.20) are known to be:

$$\bar{p} = \frac{a}{a+b}, \qquad \sigma_p^2 = \frac{ab}{(a+b)^2(a+b+1)} \tag{28.21}$$

When $a > 0$ and $b > 0$, the mode of the distribution is also known to occur at

$$\text{mode} = \frac{a-1}{a+b-2} \tag{28.22}$$

Using these facts, we derive an expression for the conditional pdf, $f_{\boldsymbol{p}|\boldsymbol{y}}(p|y)$, in Prob. 28.2 and deduce there that its peak occurs at location:

$$\widehat{p}_{\text{MAP}} \;=\; \frac{y+a-1}{N+a+b-2} \tag{28.23}$$

The bottom plot in Figure 28.1 shows the polling results from surveying $N = 500$ potential voters in the district. The simulation assumes a Beta distribution with parameters $a = 3$ and $b = 2$. The actual success probability was generated randomly according to this distribution and took the value $p = 0.5565$. Out of the $N = 500$ surveys, there were $y = 287$ votes in favor of candidate $\mathbb{A}$. Substituting into (28.23) we find that

$$\widehat{p}_{\text{MAP}} \;=\; \frac{287+3-1}{500+3+2-2} \approx 0.5746 \tag{28.24}$$

Note that we could have also estimated p by simply dividing y by N; this computation is a common solution and we will encounter it later in Prob. 31.8 where we will show that it amounts to the maximum-likelihood estimate for p denoted by:

$$\widehat{p}_{\text{ML}} \;=\; \frac{287}{500} = 0.5740 \tag{28.25}$$

This latter solution method, however, treats p as an unknown constant and not as a a random variable.

28.3 BAYES CLASSIFIER

One useful application of the MAP formulation (28.11) arises in the context of classification problems, which we will study in great detail in later chapters. In these problems, the unknown variable $\boldsymbol{x}$ is *discrete* and assumes a finite number of levels.

28.3.1 Binary Classification

We motivate classification problems by considering first the case in which $\boldsymbol{x}$ is a discrete *binary* random variable assuming one of two possible values, say, $\boldsymbol{x} \in \{\pm 1\}$. Given some possibly *vector*-valued observation $\boldsymbol{y} \in \mathbb{R}^M$ that is dependent on $\boldsymbol{x}$, we would like to infer $\boldsymbol{x}$ by determining a mapping, now called a *classifier*, $c(\boldsymbol{y})$, that maps $\boldsymbol{y}$ into one of the two *discrete* values:

$$c(\boldsymbol{y}) : \mathbb{R}^M \;\longrightarrow\; \{\pm 1\} \tag{28.26}$$

We refer to $\boldsymbol{x}$ as the *class* variable or the *label* corresponding to $\boldsymbol{y}$. The intention is to employ this mapping to deduce from the observation $\boldsymbol{y}$ whether it belongs to class $+1$ or -1. We can attain this objective by seeking the optimal estimator,

denoted by $\widehat{\boldsymbol{x}}_{\text{bayes}}$, that minimizes the probability of erroneous decisions (or misclassifications), i.e., that solves:

$$\boxed{\widehat{\boldsymbol{x}}_{\text{bayes}} \;=\; \underset{\widehat{\boldsymbol{x}}=c(\boldsymbol{y})}{\text{argmin}}\Big\{\mathbb{P}\left(c(\boldsymbol{y})\neq \boldsymbol{x}\right)\Big\}} \tag{28.27}$$

We verify below that the following classifier, known as *Bayes classifier*, solves (28.27):

$$\boxed{\widehat{x}_{\text{bayes}} \;=\; \begin{cases} +1, & \text{when } \mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=y)\geq 1/2 \\ -1, & \text{otherwise} \end{cases}} \tag{28.28}$$

which can also be written in a single equation as

$$\widehat{x}_{\text{bayes}} = 2\,\mathbb{I}\Big[\mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=y)\geq 1/2\Big]-1 \tag{28.29}$$

in terms of the indicator function:

$$\mathbb{I}[a] = \begin{cases} 1, & \text{if statement } a \text{ is true} \\ 0, & \text{otherwise} \end{cases} \tag{28.30}$$

Expression (28.28) indicates that the classifier decides in favor of $+1$ when the conditional probability of the event $\boldsymbol{x}=+1$ is at least $1/2$. In other words, the classifier $\widehat{\boldsymbol{x}}_{\text{bayes}}$ selects the value for x that maximizes the conditional probability of observing x given y, which means that $\widehat{\boldsymbol{x}}_{\text{bayes}}$ coincides with the MAP estimator:

$$\widehat{x}_{\text{bayes}} \;=\; \widehat{x}_{\text{MAP}} \;=\; \underset{x\in\{\pm 1\}}{\text{argmax}}\;\; \mathbb{P}(\boldsymbol{x}=x|\boldsymbol{y}=y) \tag{28.31}$$

Proof of (28.28): First, note that problem (28.27) is equivalent to solving

$$\widehat{\boldsymbol{x}}_{\text{bayes}} \;\overset{\Delta}{=}\; \underset{\widehat{\boldsymbol{x}}=c(\boldsymbol{y})}{\text{argmax}}\;\; \mathbb{P}\left(c(\boldsymbol{y})=\boldsymbol{x}\right) \tag{28.32}$$

where

$$\begin{aligned} \mathbb{P}\left(c(\boldsymbol{y})=\boldsymbol{x}\right) \;&=\; \int_{y\in\mathcal{Y}} \mathbb{P}\left(c(\boldsymbol{y})=\boldsymbol{x}|\boldsymbol{y}=y\right)\; f_{\boldsymbol{y}}(y)dy \\ &\overset{\Delta}{=} \int_{y\in\mathcal{Y}} \Delta(y) f_{\boldsymbol{y}}(y)dy \end{aligned} \tag{28.33}$$

where the integration is over the observation space, $y\in\mathcal{Y}$. In the above expression, the term $f_{\boldsymbol{y}}(y)$ denotes the probability density function of the observation and the shorthand notation $\Delta(y)$ denotes the conditional probability that appears multiplying $f_{\boldsymbol{y}}(y)$ in the first line. Since $f_{\boldsymbol{y}}(y)\geq 0$, we can solve (28.32) by seeking a classifier $c(y)$ that maximizes $\Delta(y)$. Now observe that, since the events $\boldsymbol{x}=+1$ and $\boldsymbol{x}=-1$ are mutually exclusive conditioned on $\boldsymbol{y}$:

$$\begin{aligned} \Delta(y) \;&\overset{\Delta}{=}\; \mathbb{P}\left(c(\boldsymbol{y})=\boldsymbol{x}|\boldsymbol{y}=y\right) \\ &=\; \mathbb{P}\left(c(\boldsymbol{y})=+1,\boldsymbol{x}=+1|\boldsymbol{y}=y\right) \;+\; \mathbb{P}\left(c(\boldsymbol{y})=-1,\boldsymbol{x}=-1|\boldsymbol{y}=y\right) \\ &=\; \mathbb{I}\left[c(y)=+1\right]\mathbb{P}\left(\boldsymbol{x}=+1|\boldsymbol{y}=y\right) \;+\; \mathbb{I}\left[c(y)=-1\right]\mathbb{P}\left(\boldsymbol{x}=-1|\boldsymbol{y}=y\right) \end{aligned} \tag{28.34}$$

For any given observation y, we need to select $c(y)$ to maximize $\Delta(y)$. There are only

two possibilities for $c(y)$ in the binary classification problem, either $c(y) = +1$ or $c(y) = -1$:

$$\text{if we set } c(y) = +1, \text{ then } \Delta(y) = \mathbb{P}(\boldsymbol{x} = +1|\boldsymbol{y} = y) \tag{28.35}$$

$$\text{if we set } c(y) = -1, \text{ then } \Delta(y) = \mathbb{P}(\boldsymbol{x} = -1|\boldsymbol{y} = y) \tag{28.36}$$

Therefore, we should set $c(y) = +1$ whenever

$$\mathbb{P}(\boldsymbol{x} = +1|\boldsymbol{y} = y) \geq \mathbb{P}(\boldsymbol{x} = -1|\boldsymbol{y} = y) \;=\; 1 - \mathbb{P}(\boldsymbol{x} = +1|\boldsymbol{y} = y) \tag{28.37}$$

which is equivalent to the condition $\mathbb{P}(\boldsymbol{x} = +1|\boldsymbol{y} = y) \geq 1/2$.

■

28.3.2 Likelihood Ratio Test

The Bayes classifier (28.28) can be expressed in an equivalent form involving a likelihood ratio test. To see this, note from expression (28.28) that deciding on whether $\boldsymbol{x}$ is $+1$ or -1 amounts to checking the inequality:

$$\mathbb{P}(\boldsymbol{x} = +1|\boldsymbol{y} = y) \;\geq\; \mathbb{P}(\boldsymbol{x} = -1|\boldsymbol{y} = y) \;\iff\; \widehat{x}_{\text{bayes}}(y) = +1 \tag{28.38}$$

Using Bayes rule (3.39) for conditional probabilities, the above inequality is equivalent to checking whether

$$f_{\boldsymbol{y}|\boldsymbol{x}}(y|\boldsymbol{x} = +1)\,\mathbb{P}(\boldsymbol{x} = +1) \;\geq\; f_{\boldsymbol{y}|\boldsymbol{x}}(y|\boldsymbol{x} = -1)\,\mathbb{P}(\boldsymbol{x} = -1) \tag{28.39}$$

Let $\pi_{\pm 1}$ denote the prior probabilities for the events $\boldsymbol{x} = +1$ and $\boldsymbol{x} = -1$, i.e.,

$$\pi_{+1} \;\triangleq\; \mathbb{P}(\boldsymbol{x} = +1), \quad \pi_{-1} \;\triangleq\; \mathbb{P}(\boldsymbol{x} = -1) \tag{28.40}$$

where

$$\pi_{-1} + \pi_{+1} = 1 \tag{28.41}$$

Let further $L(y)$ denote the likelihood ratio:

$$L(y) \;\triangleq\; \frac{f_{\boldsymbol{y}|\boldsymbol{x}}(y|\boldsymbol{x} = +1)}{f_{\boldsymbol{y}|\boldsymbol{x}}(y|\boldsymbol{x} = -1)} \tag{28.42}$$

Then, condition (28.39) translates into deciding for $\boldsymbol{x} = +1$ or $\boldsymbol{x} = -1$ depending on whether

$$\boxed{L(y) \underset{-1}{\overset{+1}{\gtrless}} \frac{\pi_{-1}}{\pi_{+1}}} \tag{28.43}$$

This test is equivalent to the Bayes classifier (28.28): it decides for $\boldsymbol{x} = +1$ when the likelihood ratio is larger than or equal to π_{-1}/π_{+}. When the classes are equally probable so that $\pi_{-1} = \pi_{+1} = 1/2$, the threshold value on the right-hand side of (28.43) reduces to one.

Example 28.3 (**Hard classifier**) Let us apply the Bayes classifier (28.28) to the situation encountered earlier in Example 27.3. In that example, we discussed recovering the class variable $\boldsymbol{x} \in \{+1, -1\}$ for cat and dog images from soft measurements $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{v}$

in the presence of additive Gaussian perturbation, $\boldsymbol{v}$.

Given $\boldsymbol{y}$, we would like now to recover $\boldsymbol{x}$ by minimizing the probability of misclassification (rather than the mean-square-error, as was done in Example 27.3). The solution is given by the Bayes classifier (28.28); its computation requires that we evaluate $\mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=y)$. This quantity has already been evaluated in Example 3.17. Indeed, from that example we know that:

$$\mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=y) \;=\; \frac{f_{\boldsymbol{v}}(y-1)}{f_{\boldsymbol{v}}(y+1)+f_{\boldsymbol{v}}(y-1)}, \;\; \text{where; } f_{\boldsymbol{v}}(v)=\mathcal{N}_{\boldsymbol{v}}(0,1) \tag{28.44}$$

Simplifying gives

$$\mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=y) \;=\; \frac{1}{\left(\frac{e^{-(y+1)^2/2}}{e^{-(y-1)^2/2}}\right)+1} \tag{28.45}$$

According to (28.28), we need to compare $\mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=y)$ against the threshold $1/2$. It is easy to verify from the above expression that

$$\mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=y) \geq 1/2 \iff y \geq 0 \tag{28.46}$$

In this way, expression (28.28) for the optimal classifier reduces to

$$\widehat{x}_{\text{bayes}} \;=\; \begin{cases} +1, & \text{when } \; y \geq 0 \\ -1, & \text{otherwise} \end{cases} \tag{28.47}$$

which is equivalent to

$$\widehat{x}_{\text{bayes}} \;=\; \text{sign}(y) \tag{28.48}$$

This is precisely the expression for the sub-optimal MSE estimator we used earlier in (27.35)! Here, we discover that this construction is actually *optimal* but relative to the misclassification criterion (28.27).

Example 28.4 (**Using the likelihood ratio test**) We reconsider the previous example from the perspective of likelihood ratios to arrive at the same conclusion (28.48). Indeed, note that the pdf of the observation $\boldsymbol{y}$ under both classes $\boldsymbol{x} \in \{\pm 1\}$ is Gaussian with means $\{\pm 1\}$ and variances equal to one:

$$f_{\boldsymbol{y}|\boldsymbol{x}}(y|\boldsymbol{x}=+1) \sim \mathcal{N}_{\boldsymbol{y}}(1,1), \qquad f_{\boldsymbol{y}|\boldsymbol{x}}(y|\boldsymbol{x}=-1) \sim \mathcal{N}_{\boldsymbol{y}}(-1,1) \tag{28.49}$$

In other words,

$$f_{\boldsymbol{y}|\boldsymbol{x}}(y|\boldsymbol{x}=+1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-1)^2} \tag{28.50a}$$

$$f_{\boldsymbol{y}|\boldsymbol{x}}(y|\boldsymbol{x}=-1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y+1)^2} \tag{28.50b}$$

so that the likelihood ratio is

$$L(y) = \frac{e^{-\frac{1}{2}(y-1)^2}}{e^{-\frac{1}{2}(y+1)^2}} \tag{28.51}$$

Assuming equally probable classes, we need to compare this ratio against one or, equivalently,

$$\exp\Big\{-\frac{1}{2}(y-1)^2+\frac{1}{2}(y+1)^2\Big\} \;\underset{-1}{\overset{+1}{\gtrless}}\; 1 \tag{28.52}$$

Computing the natural logarithms of both sides, it is straightforward to verify that the above condition reduces to

$$y \underset{-1}{\overset{+1}{\gtrless}} 0 \tag{28.53}$$

which is equivalent to (28.48).

The likelihood ratio test can be illustrated graphically as shown in Fig. 28.2. The figure shows two Gaussian distributions centered at ± 1 and with unit variances. These distributions represent the conditional pdfs (28.50a)–(28.50b) of the observation given $\boldsymbol{x}$. In the example under consideration, the means are symmetric around the origin and both distributions have equal variances. Obviously, more general situations can be considered as well — see Prob. 28.5. For the scenario illustrated in the figure, the likelihood ratio test (28.53) leads to comparing the value of y against zero. That is, given an observation y, we decide that its class is $\boldsymbol{x} = +1$ whenever $y \geq 0$ (i.e., whenever it lies to the right of the zero threshold). Likewise, we decide that its class is $\boldsymbol{x} = -1$ whenever $y < 0$ (i.e., whenever it lies to the left of the zero threshold). The figure highlights in color two small areas under the pdf curves. The smaller area to the right of the zero threshold (colored in red) corresponds to the following probability of error:

$$\begin{aligned}\mathbb{P}(\text{ deciding } \boldsymbol{x} = +1 | \boldsymbol{x} = -1\,) &= \int_0^{\infty} f_{\boldsymbol{y}|\boldsymbol{x}}(y|\boldsymbol{x} = -1)dy \\ &= \int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y+1)^2} dy \;\triangleq\; \epsilon \end{aligned} \tag{28.54}$$

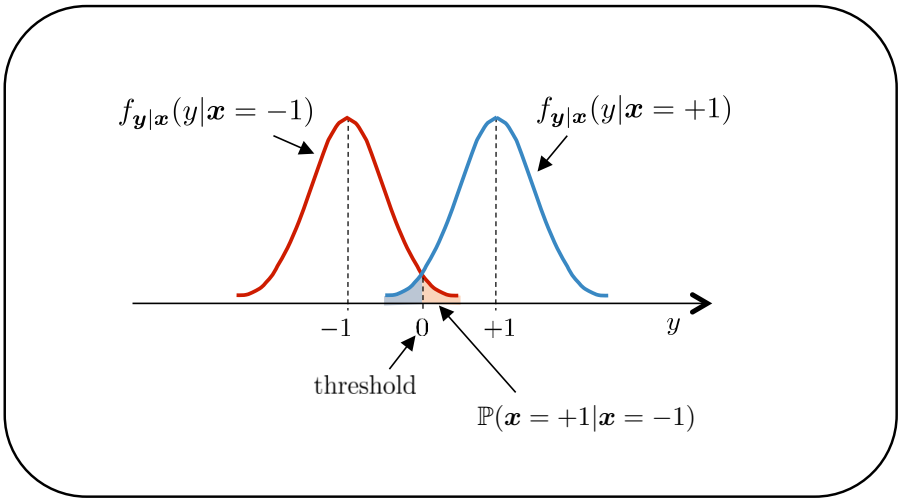

Figure 28.2 Illustration of the conditional Gaussian distributions (28.50a)–(28.50b).

That is, the probability of assigning y wrongly to class $\boldsymbol{x} = +1$ when it actually originates from class $\boldsymbol{x} = -1$ is given by the red-colored area in the figure, whose value we are denoting by ϵ. Likewise, from the same figure, the smaller area to the left of the

zero threshold (colored in blue) corresponds to the following probability of error:

$$\begin{aligned}\mathbb{P}(\text{ deciding } \boldsymbol{x}=-1|\boldsymbol{x}=+1\,) &= \int_{-\infty}^{0} f_{\boldsymbol{y}|\boldsymbol{x}}(y|\boldsymbol{x}=+1)dy \\ &= \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-1)^2} dy \;=\; \epsilon \end{aligned} \tag{28.55}$$

In this example, both error probabilities (or areas) are equal in size and we denote each one of them by ϵ. The probabilities can be combined to determine an expression for the probability of error of the Bayes classifier since:

$$\begin{aligned}\mathbb{P}(\widehat{\boldsymbol{x}}_{\text{bayes}} \neq \boldsymbol{x}) &= \mathbb{P}(\text{ erroneous decisions }) \\ &= \frac{1}{2}\,\mathbb{P}(\text{ deciding } \boldsymbol{x}=+1|\boldsymbol{x}=-1) + \frac{1}{2}\,\mathbb{P}(\text{deciding } \boldsymbol{x}=-1|\boldsymbol{x}=+1) \\ &= \epsilon/2 + \epsilon/2 \\ &\overset{(28.41)}{=} \epsilon \end{aligned} \tag{28.56}$$

Example 28.5 (**Classifying iris flowers**) We reconsider the iris flower dataset encountered earlier in Example 27.4. The top row in Figure 28.3 shows two histogram distributions for the petal length measured in cm for two types of flowers: iris setosa and iris virginica. Each histogram constructs 5 bins based on 50 measurements for each flower type. The width of the bin is 0.32cm for setosa flowers and 0.70cm for virginica flowers. The bottom row shows the same histograms normalized by dividing each bin value by the number of samples (which is 50) and by the bin width (0.32 for setosa flowers and 0.70 for virginica flowers). This normalization results in approximations for the probability density functions. We assume that a flower can only be one of two kinds: either setosa or virginica. Given an observation of a flower with petal length equal to 5.5cm, we would like to decide whether it is of one type or the other. We will be solving classification problems of this type in a more structured manner in later chapters, and in many different ways. The current example is only meant to illustrate Bayes classifiers.

Let $\boldsymbol{x}$ denote the class label, namely, $\boldsymbol{x}=+1$ if the flower is iris setosa and $\boldsymbol{x}=-1$ if the flower is iris virginica. We model the petal length as a random variable $\boldsymbol{y}$. According to the Bayes classifier (28.28), we need to determine the conditional probability $\mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=5.5)$. To do so, we assume the flowers are equally distributed so that

$$\mathbb{P}(\boldsymbol{x}=+1) = \mathbb{P}(\boldsymbol{x}=-1) = 1/2 \tag{28.57}$$

According to Bayes rule (3.42b), we have:

$$\mathbb{P}(\boldsymbol{x}=x|\boldsymbol{y}=y) = \frac{\mathbb{P}(\boldsymbol{x}=x)\; f_{\boldsymbol{y}|\boldsymbol{x}}(y|\boldsymbol{x}=x)}{f_{\boldsymbol{y}}(y)} \tag{28.58}$$

Therefore, we need to evaluate the pdfs $f_{\boldsymbol{y}|\boldsymbol{x}}(y|x)$ and $f_{\boldsymbol{y}}(y)$ that appear on the right-hand side. We do not have these pdfs but we will estimate them from the data measurements by assuming they follow Gaussian distributions. For that purpose, we only need to identify the mean and variance parameters for these distributions; in later chapters, we will learn how to fit more complex distributions into data measurements such as mixtures of Gaussian models.

The sample means and variances for the petal length computed from the respective 50

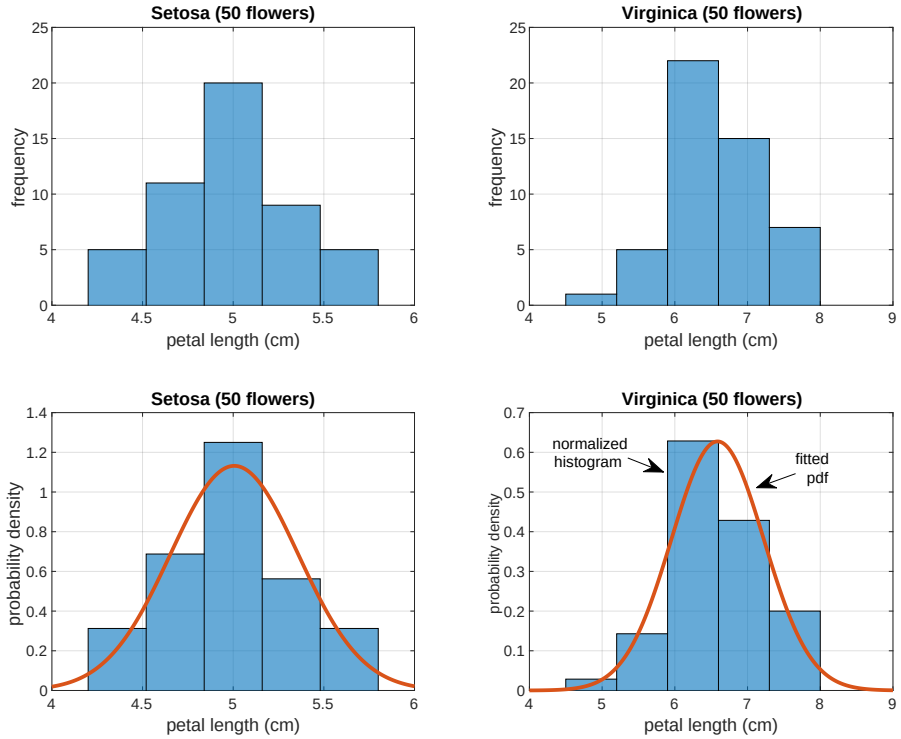

Figure 28.3 (*Top*) Histogram distribution of the petal length measured in cm for iris setosa flowers on the left and for iris virginica flowers on the right. (*Bottom*) The same histogram plots are normalized by dividing the value for each bin by the bin size and by the total number of 50 samples to generate approximate probability distributions for the petal length variable. (*Bottom*) Two Gaussian distributions are fitted on top of the normalized histograms.

measurements for each flower type are found to be:

$$\mathbb{E}\,(\text{petal length}\,\big|\,\text{flower} = \text{setosa}) \approx 5.0060 \tag{28.59a}$$

$$\mathbb{E}\,(\text{petal length}\,\big|\,\text{flower} = \text{virginica}) \approx 6.5880 \tag{28.59b}$$

$$\text{var}(\text{petal length}\,\big|\,\text{flower} = \text{setosa}) \approx 0.1242 \tag{28.59c}$$

$$\text{var}(\text{petal length}\,\big|\,\text{flower} = \text{virginica}) \approx 0.4043 \tag{28.59d}$$

where, for example, the sample mean and variance for the setosa flower are computed by using:

$$\mathbb{E}\,(\text{petal length}\,\big|\,\text{flower} = \text{setosa}) \approx \frac{1}{50}\sum_{n=1}^{50} y_n \;\triangleq\; \bar{y}_{\text{setosa}} \tag{28.60}$$

$$\text{var}(\text{petal length}\,\big|\,\text{flower} = \text{setosa}) \approx \frac{1}{49}\sum_{n=1}^{50}(y_n - \bar{y}_{\text{setosa}})^2 \tag{28.61}$$

Here, the sum is over the 50 setosa samples and y_n is the petal length for the $n-$th

setosa sample.

The bottom row in Figure 28.3 shows two Gaussian distributions with these means and variances fitted on top of the histograms. These are used as approximations for the conditional pds $f_{\boldsymbol{y}|\boldsymbol{x}}(y|\boldsymbol{x}=x)$, namely,

$$f_{\boldsymbol{y}|\boldsymbol{x}}(y|\boldsymbol{x}=\text{setosa}) = \frac{1}{\sqrt{2\pi \times 0.1242}} \exp\left\{-\frac{1}{2 \times 0.1242}(y-5.0060)^2\right\} \tag{28.62}$$

$$f_{\boldsymbol{y}|\boldsymbol{x}}(y|\boldsymbol{x}=\text{virginica}) = \frac{1}{\sqrt{2\pi \times 0.4043}} \exp\left\{-\frac{1}{2 \times 0.4043}(y-6.5880)^2\right\} \tag{28.63}$$

The combined distribution for the petal length variable can then be approximated by

$$f_{\boldsymbol{y}}(y) = \frac{1}{2}\frac{1}{\sqrt{2\pi \times 0.1242}} \exp\left\{-\frac{1}{2 \times 0.1242}(y-5.0060)^2\right\} + \frac{1}{2}\frac{1}{\sqrt{2\pi \times 0.4043}} \exp\left\{\frac{1}{2 \times 0.4043}(y-6.5880)^2\right\} \tag{28.64}$$

since it is equally likely for a petal length to arise from one Gaussian distribution or the other. Figure 28.4 shows the normalized histogram distribution for all 100 petal lengths and fits the sum of two Gaussian distributions on top of it.

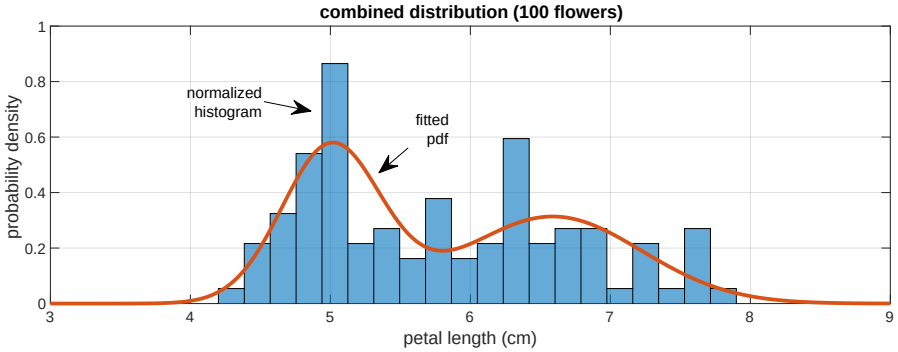

Figure 28.4 Combined normalized histogram for the distribution of the petal length measured in cm for both classes of iris setosa and iris virginica flowers. The sum of two Gaussian distributions is fitted on top of the histogram.

We now have all the elements needed to evaluate the right-hand side of (28.58) for the given petal length of $y = 5.5$ cm. Indeed,

$$\begin{aligned}\mathbb{P}(\boldsymbol{x}=\text{setosa}|\boldsymbol{y}=5.5) &= \frac{\mathbb{P}(\boldsymbol{x}=\text{setosa})\; f_{\boldsymbol{y}|\boldsymbol{x}}(y=5.5|\boldsymbol{x}=\text{setosa})}{f_{\boldsymbol{y}}(y=5.5)} \\ &= \frac{0.5 \times 0.3309}{0.3498} \approx 0.4730\end{aligned} \tag{28.65}$$

This value is less than $1/2$ and we therefore classify the flower as being of the iris virginica type.

28.3.3 Multiclass Classification

Problem 28.3 at the end of the chapter extends conclusion (28.28) to multiclass classification problems where $\boldsymbol{x}$ could assume one of $R \geq 2$ discrete values, say, $\boldsymbol{x} \in \{1, 2, \ldots, R\}$. In this case, the classifier maps the observation vector $\boldsymbol{y}$ to integer values in the range $\{1, 2, \ldots, R\}$, i.e.,

$$c(y) : \mathbb{R}^M \rightarrow \{1, 2, \ldots, R\} \tag{28.66}$$

and its optimal construction is now given by the MAP formulation:

$$\boxed{\widehat{x}_{\text{bayes}} \;=\; \widehat{x}_{\text{MAP}} \;=\; \underset{x \in \{1,2,\ldots,R\}}{\text{argmax}} \;\; \mathbb{P}(\boldsymbol{x} = x | \boldsymbol{y} = y)} \tag{28.67}$$

which is the natural generalization of (28.31). This construction seeks the class x that maximizes the posterior probability given the observation. Since the observation $\boldsymbol{y}$ is random, the resulting classifier $\widehat{\boldsymbol{x}}_{\text{bayes}}$ is also random, i.e., each realization value $\boldsymbol{y} = y$ results in a realization $\widehat{x}_{\text{bayes}}$.

We can assess the probability of erroneous decisions by the Bayes classifier as follows. If the true label corresponding to $\boldsymbol{y} = y$ is x, then the probability of error for this observation is

$$\begin{aligned} \mathbb{P}(\text{error} | \boldsymbol{y} = y) \;&\triangleq\; \mathbb{P}(\widehat{\boldsymbol{x}}_{\text{bayes}} \neq x | \boldsymbol{y} = y) \\ &=\; 1 - \mathbb{P}(\widehat{\boldsymbol{x}}_{\text{bayes}} = x | \boldsymbol{y} = y) \end{aligned} \tag{28.68}$$

If we average over the distribution of the observations, we remove the conditioning over $\boldsymbol{y}$ and arrive at the probability of error for the Bayes classifier denoted by

$$P_e^{\text{bayes}} \;\triangleq\; \mathbb{P}(\widehat{\boldsymbol{x}}_{\text{bayes}} \neq \boldsymbol{x}) \;=\; \int_{y \in \mathcal{Y}} \Big(1 - \mathbb{P}(\widehat{\boldsymbol{x}}_{\text{bayes}}(y) = x_y | \boldsymbol{y} = y)\Big) f_{\boldsymbol{y}}(y) dy \tag{28.69}$$

Here, we are writing x_y, with an explicit subscript y, inside the integral expression to emphasize that x_y is the label that corresponds to the observation y. The following bound holds.

Theorem 28.1. (Performance of Bayes classifier) *Consider a multiclass classification problem with R labels, $x = 1, 2, \ldots, R$. It holds that*

$$P_e^{\text{bayes}} \;\leq\; \frac{R-1}{R} \tag{28.70}$$

Proof: We employ the result of future Theorem 52.1 in the following argument. By construction, the Bayes classifier minimizes the probability of erroneous decisions, i.e.,

$$\widehat{\boldsymbol{x}}_{\text{bayes}} \;=\; \underset{\widehat{\boldsymbol{x}} = c(\boldsymbol{y})}{\text{argmin}} \; \mathbb{P}(c(\boldsymbol{y}) \neq \boldsymbol{x}) \tag{28.71}$$

In future Theorem 52.1, we will study one particular suboptimal classifier called the nearest-neighbor rule. It is suboptimal in the sense that it does not minimize the probability of error. We denote its probability of error by P_e, which is of course worse than

that of the *optimal* Bayes classifier, i.e., $P_e^{\text{bayes}} \leq P_e$. We will establish in Theorem 52.1 that the probability of error for the nearest-neighbor classifier is upper bounded by

$$P_e \;\leq\; P_e^{\text{bayes}}\left(2 - \frac{R}{R-1}P_e^{\text{bayes}}\right) \tag{28.72}$$

The right-hand side is a quadratic function in P_e^{bayes}; its maximum is attained at the location $P_e^{\text{bayes}} = (R-1)/R$. Substituting this value into the upper bound we get

$$P_e^{\text{bayes}} \;\leq\; P_e \;\leq\; \frac{R-1}{R}\left(2 - \frac{R}{R-1}\frac{R-1}{R}\right) = \frac{R-1}{R} \tag{28.73}$$

as claimed.

■

28.3.4 Discriminant Function Structure

Regardless of whether we are dealing with a binary or multiclass classification problem, both solutions (28.31) and (28.67) admit a *discriminant* function interpretation. The solution first associates a discriminant function with each discrete class x, which we denote by:

$$d_x(y) \;\triangleq\; \mathbb{P}(\boldsymbol{x} = x|\boldsymbol{y} = y), \quad x = 1, 2, \ldots, R \tag{28.74}$$

This function measures the likelihood that observation y belongs to class x. Then, the optimal classifier selects the class label, $\widehat{x}_{\text{bayes}}$, with the largest discrimination value — see Fig. 28.5. We will encounter this type of structure multiple times in our treatment — see, e.g., future expression (56.10), which will arise in the design of linear discriminant classifiers; see also future Prob. 56.1.

28.4 LOGISTIC REGRESSION INFERENCE

We will encounter other choices for the loss function $Q(\boldsymbol{x}, \widehat{\boldsymbol{x}})$ in our future development. One of them is the logistic regression loss for binary variables $\boldsymbol{x} \in \{\pm 1\}$ defined by

$$Q(\boldsymbol{x}, \widehat{\boldsymbol{x}}) = \ln(1 + e^{-\boldsymbol{x}\widehat{\boldsymbol{x}}}), \quad \widehat{\boldsymbol{x}} = c(\boldsymbol{y}) \tag{28.75}$$

We will study logistic regression in greater detail later in Chapter 59. Here we provide some motivation based on the following theorem, which provides an expression for the optimal estimator $\widehat{\boldsymbol{x}}$ that follows from minimizing the logistic risk.

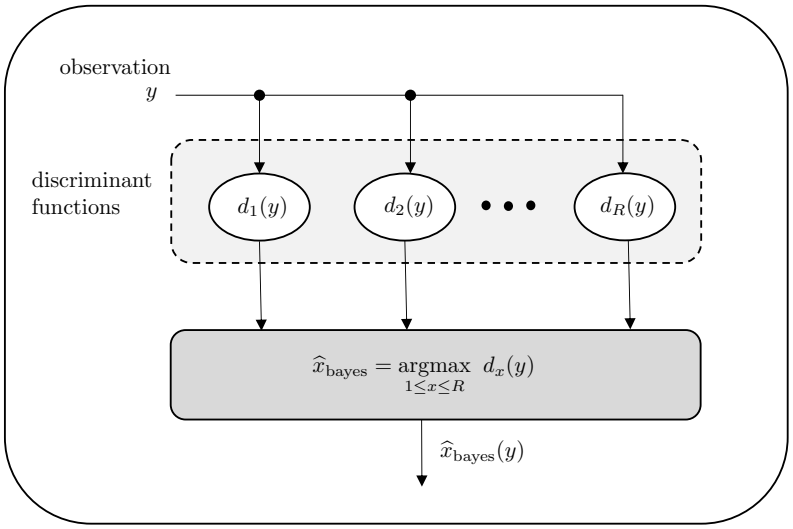

Figure 28.5 Classifier structure in the form of a collection of discriminant functions, $d_x(y)$; one for each discrete value x. For each observation vector, y, the optimal classifier is obtained by selecting the class label $\widehat{x}_{\text{bayes}}$ with the largest discrimination value. This value is denoted by $\widehat{x}_{\text{bayes}}(y)$ at the bottom of the figure.

Theorem 28.2. (Minimizer of logistic risk) *Consider a binary classification problem where* $\boldsymbol{x} \in \{\pm 1\}$ *and the following Bayesian inference problem:*

$$\widehat{\boldsymbol{x}}_{\text{LR}} \;=\; \underset{\widehat{\boldsymbol{x}}=c(\boldsymbol{y})}{\text{argmin}}\ \mathbb{E}\,\ln(1+e^{-\boldsymbol{x}\widehat{\boldsymbol{x}}}) \tag{28.76}$$

The optimal estimator that minimizes the above risk is given by

$$\widehat{x}_{\text{LR}} = c^o(y) = \ln\left(\frac{\mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=y)}{\mathbb{P}(\boldsymbol{x}=-1|\boldsymbol{y}=y)}\right) \;\triangleq\; \text{logit}(y) \tag{28.77}$$

where we are denoting the ratio by the notation $\text{logit}(y)$*. The sign of* $\widehat{x}_{\text{LR}}$ *determines the logistic classifier for* $\boldsymbol{x}$*.*

Proof: Let $R(c) = \mathbb{E}\,\ln(1+e^{-\boldsymbol{x}\widehat{\boldsymbol{x}}})$ denote the logistic risk. We recall the conditional mean property from Prob. 3.25 that $\mathbb{E}\,\boldsymbol{a} = \mathbb{E}\,[\mathbb{E}\,(\boldsymbol{a}|\boldsymbol{b})]$, for any two random variables $\boldsymbol{a}$ and $\boldsymbol{b}$. Applying this property to the logistic risk we get

$$R(c) = \mathbb{E}_{\boldsymbol{y}}\Big\{\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\Big(\ln(1+e^{-\boldsymbol{x}\widehat{\boldsymbol{x}}})\,|\,\boldsymbol{y}\Big)\Big\} \tag{28.78}$$

where the inner expectation is over the conditional pdf $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$, while the outer expectation is over the distribution of $\boldsymbol{y}$. The inner expectation is always nonnegative. Therefore, it is sufficient to examine the problem of minimizing its value to arrive at a minimizer for $R(c)$. Since $\boldsymbol{x}$ assumes the discrete values ± 1, we can assess the inner expectation and write

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\Big(\ln(1+e^{-\boldsymbol{x}\widehat{\boldsymbol{x}}})\,|\,\boldsymbol{y}=y\Big)\\ &=\mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=y)\ln(1+e^{-\widehat{x}})\;+\;\mathbb{P}(\boldsymbol{x}=-1|\boldsymbol{y}=y)\ln(1+e^{\widehat{x}})\end{aligned} \tag{28.79}$$

Differentiating over $\widehat{x}$ and setting the derivative to zero at $\widehat{x}_{\rm LR} = c^o(y)$ gives

$$-\mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=y)\frac{e^{-\widehat{x}_{\rm LR}}}{1+e^{-\widehat{x}_{\rm LR}}} + \mathbb{P}(\boldsymbol{x}=-1|\boldsymbol{y}=y)\frac{e^{\widehat{x}_{\rm LR}}}{1+e^{\widehat{x}_{\rm LR}}} = 0 \tag{28.80}$$

or, equivalently,

$$\frac{\mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=y)}{\mathbb{P}(\boldsymbol{x}=-1|\boldsymbol{y}=y)} = \frac{e^{\widehat{x}_{\rm LR}}\left(1+e^{-\widehat{x}_{\rm LR}}\right)}{e^{-\widehat{x}_{\rm LR}}\left(1+e^{\widehat{x}_{\rm LR}}\right)} = \frac{e^{\widehat{x}_{\rm LR}}\left(1+e^{-\widehat{x}_{\rm LR}}\right)}{1+e^{-\widehat{x}_{\rm LR}}} = e^{\widehat{x}_{\rm LR}} = e^{c^o(h)} \tag{28.81}$$

from which we arrive at (28.77). We explain in (28.85) that the sign of $\widehat{x}_{\rm LR}$ determines the Bayes classifier for $\boldsymbol{x}$.

■

The reason for the qualification "logistic" is because the solution (28.77) expresses the conditional label probabilities in the form of logistic functions evaluated at $\widehat{x}_{\rm LR}$. Indeed, it follows from (28.77) that

$$\mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=y) = \frac{1}{1+e^{-\widehat{x}_{\rm LR}}} \tag{28.82a}$$

$$\mathbb{P}(\boldsymbol{x}=-1|\boldsymbol{y}=y) = \frac{1}{1+e^{+\widehat{x}_{\rm LR}}} \tag{28.82b}$$

Figure 28.6 illustrates the logistic functions $1/(1+e^{-z})$ and $1/(1+e^{z})$. Note that these functions return values between 0 and 1 (as befits a true probability measure).

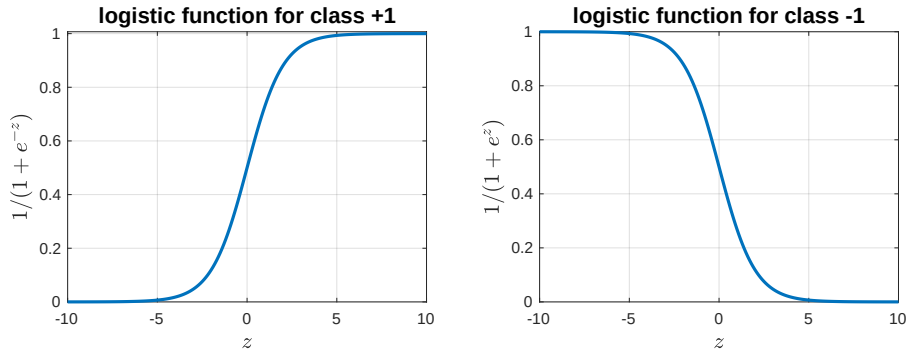

Figure 28.6 Typical behavior of logistic functions for two classes. The figure shows plots of the functions $1/(1+e^{-z})$ (*left*) and $1/(1+e^{z})$ (*right*) assumed to correspond to classes $+1$ and -1, respectively.

Note that the logit of y is the logarithm of the odds of y belonging to one class or the other:

$$\text{odds}(y) \triangleq \frac{\mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=y)}{\mathbb{P}(\boldsymbol{x}=-1|\boldsymbol{y}=y)} = \frac{\mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=y)}{1-\mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=y)} \tag{28.83}$$

so that

$$\text{odds}(y) \geq 1 \iff \mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=y) \geq 1/2 \tag{28.84}$$

which agrees with the condition used by the Bayes classifier. Therefore, once the logarithm is applied to the odds function, the value of $\widehat{x}_{\rm LR}$ will be nonnegative when $\mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=y)\geq 1/2$ and negative otherwise. For this reason, we can use the logistic estimator $\widehat{\boldsymbol{x}}_{\rm LR}$ to deduce the value of the Bayes classifier by rewriting (28.28) in the form:

$$\widehat{x}_{\rm bayes} \;=\; \begin{cases} +1, & \text{when } \widehat{x}_{\rm LR}=\text{logit}(y)\geq 0 \\ -1, & \text{otherwise} \end{cases} \tag{28.85}$$

Example 28.6 (**Exponential loss and boosting**) We will encounter later in Chapter 62, while studying boosting algorithms for learning, the exponential loss function $Q(\boldsymbol{x},\widehat{\boldsymbol{x}})=e^{-\boldsymbol{x}\widehat{\boldsymbol{x}}}$. Consider again a binary classification problem where $\boldsymbol{x}\in\{\pm 1\}$ and assume we seek to solve

$$\widehat{\boldsymbol{x}}_{\rm EXP} \;=\; \underset{\widehat{\boldsymbol{x}}=c(\boldsymbol{y})}{\text{argmin}} \left\{\mathbb{E}\, e^{-\boldsymbol{x}\widehat{\boldsymbol{x}}}\right\} \tag{28.86}$$

Then, the same derivation will lead to

$$\widehat{x}_{\rm EXP}=\frac{1}{2}\ln\left(\frac{\mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=y)}{\mathbb{P}(\boldsymbol{x}=-1|\boldsymbol{y}=y)}\right)=\frac{1}{2}\text{logit}(y) \tag{28.87}$$

with an additional scaling by $1/2$.

Example 28.7 (**Motivating the logistic risk**) One way to motivate the logistic risk function used in (28.76) is to invoke the Kullback-Leibler (KL) divergence measure. Let $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$ denote some unknown conditional pdf that we wish to estimate, where we are using the pdf notation $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$ instead of the more explicit form $\mathbb{P}(\boldsymbol{x}=+1|\boldsymbol{y}=y)$ for convenience. Assume we opt to use a sigmoid function to approximate the unknown pdf and choose the approximation to be of the form:

$$g_{\boldsymbol{x}|\boldsymbol{y}}(x|y)=\frac{1}{1+e^{-xc(y)}} \tag{28.88}$$

for some function $c(y)$ to be determined. Such sigmoidal functions are particularly useful to model distributions for binary-valued discrete variables $\boldsymbol{x}\in\{\pm 1\}$ since they return values between 0 and 1 (as befitting a true probability measure).

Recall from the discussion in Chapter 6 that the KL divergence is a useful measure of closeness between probability distributions. Accordingly, we can choose $c(y)$ to minimize the KL divergence between $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$ and $g_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$, i.e.,

$$c^o(y) \;\triangleq\; \underset{c(y)}{\text{argmin}} \;\; \mathbb{E}_f \left\{\ln\left(\frac{f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)}{g_{\boldsymbol{x}|\boldsymbol{y}}(x|y)}\right)\right\} \tag{28.89}$$

where the expectation is relative to the unknown distribution $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$. But since this distribution is independent of $c(y)$, the above problem is equivalent to

$$c^o(y)=\underset{c(y)}{\text{argmin}} \;\; \left\{-\mathbb{E}_f \ln g_{\boldsymbol{x}|\boldsymbol{y}}(x|y)\right\} \tag{28.90}$$

Substituting the assumed form (28.88) for $g_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$, we arrive at

$$c^o(y)=\underset{c(y)}{\text{argmin}} \;\; \mathbb{E}\ln\left(1+e^{-\boldsymbol{x}c(\boldsymbol{y})}\right) \tag{28.91}$$

which agrees with the logistic risk formulation (28.76).

28.5 DISCRIMINATIVE AND GENERATIVE MODELS

The solution of Bayesian inference problems requires knowledge of the conditional distribution $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$, as is evident from (28.5). For example, the mean-square error estimator, $\widehat{\boldsymbol{x}}_{\rm MSE}$, corresponds to the mean of this conditional distribution, while the maximum a-posteriori estimator, $\widehat{\boldsymbol{x}}_{\rm MAP}$, corresponds to the location of its mode. The same is true for the Bayes classifier when $\boldsymbol{x}$ is discrete since it requires knowledge of the conditional probabilities $\mathbb{P}(\boldsymbol{x} = r|\boldsymbol{y} = y)$.

Implementing inference solutions that depend on knowledge of the conditional distribution $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$ can be challenging, as explained below. For this reason, in future chapters we will be pursuing various methodologies that attempt to solve the inference problem of predicting $\boldsymbol{x}$ from $\boldsymbol{y}$ in different ways, either by insisting on approximating the conditional pdf $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$ or by ignoring it altogether and working directly with data realizations instead. Four broad classes of approaches stand out:

(a) (**Approaches based on discriminative models**). Even if $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$ were known in closed-form, computing its mean or mode locations can be demanding and need not admit closed-form solutions. In later chapters, we will assume that this conditional distribution has particular forms that are easy to work with. These approximate techniques will belong to the class of *discriminative methods* because they assume models for the conditional pdf $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$ and allow us to discriminate between classes.

(b) (**Approaches based on generative models**). In some other instances, we may actually have more information than $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$ and know the full joint distribution $f_{\boldsymbol{x},\boldsymbol{y}}(x, y)$. In principle, this joint distribution should be sufficient to determine the conditional pdf since, from Bayes rule:

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) = \frac{f_{\boldsymbol{x},\boldsymbol{y}}(x, y)}{f_{\boldsymbol{y}}(y)} \tag{28.92}$$

where the distribution for $\boldsymbol{y}$ (also called its *evidence*), and which appears in the denominator, can be determined by marginalizing the joint distribution:

$$f_{\boldsymbol{y}}(y) = \int_{y\in\mathcal{Y}} f_{\boldsymbol{x},\boldsymbol{y}}(x, y)dy \tag{28.93}$$

The difficulty, however, lies in the fact that this marginalization does not always admit a tractable closed-form expression. In later chapters, we will describe various approximation methods to forgo the need to evaluate the evidence, such as the Laplace method, the Markov chain Monte Carlo method, and the expectation propagation method, in addition to variational inference techniques.

Besides solving inference problems, knowledge of the joint distribution can also be used to determine the *generative distribution*, $f_{\boldsymbol{y}|\boldsymbol{x}}(y|x)$, which allows us to generate samples y from knowledge of x. We will encounter

many examples of this approach in the form of Gaussian mixture models (GMM), restricted Boltzmann machines (RBMs), hidden Markov models (HMMs), and variational autoencoders. These techniques belong to the class of *generative methods* because they allow us to determine models for the *reverse* conditional pdf $f_{\boldsymbol{y}|\boldsymbol{x}}(y|x)$.

Observe that the main difference between the *discriminative* and *generative* approaches is that the former works with (or approximates) $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$ while the latter works with (or approximates) $f_{\boldsymbol{y}|\boldsymbol{x}}(y|x)$:

$$\textbf{discriminative approach} \Longrightarrow \textbf{works with } f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) \tag{28.94a}$$

$$\textbf{generative approach} \Longrightarrow \textbf{works with } f_{\boldsymbol{y}|\boldsymbol{x}}(y|x) \tag{28.94b}$$

(c) (**Approaches based on model-based inference**). The inference methods under (a) and (b) work directly with joint or conditional distributions for the variables involved, namely, $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$ and $f_{\boldsymbol{x},\boldsymbol{y}}(x,y)$. These distributions are either known or approximated. The approximations can take different forms. For example, one can assume a parametric model for the conditional pdf $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y;\theta)$, assumed parameterized by some θ (such as assuming a Gaussian form with its mean and variance playing the role of the parameter θ). One can then seek to estimate θ in order to fit the assumed distribution model onto the data, and proceed from there to perform inference. The maximum-likelihood technique and Gaussian mixture models are examples of this approach. Alternatively, one can assume a model relating the variables $\{\boldsymbol{x},\boldsymbol{y}\}$ directly, such as a state-space model or a linear regression model that tells us how $\boldsymbol{x}$ generates $\boldsymbol{y}$. The Kalman and particle filter solutions are prominent examples of this approach. The assumed state-space models implicitly define a conditional distribution linking $\boldsymbol{x}$ and $\boldsymbol{y}$. One can then work with the model equations to perform inference. In many instances of interest, the assumed model removes the need to know the full conditional pdf $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$, and only some of its moments are necessary. We will encounter our first example of this scenario in the next chapter. There, by assuming a linear regression model, it will be seen that the Bayesian solution only requires knowledge of the first and second-order moments of the variables $\{\boldsymbol{x},\boldsymbol{y}\}$, namely, their means, cross-covariance, and variances. Our treatment of inference methods will cover steps (a)–(c) in some detail, and introduce various techniques that fit into one of these approaches starting from the next chapter.

(d) (**Approaches based on data-driven inference**). Model-based solutions can be complex and computationally demanding; for example, it is not unusual for these implementations to involve the computation of challenging integrals or to require fitting complex distributions onto data. Moreover, in a large number of applications, designers do not know the general forms of the conditional or joint probability distributions, or even models linking the

variables, and will only have access to data realizations $\{x(n), y_n\}$ that arise from these distributions or models. For this reason, we will be motivated to introduce a variety of *learning methods* that perform inference directly from data. In contrast to the inference methods under (a)–(c), which attempt to approximate or emulate the underlying distributions or models, learning algorithms will be largely data-driven and will arrive at inference conclusions without the need to know or determine explicitly the forms of the underlying distributions or models.

The learning methods will differ by how they process the data. Some methods will operate directly on $\{x(n), y_n\}$ to estimate values for the conditional probabilities (rather than their actual forms). Examples include the nearest neighbor (NN) rule and self-organizing maps (SOMs). Other learning methods will go a step further. They will require the mapping $c(\boldsymbol{y})$ to be an affine model of the observations, say, $c(\boldsymbol{y}) = y^{\mathsf{T}}w - \theta$, for some parameters $w \in \mathbb{R}^M$ and $\theta \in \mathbb{R}$, or use some more involved nonlinear models as happens with kernel methods and neural networks. For affine models, the Bayesian inference problem (28.1) will reduce to minimizing over the parameters (w, θ):

$$(w^o, \theta^o) \;=\; \underset{w,\theta}{\operatorname{argmin}}\; \mathbb{E}\, Q(w, \theta; \boldsymbol{x}, \boldsymbol{y}) \tag{28.95}$$

For example, in the mean-square error case, the loss function will take the form (for scalar $\boldsymbol{x}$):

$$\begin{aligned} Q(\boldsymbol{x}, \widehat{\boldsymbol{x}}) &= (\boldsymbol{x} - \widehat{\boldsymbol{x}})^2 \\ &= (\boldsymbol{x} - \boldsymbol{y}^{\mathsf{T}}w + \theta)^2 \\ &= Q(w, \theta; \boldsymbol{x}, \boldsymbol{y}) \end{aligned} \tag{28.96}$$

which shows that the loss is dependent on the parameters (w, θ) and on the variables $\{\boldsymbol{x}, \boldsymbol{y}\}$. Formulation (28.95) is an optimization problem with a stochastic risk. If we observe a collection of realizations $\{x(n), y_n\}$ arising from the underlying (but unknown) distribution $f_{\boldsymbol{x},\boldsymbol{y}}(x, y)$, then we already know how to run stochastic gradient algorithms and many variations thereof to seek the optimizers (w^o, θ^o).

We will also consider empirical risk versions of problem (28.95) such as

$$(w^\star, \theta^\star) \;=\; \underset{w,\theta}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{n=0}^{N-1} Q(w, \theta; x(n), y_n) \right\} \tag{28.97}$$

Most learning algorithms discussed in later chapters will correspond to stochastic approximation methods applied to the minimization of the stochastic or empirical formulations similar to (28.95) or (28.97). We will encounter a variety of methods that fit into this paradigm such as support vector machines, the Perceptron, kernel methods, and neural networks. These methods will differ by their choice of the loss function.

For ease of reference, we represent the various inference and learning methods described above in the diagram shown in Fig. 28.7, where we also embedded the encoder and decoder cells that map the variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ to each other.

28.6 COMMENTARIES AND DISCUSSION

Bayesian and non-Bayesian formulations. In statistics, there is a clear distinction between the classical approach and the Bayesian approach to estimation. In the classical approach, the unknown quantity to be estimated is modeled as a deterministic but unknown constant. One popular non-Bayesian technique is the *maximum likelihood* (ML) approach discussed later in Chapter 31. This approach was developed by the English statistician **Ronald Fisher (1890–1962)** in the works by Fisher (1912,1922,1925) — see the presentations by Pratt (1976), Savage (1976), and Aldrich (1997). The maximum likelihood formulation does not assume any prior distribution for the unknown x and relies on maximizing a certain *likelihood function*

The Bayesian approach, on the other hand, models both the unknown quantity and the observation as random variables. It allows the designer to incorporate prior knowledge about the unknown into the solution, such as information about its probability density function. This fact helps explain why Bayesian techniques are dominant in many successful filtering and estimation designs. We will provide a more detailed comparison of the maximum-likelihood and Bayesian approaches in the comments at the end of Chapter 31. For additional information on Bayesian and non-Bayesian techniques, readers may refer to the texts by Zacks (1971), Box and Tiao (1973), Scharf (1991), Kay (1993), Cassella and Berger (2002), Cox (2006), Hogg and McKean (2012), and Van Trees (2013).

Bayes classifiers. The Bayes classifier (28.67) is one notable application of the method of Bayesian inference in statistical analysis. Some early references on the application of Bayesian inference to classification problems include the works by Chow (1957) and Miller (1962) and the texts by Davenport and Root (1958), Middleton (1960), and Wald (1950). For readers interested in learning more about Bayes classifiers and Bayesian inference, there are many available treatments in the literature including, among others, the textbooks by Bernardo and Smith (2000), Lee (2002), DeGroot (2004), Cox (2006), Bolstad (2007), Robert (2007), Hoff (2009), and Young and Smith (2010).

Likelihood ratio tests. In expression (28.31) we showed that the optimal classifier that minimizes the probability of misclassification can be obtained by maximizing the posterior probability of the variable $\boldsymbol{x}$ given the observation $\boldsymbol{y} = y$. This construction provides a useful interpretation for the Bayes classifier as a maximum a-posteriori (MAP) solution — see Duda, Hart, and Stork (2000), Webb (2002), Bishop (2007), and Theodoridis and Koutroumbas (2008). In Sec. 28.3.2, we explained how the Bayes classifier (28.28) can be recast in terms of the likelihood ratio test (28.43). This reformulation brings forth connections with another notable framework in statistical analysis, namely, the solution of detection problems by evaluating likelihood ratios and comparing them against threshold values. There is an extensive literature on this important topic, starting with the seminal works by Neyman and Pearson (1928,1933), which laid the foundation for most of the subsequent development in this field. In one of its most basic forms, the Neyman-Pearson construction allows us to select between two simple hypotheses represented by parameter values x_o and x_1. For example, in the context of binary classification, the parameter x_o could be chosen to represent class $+1$, while the parameter x_1 could be chosen to represent class -1. The two hypotheses are then

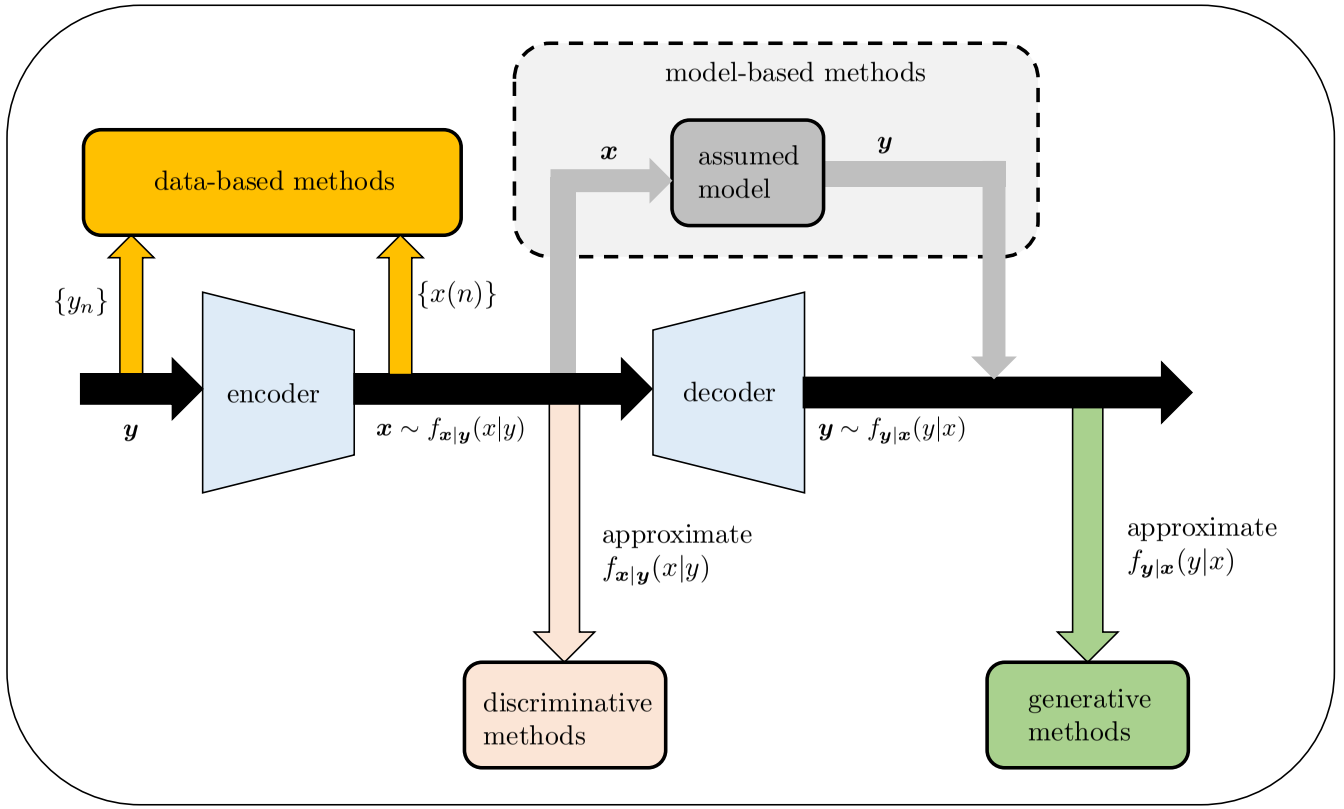

Figure 28.7 Schematic representation of inference and learning approaches based on discriminative methods, generative methods, model-based methods, and data-based methods.

stated as follows:

$$\begin{cases} H_o : x = x_o, & (\textbf{null or positive hypothesis}) \\ H_1 : x = x_1, & (\textbf{alternative or negative hypothesis}) \end{cases} \tag{28.98}$$

It is customary to refer to H_o as the null or positive hypothesis, while H_1 is the alternative or negative hypothesis. Given an observation y, the Neyman-Pearson test would accept H_o in lieu of H_1 (i.e., declare that the null hypothesis is valid) when the following likelihood ratio exceeds some threshold value η:

$$L(y) \triangleq \frac{f_{\boldsymbol{y}|\boldsymbol{x}}(y|\boldsymbol{x}=x_o)}{f_{\boldsymbol{y}|\boldsymbol{x}}(y|\boldsymbol{x}=x_1)} \underset{H_1}{\overset{H_o}{\gtrless}} \eta \tag{28.99}$$

The value of η is usually selected to ensure that some upper bound, denoted by α, is imposed on the probability of erroneously rejecting H_o when H_o is true. The resulting type-I error, or the probability of falsenegatives or *missed detection*, is given by:

$$\begin{aligned} &(\textbf{false negative or type-I error}) \\ &\mathbb{P}(L(y) < \eta | H_o) \;=\; \mathbb{P}\,(\text{reject } H_o | \text{when } H_0 \text{ is true}) \;<\; \alpha \end{aligned} \tag{28.100}$$

On the other hand, the probability of false positives or *false alarm*, also called type-II error, corresponds to

$$\begin{aligned} &(\textbf{false positive or type-II error}) \\ &\beta = \mathbb{P}(L(y) \geq \eta | H_1) \;=\; \mathbb{P}\,(\text{accept } H_o | \text{when } H_1 \text{ is true}) \end{aligned} \tag{28.101}$$

The Neyman-Pearson theory establishes that the likelihood test (28.99) is the *most powerful* test at level α. This means that it is the test that results in the largest *power* defined as the following probability:

$$\text{power} \triangleq \mathbb{P}\,(\text{reject } H_o | \text{when } H_1 \text{ is true}) \;=\; 1 - \beta \tag{28.102}$$

which measures the ability of the test to reject H_o when H_1 is present. Table 28.1 summarizes the various decision possibilities and their respective probabilities.

Table 28.1 Definitions of the probabilities of false negatives, false positives, and the power of a hypothesis test.

	H_o **is true**	H_1 **is true**
accept H_o	correct decision, $1-\alpha$.	type-II error (false positive), β.
reject H_o	type-I error (false negative), α.	correct decision (power), $1-\beta$.

Returning to the Bayes classifier, we noted in expression (28.43) that the threshold value η should be selected as the ratio $\eta = \pi_{-1}/\pi_{+1}$, in terms of the priors for the classes ± 1. Moreover, from expressions (28.54)–(28.55), we can deduce the probabilities of errors of types-I and II (i.e., the fraction of false negatives and false positives by the classifier) for the situation discussed in the example:

$$\mathbb{P}\,(\text{deciding } \boldsymbol{x} = -1 | \boldsymbol{x} = +1) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-1)^2} dy \;=\; \epsilon, \quad (\textbf{type-I}) \tag{28.103a}$$

$$\mathbb{P}\,(\text{deciding } \boldsymbol{x} = +1 | \boldsymbol{x} = -1) = \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y+1)^2} dy \;=\; \epsilon, \quad (\textbf{type-II}) \tag{28.103b}$$

Consequently, for this example, $\alpha = \beta = \epsilon$, and the resulting power level is

$$\mathbb{P}\,(\text{deciding } \boldsymbol{x} = -1 | \boldsymbol{x} = -1) \;=\; 1 - \epsilon \tag{28.104}$$

For further reading on hypothesis testing and statistical inference, some useful references include the texts by Kay (1998), Poor (1998), Cassella and Berger (2002), DeGroot (2004), Lehmann and Romano (2005), Cox (2006), Levy (2008), Young and Smith (2010), and Van Trees (1968,2013).

Beta distribution. The Beta distribution (28.20), also known as the Beta distribution of first-kind, is very useful to model random variables that are confined to the finite interval $[0, 1]$. It is parameterized by two positive shape parameters a and b and includes the uniform distribution as a special case. It is often used as a prior in Bayesian inference, as was illustrated in Example 28.2. For more information on the Beta distribution, the reader may consult the texts by Hahn and Shapiro (1994), Johnson, Kotz, and Balakrishnan (1995), and Gupta and Nadarajah (2004).

PROBLEMS

28.1 Motivated by the $0/1-$loss (28.12), consider the alternative loss function:

$$Q(\boldsymbol{x}, \widehat{\boldsymbol{x}}) \triangleq \begin{cases} 1, & |\boldsymbol{x} - \widehat{\boldsymbol{x}}| > \epsilon \\ 0, & |\boldsymbol{x} - \widehat{\boldsymbol{x}}| \leq \epsilon \end{cases}$$

for some small $\epsilon > 0$. Show that

$$\mathbb{E}\, Q(\boldsymbol{x}, \widehat{\boldsymbol{x}} | \boldsymbol{y} = y) \;=\; 1 - \int_{\widehat{x}-\epsilon}^{\widehat{x}+\epsilon} f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) dx$$

28.2 Consider a collection of N independent Gaussian realizations $\{\boldsymbol{y}_n\}$ with mean μ and unit variance, i.e., $\boldsymbol{y}_n \sim \mathcal{N}_{\boldsymbol{y}_n}(\mu, 1)$ for $n = 0, 1, \ldots, N-1$. The mean μ is unknown but arises from a Gaussian prior distribution $\boldsymbol{\mu} \sim \mathcal{N}_{\boldsymbol{\mu}}(0, \sigma_\mu^2)$ with known variance.

(a) Determine the posterior distribution $f_{\boldsymbol{\mu}|\boldsymbol{y}_0,\ldots,\boldsymbol{y}_{N-1}}(\mu | y_0, y_1, \ldots, y_{N-1})$.

(b) Determine the optimal mean-square error (MSE) estimator of μ.

(c) Determine the maximum a-posterior (MAP) estimator for μ.

(d) Determine the mean absolute error (MAE) estimator for μ.

28.3 Refer to the derivation of the Bayes classifier (28.28). We wish to extend the solution to multiclass classification problems consisting of R classes, say, $\boldsymbol{x} \in \{1, 2, \ldots, R\}$. Given $\boldsymbol{y}$, we again seek to solve over all possible classifiers: $\min_{c(y)} \; \mathbb{P}\,(c(\boldsymbol{y}) \neq \boldsymbol{x})$. Show that the Bayes classifier in this case is given by the MAP construction

$$\widehat{x}_{\text{bayes}} \;=\; \underset{1 \leq x \leq R}{\text{argmax}}\; \mathbb{P}(\boldsymbol{x} = x | \boldsymbol{y} = y)$$

28.4 A binary label $\boldsymbol{x} \in \{+1, -1\}$ is observed under zero-mean additive Gaussian noise $\boldsymbol{v}$ with variance σ_v^2. The observation is denoted by $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{v}$. Assume $\boldsymbol{x} = +1$ with probability p and $\boldsymbol{x} = -1$ with probability $1 - p$. Determine the form of the Bayes classifier. Compare with the result of Example 28.3.

28.5 Consider a binary classification problem in which $\boldsymbol{x} = +1$ with probability p and $\boldsymbol{x} = -1$ with probability $1 - p$. The observation is scalar valued, $\boldsymbol{y} \in \mathbb{R}$, and it has a Gaussian distribution with mean m_{+1} and variance σ_{+1}^2 when $\boldsymbol{x} = +1$, and mean m_{-1} and variance σ_{-1}^2 when $\boldsymbol{x} = -1$.

(a) Determine the form of the Bayes classifier.

(b) Assume $\sigma_{+1}^2 = \sigma_{-1}^2 = \sigma^2$ and $m_{+1} > m_{-1}$. Determine an expression for the probability of error of this classifier.

28.6 Consider a binary classification problem in which $\boldsymbol{x} = +1$ with probability p and $\boldsymbol{x} = -1$ with probability $1 - p$. The observation $\boldsymbol{y}$ is $M-$dimensional, $\boldsymbol{y} \in \mathbb{R}^M$, and it has a Gaussian distribution with mean m_{+1} and covariance matrix Σ_{+1} when $\boldsymbol{x} = +1$,

and mean m_{-1} and covariance matrix Σ_{-1} when $\boldsymbol{x} = -1$. Follow the log-likelihood ratio test of Sec. 28.3.2 to determine the form of the Bayes classifier.

28.7 We consider binary classification problems with $\boldsymbol{x} = \pm 1$. The Bayes classifier was derived by minimizing the probability of erroneous decisions, as defined by (28.27). There are two types of error that can occur: assigning an observation to $\boldsymbol{x} = +1$ when it actually arises from $\boldsymbol{x} = -1$ or, conversely, assigning an observation to $\boldsymbol{x} = -1$ when it arises from $\boldsymbol{x} = +1$. The formulation (28.27) treats these two errors equally. However, there are situations where one type of error is more serious than the other, e.g., in deciding whether a person has a particular disease or not. To address such situations, we can assign weights to the errors and define instead a weighted risk function, also called the *Bayes risk*, for the classifier $c(\boldsymbol{y})$ as follows:

$$R(c) \triangleq \alpha_{+1,-1}\,\pi_{+1}\,\mathbb{P}(c(\boldsymbol{y}) = -1|\boldsymbol{x} = +1) \;+\; \alpha_{-1,+1}\,\pi_{-1}\,\mathbb{P}(c(\boldsymbol{y}) = +1|\boldsymbol{x} = -1)$$

In this expression, the nonnegative scalar $\alpha_{+1,-1}$ weighs the error in assigning an observation from $\boldsymbol{x} = +1$ to $\boldsymbol{x} = -1$; similarly, for $\alpha_{-1,+1}$. Moreover, the scalars $\pi_{\pm 1}$ denote the prior probabilities for $\boldsymbol{x} = \pm 1$.

(a) Follow arguments similar to those in Sec. 28.3 to determine the optimal classifier that minimizes the above weighted risk, $R(c)$. Verify that the expression reduces to (28.28) when the weights $\{\alpha_{+1,-1}, \alpha_{-1,+1}\}$ are equal.

(b) Follow the derivation of the log-likelihood ratio test of Sec. 28.3.2 to show that the optimal classifier admits the following equivalent representation:

$$L(y) \underset{-1}{\overset{+1}{\gtrless}} \left(\frac{\alpha_{-1,+1}}{\alpha_{+1,11}}\right)\left(\frac{\pi_{-1}}{\pi_{+1}}\right)$$

where $L(y)$ is the likelihood ratio defined by (28.42).

28.8 We continue with the setting of Prob. 28.7, except that now we consider situations where it is also important to emphasize correct decisions in addition to erroneous decisions. There are two types of correct decisions: assigning an observation to $\boldsymbol{x} = +1$ when it arises from $\boldsymbol{x} = +1$ or assigning it to $\boldsymbol{x} = -1$ when it arises from $\boldsymbol{x} = -1$. Again, there are situations where one type of correct decisions is more relevant than the other. We can address these scenarios by defining a general weighted risk function as follows:

$$\begin{aligned} R(y) \triangleq\; & \alpha_{+1,-1}\,\pi_{+1}\,\mathbb{P}(c(\boldsymbol{y}) = -1|\boldsymbol{x} = +1) \;+ \\ & \alpha_{-1,+1}\,\pi_{-1}\,\mathbb{P}(c(\boldsymbol{y}) = +1|\boldsymbol{x} = -1) \;+ \\ & \alpha_{+1,+1}\,\pi_{+1}\,\mathbb{P}(c(\boldsymbol{y}) = +1|\boldsymbol{x} = +1) \;+ \\ & \alpha_{-1,-1}\,\pi_{-1}\,\mathbb{P}(c(\boldsymbol{y}) = -1|\boldsymbol{x} = -1) \end{aligned}$$

Given $\alpha_{-1,+1} > \alpha_{-1,-1}$ and $\alpha_{+1,-1} > \alpha_{+1,+1}$, follow the derivation of the log-likelihood ratio test from Sec. 28.3.2 to show that the optimal classifier admits the following equivalent representation:

$$L(y) \underset{-1}{\overset{+1}{\gtrless}} \left(\frac{\alpha_{-1,+1} - \alpha_{-1,-1}}{\alpha_{+1,-1} - \alpha_{+1,+1}}\right)\left(\frac{\pi_{-1}}{\pi_{+1}}\right)$$

where $L(y)$ is the likelihood ratio defined by (28.42).

28.9 Let π_{+1} and π_{-1} denote the prior probabilities for $\boldsymbol{x} = \pm 1$, i.e., $\mathbb{P}(\boldsymbol{x} = +1) = \pi_{+1}$ and $\mathbb{P}(\boldsymbol{x} = -1) = \pi_{-1}$. Introduce the conditional pdfs:

$$f_{+1}(y) \triangleq f_{\boldsymbol{y}|\boldsymbol{x}}(y|\boldsymbol{x} = +1), \quad f_{-1}(y) \triangleq f_{\boldsymbol{y}|\boldsymbol{x}}(y|\boldsymbol{x} = -1)$$

Let $t(y) = \mathbb{P}(\boldsymbol{x} = +1|\boldsymbol{y} = y)$. Show that

$$t(y) = \frac{\pi_{+1} f_{+1}}{\pi_{+1} f_{+1} + \pi_{-1} f_{-1}}$$

Conclude that the test $t(y) > 1/2$ is equivalent to checking for the condition $\pi_{+1} f_{+1}(y) \geq \pi_{-1} f_{-1}(y)$.

28.10 Let $\boldsymbol{y}$ denote a random variable that is distributed according to a Poisson distribution with mean $\lambda \geq 0$, i.e.,

$$\mathbb{P}(\boldsymbol{y} = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

where λ is the average number of events occurring in an interval of time. We model λ as a random variable following an exponential distribution with mean equal to one, i.e., $f_{\boldsymbol{\lambda}}(\lambda) = e^{-\lambda}$ for $\lambda \geq 0$. Assume we collect N independent and identically distributed observations $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$. We would like to estimate $\boldsymbol{\lambda}$ from the observations.

(a) Let $S = \sum_{n=1}^{N} y_n$. Verify that

$$f_{\boldsymbol{\lambda}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}(\lambda|y_1, \ldots, y_N) \propto e^{-\lambda(N+1)} \lambda^S$$

where $\propto$ denotes proportionality. Conclude that the conditional pdf of $\boldsymbol{\lambda}$ given the observations follows a Gamma distribution (which we defined earlier in Prob. 5.2). Determine the mean of this distribution and conclude that the MSE estimate for $\boldsymbol{\lambda}$ is given by

$$\widehat{\lambda}_{\text{MSE}} = \frac{1}{N+1} \left(1 + \sum_{n=1}^{N} y_n\right)$$

(b) Show that the maximum a-posteriori (MAP) estimate for $\boldsymbol{\lambda}$ is given by

$$\widehat{\lambda}_{\text{MAP}} = \frac{1}{N+1} \sum_{n=1}^{N} y_n$$

(c) Show that the mean-absolute error (MAE) estimate for $\boldsymbol{\lambda}$ is found by solving the integral equation

$$\frac{1}{S!} \int_0^{\widehat{\lambda}_{\text{MAE}}} \lambda^S (N+1)^{S+1} e^{-\lambda(N+1)} d\lambda = 1/2$$

(d) Which of the estimators found in parts (a)–(c) is unbiased?

28.11 A random variable $\boldsymbol{y}$ follows a binomial distribution with parameters N and p, i.e., the probability of observing k successes in N trials is given by:

$$\mathbb{P}(\boldsymbol{y} = k) = \binom{N}{k} p^k (1-p)^{N-k}, \quad k = 0, 1, \ldots, N$$

Having observed $\boldsymbol{y} = y$, we wish to estimate the probability of success, p, using a MAP estimator. To do so, and as explained in Example 28.2, we assume that the marginal distribution of $\boldsymbol{p}$ follows the Beta distribution (28.20).

(a) Using the assumed forms for the distributions of $\boldsymbol{p}$ and $\boldsymbol{y}$, determine an expression for the conditional pdf $f_{\boldsymbol{p}|\boldsymbol{y}}(p|y)$.

(b) Show that the peak of $f_{\boldsymbol{p}|\boldsymbol{y}}(p|y)$ occurs at location

$$\widehat{p}_{\text{MAP}} = \frac{y + a - 1}{N + a + b - 2}$$

(c) Compare the MAP solution to the ML solution in future Prob. 31.8.

28.12 Consider the same setting of Prob. 28.11. Show that the MSE estimate of $\boldsymbol{p}$ given $\boldsymbol{y} = y$ (i.e., the conditional mean estimate) is given by:

$$\widehat{p}_{\text{MSE}} = \mathbb{E}\,(\boldsymbol{p}|\boldsymbol{y} = y) = \frac{y + a}{N + a + b}$$

Find the resulting mean-square-error. Compare the MSE solution to the ML solution from future Prob. 31.8.

REFERENCES

Aldrich, J. (1997), "R. A. Fisher and the making of maximum likelihood 1912–1922," *Statistical Science*, vol. 12, no. 3, pp. 162–176.

Bernardo, J. M. and A. F. M. Smith (2000), *Bayesian Theory*, Wiley, UK.

Bishop, C. (2007), *Pattern Recognition and Machine Learning*, Springer, NY.

Bolstad, W. M. (2007), *Bayesian Statistics*, 2nd edition, Wiley, NY.

Box, G. E. P. and G. C. Tiao (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, MA.

Cassella, G. and R. L. Berger (2002), *Statistical Inference*, Duxbury, CA.

Chow, C. K. (1957), "An optimum character recognition system using decision functions," *IRE Trans. Electronic Computers*, vol. 6, pp. 247–254.

Cox, D. R. (2006), *Principles of Statistical Inference*, Cambridge University Press.

Davenport, W. B. and W. L. Root (1958), *The Theory of Random Signals and Noise*, McGraw-Hill, NY.

DeGroot, M. H. (2004), *Optimal Statistical Decisions*, Wiley Classic Library Edition, Wiley, NY.

Duda, R. O., P. E. Hart, and D. G. Stork (2000), *Pattern Classification*, 2nd edition, Wiley, NY.

Fisher, R. A. (1912), "On an absolute criterion for fitting frequency curves," *Messeg. Math.*, vol. 41, pp. 155–160.

Fisher, R. A. (1922), "On the mathematical foundations of theoretical statistics," *Philos. Trans. Roy. Soc. London Ser. A.*, vol. 222, pp. 309–368.

Fisher, R. A. (1925), "Theory of statistical estimation," *Proc. Cambridge Philos. Soc.*, vol. 22, pp. 700–725.

Gupta A. K. and S. Nadarajah (2004) *Handbook of Beta Distribution and its Applications*, Marcel Dekker, NY.

Hahn, G. J. and S. S. Shapiro (1994), *Statistical Models in Engineering*, Wiley, NY.

Hoff, P. D. (2009), *A First Course in Bayesian Statistical Methods,* Springer, NY.

Hogg, R. V. and J. McKean (2012), *Introduction to Mathematical Statistics*, 7th edition, Pearson.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1994), *Continuous Univariate Distributions*, vol. 1, Wiley, NY.

Kay, S. (1993), *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, NJ.

Kay, S. (1998), *Fundamentals of Statistical Signal Processing: Detection Theory*, Prentice Hall, NJ.

Lee, P. M. (2002), *An Introduction to Bayesian Statistics,* 4th edition, Wiley, NY.

Lehmann, E. L. and J. P. Romano (2005), *Testing Statistical Hypothesis,* 3rd edition, Springer, NY.

Levy, B. C. (2008), *Principles of Signal Detection and Parameter Estimation*, Springer, NY.

Middleton, D. (1960), *An Introduction to Statistical Communication Theory*, McGraw-Hill, NY.

Miller, R. G. (1962), "Statistical prediction by discriminant analysis," *Meteorological Monographs*, vol. 4, no. 25, pp. 1–54.

Neyman, J. and E. Pearson (1928), "On the use and interpretation of certain test criteria for purposes of statistical inference — Part I," *Biometrika*, vol. 20A, no.1/2, pp.175–240.

Neyman, J. and E. Pearson (1933), "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London*, vol. 231, pp. 289-337.
Poor, H. V. (1998), *An Introduction to Signal Detection and Estimation*, Springer, NY.
Pratt, J. W. (1976), "F. Y. Edgeworth and R. A. Fisher on the efficiency of maximum likelihood estimation," *Annals of Statistics*, vol. 4, no. 3, pp. 501–514.
Robert, C. P. (2007), *The Bayesian Choice,* 2nd edition, Springer, NY.
Savage, L. J. (1954), *The Foundations of Statistics*, Wiley, NY.
Scharf, L. L. (1991), *Statistical Signal Processing: Detection, Estimation, and Time-Series Analysis*, Addison-Wesley, Reading, MA.
Theodoridis, S. and K. Koutroumbas (2008), *Pattern Recognition*, 4th edition, Academic Press.
Van Trees, H. L. (1968), *Detection, Estimation, and Modulation Theory*, Wiley, NY.
Van Trees, H. L. (2013), *Detection, Estimation, and Modulation Theory*, Part I, 2nd edition, Wiley, NY.
Wald, A. (1950), *Statistical Decision Functions*, Wiley, NY.
Webb, A. (2002), *Statistical Pattern Recognition*, Wiley, NY.
Young, G. A. and R. L. Smith (2010), *Essentials of Statistical Inference*, Cambridge University Press.
Zacks, S. (1971), *The Theory of Statistical Inference*, Wiley, NY.