Reading Notes of Pattern Classification and Machine Learning

Tianyi Cui

August 15, 2012

1 Introduction

Different kinds of tasks of machine learning:

- supervised learning: known input and target vectors
- classification: output is one of a finite number of discrete categories
  - regression: output is one or more continuous variables
- unsupervised learning: no corresponding target values
  - clustering: discover groups of similar examples within the data
  - density estimation: determine the distribution of data within the input space
  - dimension reduction
- reinforcement learning: finding suitable actions to take in a given situation in order to maximize a reward

1.1 Example: Polynomial Curve Fitting

In regression problems, we can use a polynomial function

$$y(x, \mathbf{w}) = w_0 + w_1 + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j \tag{1.1}$$

to fit the underlying function.

We need to minimize the *error function*

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 \tag{1.2}$$

in which unique solution $\mathbf{w}^*$ can be found in closed form.

The root-mean-square (RMS) error is defined by

$$E_{RMS} = \sqrt{2E(\mathbf{w}^*)/\mathbf{N}} \tag{1.3}$$

When M is large, *over-fitting* occurs, i.e. E_{RMS} against test data becomes large. One technique to control over-fitting is *regularization*, by adding a penalty term to the error function (1.2) in order to discourage the coefficients from reaching large values:

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2 \tag{1.4}$$

1.2 Probability Theory

Equations for probability:

- Sum rule
$$p(X) = \sum_Y p(X, Y) \tag{1.5}$$
- Product rule
$$p(X, Y) = p(Y|X)p(X) \tag{1.6}$$
- Bayes' theorem
$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \tag{1.7}$$

The denominator in (1.7) can be expressed in terms of the quantities appearing in the numerator
$$p(X) = \sum_Y p(X|Y)p(Y) \tag{1.8}$$

We can view the denominator in Bayes' theorem as being the normalization constant required to ensure that the sum of the conditional probability on the left-hand side of (1.7) over all values of Y equals 1.

Before any observation, we have a probability of a certain event Y, this is called *prior probability* $p(Y)$, after some observation X, the probability of event Y becomes the *posterior probability* $p(Y|X)$.

X and Y are said to be *independent* if $p(X, Y) = p(X)p(Y)$, which is equivalent to $P(Y|X) = p(Y)$.

1.2.1 Probability densities

If the probability that x will lie in (a, b) is given by
$$p(x \in (a, b)) = \int_a^b p(x)\mathrm{d}x \tag{1.9}$$
then $p(x)$ is called the *probability density* over x.

Apparently $p(x) \geq 0$ and $\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$.

Under a nonlinear change of variable, a probability density transforms differently from a simple function, due to the Jacobian factor. If $x = g(y)$, since $p_x(x)\mathrm{d}x = p_y(y)\mathrm{d}y$, hence
$$\begin{aligned} p_y(y) &= p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| \\ &= p_x(g(y))\left|g'(y)\right| \end{aligned} \tag{1.10}$$

The *cumulative distribution function*
$$P(z) = \int_{-\infty}^{z} p(x)\mathrm{d}x \tag{1.11}$$

For several continuous variables $x_1, \ldots, x_D$, denoted collectively by the vector $\mathbf{x}$, then we can define a joint probability density $p(\mathbf{x})$ such that $p\left(\mathbf{x} \in (\mathbf{x}_0, \mathbf{x}_0 + \delta\mathbf{x})\right) = p(\mathbf{x}_0)\delta\mathbf{x}$.

1.2.2 Expectations and covariances

The average value of some function $f(x)$ under a probability distribution $p(x)$ is called the *expectation* of $f(x)$ and denoted by $\mathbb{E}[f]$. For a discrete distribution,

$$\mathbb{E}[f] = \sum_x p(x)f(x) \tag{1.12}$$

For continuous variables,

$$\mathbb{E}[f] = \int p(x)f(x)\mathrm{d}x \tag{1.13}$$

In either case, the expectation can be approximated given N samples,

$$\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n) \tag{1.14}$$

When considering expectations of functions of several variables, we use subscript to indicate which variable is being averaged over, e.g. $\mathbb{E}_x[f(x,y)]$ is a function of y.

We can also consider *conditional expectation*

$$\mathbb{E}_x[f|y] = \sum_x p(x|y)f(x) \tag{1.15}$$

The *variance* of $f(x)$ is defined by

$$\begin{aligned} \mathrm{var}[f] &= \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)]\right]^2 & (1.16)\\ &= \mathbb{E}\left[f(x)^2\right] - \mathbb{E}\left[f(x)\right]^2 & (1.17) \end{aligned}$$

The *covariance* of two random variable x and y is defined by

$$\begin{aligned} \mathrm{cov}[x,y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] & (1.18) \end{aligned}$$

which expresses the extent to which x and y vary together. If they are independent, then the covariance vanishes.

In the case of two vectors of random variables $\mathbf{x}$ and $\mathbf{y}$, the covariance is a matrix

$$\begin{aligned} \mathrm{cov}[\mathbf{x},\mathbf{y}] &= \mathbb{E}_{x,y}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\left\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\right\}\right] \\ &= \mathbb{E}_{x,y}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}] & (1.19) \end{aligned}$$

1.2.3 Bayesian probabilities

In the *classical* or *frequentist* interpretation of probability, probabilities is viewed in terms of the frequencies of random, repeatable events. In the more general *Beyesian* view, probabilities provide a quantification of uncertainty, so we can say the probability of an uncertain event, like whether the Arctic ice cap will have disappeared by the end of the century, which is not events that can be repeated.

In the polynomial curve fitting example, we assume the parameters $\mathbf{w}$ have a prior probability distribution $p(\mathbf{w})$, then given the observed data $\mathcal{D}$, the posterior probability is

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \tag{1.20}$$

where the quantity $p(\mathcal{D}|\mathbf{w})$ is called the *likelihood function*, which expresses how probable the observed data set is for different settings of the parameter vector $\mathbf{w}$. The likelihood is not a probability distribution over $\mathbf{w}$, and its integral does not necessarily equal one.

Given the definition of likelihood, we can state Bayes' theorem in words

$$\text{posterior} \propto \text{likelihood} \times \text{prior} \tag{1.21}$$

where all of these quantities are viewed as functions of $\mathbf{w}$.

In the likelihood function $p(\mathcal{D}|\mathbf{w})$, in the frequentist setting, $\mathbf{w}$ is considered to be a fixed parameter, whose value is determines by some form of 'estimator', and error bars on this estimate are obtained by considering the distribution of possible data sets $\mathcal{D}$. By contrast, from Bayesian viewpoint there is only a single data set $\mathcal{D}$ (the one actually observed), and the uncertainty in the parameters is expressed through a probability distribution over $\mathbf{w}$.

1.2.4 The Gaussian distribution

The Gaussian distribution on a single real-valued variable x is defined by

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{1.22}$$

which is governed by two parameters: μ the *mean* and σ^2 the *variance*. σ is called the *standard deviation*, and $\beta = 1/\sigma^2$ is called the *precision*. The mean of x is given by $\mathbb{E}[x] = \mu$ and the variance of x is given by $\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$.

The Gaussian distribution defined over a D-dimensional vector $\mathbf{x}$ of continuous variables is given by

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{1.23}$$

Suppose we have a data set of observation $\mathbf{x} = (x_1, \ldots, x_N)^{\mathrm{T}}$ which is *independent and identically distributed* (often abbreviated to i.i.d.) from a Gaussian distribution. The

likelihood of the data set, which is a function of μ and σ^2, is in the form

$$p\left(\mathbf{x}|\mu,\sigma^2\right) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2) \tag{1.24}$$

One common criterion for determining the parameters in a probability distribution using an observed data set is to find the parameter values that maximize the likelihood function.

In practice, for mathematical and numerical reasons, it's more convenient to maximize the log of the likelihood functions

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi) \tag{1.25}$$

Maximizing (1.25) with respect to μ gives the maximum likelihood solution

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{1.26}$$

which is the *sample mean*. Similarly, Maximize (1.25) with respect to σ^2gives

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2 \tag{1.27}$$

which is the *sample variance.*

The maximum likelihood approach systematically underestimates the variance of the distribution. This is an example of a phenomenon called *bias* and is related to the problem of over-fitting encountered in the context of polynomial curve fitting. First, we note that μ_{ML} and σ^2_{ML} are functions of the data set values $x_1,\ldots,x_N$. Consider the expectations of these quantities with respect to the data set values, which themselves come from a Gaussian distribution with parameters μ and σ^2

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu \tag{1.28}$$

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2 \tag{1.29}$$

so on average the maximum likelihood approach will underestimate the true variance by a factor $(N-1)/N$.

From (1.29) we see the following estimate for the variance parameter is unbiased

$$\widetilde{\sigma}^2 = \frac{N}{N-1}\sigma^2_{\mathrm{ML}} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2 \tag{1.30}$$

this result arises automatically when we adopt a Bayesian approach (Section 10.1.3).

1.2.5 Curve fitting re-visited

The goal of curve fitting problem is to make predictions for the target variable t given some new value of the input variable x on the basis of a set of training data $\mathbf{x} = (x_1, \ldots, x_N)^{\mathrm{T}}$ and $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$. We can express out uncertainty over the value of the target variable using a probability distribution. Assume that, given the value of x, the corresponding value of t has a Gaussian distribution with a mean equal to the value $y(x, \mathbf{w})$ given by (1.1). Thus we have

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right) \tag{1.31}$$

where β is the precision parameter.

Use the training data $\{\mathbf{x}, \mathbf{t}\}$ to determine the values of the unknown parameters $\mathbf{w}$ and β by maximum likelihood, the likelihood function is

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n|y(x_n, \mathbf{w}), \beta^{-1}\right) \tag{1.32}$$

and its logarithm is

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) \tag{1.33}$$

Maximizing (1.33) with respect to $\mathbf{w}$ gives us $\mathbf{w}_{\mathrm{ML}}$, which is the same as minimize the *sum-of-squares error function* defined by (1.2).

Maximizing (1.33) with respect to β gives

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{n=1}^{N} \{y(x_n, \mathrm{w}_{\mathrm{ML}}) - t_n\}^2 \tag{1.34}$$

Having determined the parameters $\mathbf{w}$ and β, we can now make predictions for new values of x, and in probabilistic model, these are expressed in terms of the *predictive distribution* that gives the probability distribution over t

$$p(t|x, \mathrm{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{ML}), \beta_{\mathrm{ML}}^{-1}\right) \tag{1.35}$$

In a more Bayesian approach, we introduce a Gaussian prior distribution over the polynomial coefficients $\mathbf{w}$

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\} \tag{1.36}$$

where α is the precision of the distribution and $M + 1$ is the number of elements in $\mathbf{w}$. Values such as α, which controls the distribution of model parameters, are called *hyperparameters*.

Using Bayes' theorem

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}\,, \alpha, \beta) \propto p(\mathbf{t}\,|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\alpha) \tag{1.37}$$

We can now determine $\mathbf{w}$ by finding the most probable value of $\mathbf{w}$ given the data, in other words by maximizing the posterior distribution. This technique is called *maximum posterior*, or simply *MAP*.

Taking the negative logarithm of (1.37) and combining with (1.33) and (1.36), we find that the maximum of the posterior is given by the minimum of

$$\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2} \mathbf{w}^{\mathrm{T}} \mathbf{w} \tag{1.38}$$

Thus we see that maximizing the posterior distribution is equivalent to minimizing the regularized sum-of-squares error function (1.4).

1.2.6 Bayesian curve fitting

Although we have included a prior distribution $p(\mathbf{w}|\alpha)$, we are still making a point estimate of $\mathbf{w}$ and so this does not yet amount to a Bayesian treatment. In a fully Bayesian approach, we should consistently apply the sum and product rules of probability, which requires, as we shall see shortly, that we integrate over all values of $\mathbf{w}$. Such marginalizations lie at the heart of Bayesian methods for pattern recognition.

In curve fitting, we are given the training data $\{\mathbf{x}, \mathbf{t}\}$, along with a new test point x, and our goal is to predict the value of t. Assuming the parameters α and β are fixed and known in advance by now, we wish the evaluate the predictive distribution $p(t|\mathbf{x}, \mathbf{t})$. Using the product rules of probability

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{t}) \mathrm{d}\mathbf{w} \tag{1.39}$$

Here $p(t|x, \mathbf{w})$ and $p(\mathbf{w}|\mathbf{x}, \mathbf{t})$ are given by (1.31) and normalizing the right-hand side of (1.37).

The calculation and integration in (1.39) can be performed analytically with the result in a Gaussian distribution

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}\left(t|m(x), s^2(x)\right) \tag{1.40}$$

where the mean and variance are given by

$$m(x) = \beta \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n \tag{1.41}$$

$$s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \boldsymbol{\phi}(x) \tag{1.42}$$

Here the matrix $\mathbf{S}$ is given by

$$\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x)^{\mathrm{T}} \tag{1.43}$$

and we have defined the vector $\boldsymbol{\phi}(x)$ with elements $\phi_i(x) = x^i$ for $i = 0, \ldots, M$.

The matrix and the mean of the predictive distribution in (1.40) is dependent on x. The first term in (1.42) represents the uncertainty due to the noise on the target variables, and the second term arises from the uncertainty in the parameters $\mathbf{w}$ and is a consequence of the Bayesian treatment.

1.3 Model Selection

Model selection is to find the appropriate values of complexity parameters within a given model and to find the best model for a particular application.

Due to the problem of over-fitting, performance on the training set is not a good indicator of predictive performance. If data is plentiful, we can set aside a *validation set* for comparing models. If the model design is iterated many times using a limited size data set, some over-fitting to the validation data can occur so it may be necessary to keep aside a third *test set* on which the performance of the selected model is finally evaluated.

But the supply of data for training and testing will be limited. To use as much of the available data as possible for training, one solution is to use *cross-validation*, which is, to divide the data into S sets, and use $S - 1$ sets for training and 1 set for validation, in total S runs. When $S = N$, it's called the *leave-one-out* technique.

One major drawback of cross-validation is that the number of training runs is increased by a factor of S, and this can be problematic when training is computationally expensive. And when there are multiple parameters to explore, required number of training runs is exponential in the number of parameters. We therefore need a measure of performance which depends only on the training data (i.e. not validation-based) and which does not suffer from bias due to over-fitting.

Historically various 'information criteria' have been proposed that attempt to correct for the bias of maximum likelihood by the addition of a penalty term to compensate for the over-fitting of more complex models. For example, the *Akaike information criterion*, or AIC, chooses the model for which the quantity

$$\ln p(\mathcal{D}|\mathbf{w}_{\mathrm{ML}}) - M \tag{1.44}$$

is largest. Later we'll see how complexity penalties arise in a natural and principled way in a fully Bayesian approach.

1.4 The Curse of Dimensionality

In the polynomial curve fitting example we had just one input variable x, but in practice we will deal with spaces of high dimensionality comprising many input variables. This poses some serious challenges and is an important factor influencing the design of pattern recognition techniques.

For example, a simple approach for classification is to divide the input space into regular cells and classify each cell independently. But the number of cells grows exponentially

with the dimensionality of the space, so we need exponentially large quantity of training data in order to ensure that the cells are not empty, which is not practical in a space of more than a few variables. High-dimensional general polynomial curve fitting have similar problems, as D the number of input variables increases, the number of independent coefficients grows proportionally to D^M for a polynomial of order M.

Our geometrical intuitions formed from life can fail badly when we consider spaces of higher dimensionality. For example, consider a sphere of radius $r = 1$ in a space of D dimensions, the fraction of the volume of the sphere that lies between radius $r = 1 - \epsilon$ and $r = 1$ is given by $1 - (1 - \epsilon)^D$. For large D, this fraction tends to 1 even for small values of ϵ. Thus, in spaces of high dimensionality, most of the volume of a sphere is concentrated in a thin shell near the surface!

Similarly, consider Gaussian distribution in high-dimensional space. If we transform from Cartesian to polar coordinates, and then integrate out the directional variables, we obtain an expression for the density $p(r)$ as a function of radius r from the origin. We can see that for large D the probability mass of the Gaussian is concentrated in a thin shell.

The severe difficulty that can arise in spaces of many dimensions is sometimes called the *curse of dimensionality*. But it does not prevent us from finding effective techniques applicable to high-dimensional spaces. First, real data will often be confined to a region of the space having lower effective dimensionality, and in particular the directions over which important variations in the target variables occur may be so confined. Second, real data will typically exhibit some smoothness properties (at least locally) so that for the most part small changes in the input variables will produce small changes in the target variables, and so we can exploit local interpolation-like techniques to allow us to make predictions of the target variables for new values of the input variables. Successful pattern recognition techniques exploit one or both of these properties. For example, an application in manufacturing in which images are captured of identical planar objects on a conveyor belt, in which the goal is to determine their orientation. Each image is a point in a space whose dimensionality is determined by the number of pixels. But since there are three degrees of freedom of variability between images, actually a set of images will live on a three dimensional *manifold* embedded within the high-dimensional space.

1.5 Decision Theory

Decision theory, when combined with probability theory, allows us to make optimal decisions in situations involving uncertainty.

Suppose we have an input vector $\mathbf{x}$ together with a corresponding vector $\mathbf{t}$ of target variables, and our goal is to predict $\mathbf{t}$ given a new value for $\mathbf{x}$. $\mathbf{t}$ are continuous variables or class labels for regression and classification problems. The joint probability distribution $p(\mathbf{x}, \mathbf{t})$ provides a complete summary of the uncertainty associated with these variables. Determination of $p(\mathbf{x}, \mathbf{t})$ from a set of training data is an example of *inference* and is typically very difficult. In practice, what we need is the prediction of $\mathbf{t}$, or more generally take a specific action based on our understudying of values $\mathbf{t}$ is likely to take, and this

aspect is the subject of decision theory.

Consider, for example, a medical diagnosis problem, we have a X-ray image input vector $\mathbf{x}$, and output value t to be a binary variable such that $t = 0$ corresponds to class $\mathcal{C}_1$, the presence of cancer, and $t = 1$ corresponds to $\mathcal{C}_2$. The general inference problem involves determining the joint distribution $p(\mathbf{x}, \mathcal{C}_k)$, or equivalently $p(\mathbf{x}, t)$. Although this can be very useful and informative, in the end we must decide whether to give treatment, and we would like this choice to be optimal in some appropriate sense. This is the *decision* step.

When we obtained $\mathbf{x}$, we're interested in the probabilities of the two classes given the image, which are given by $p(\mathcal{C}_k|\mathbf{x})$, using Bayes' theorem, it can be expressed in the form

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} \tag{1.45}$$

If our aim is to minimize the chance of assigning $\mathbf{x}$ to the wrong class, then intuitively we would choose the class having the higher posterior probability.

1.5.1 Minimizing the misclassification rate

We need a rule to assign each value of $\mathbf{x}$ to one of the available classes. Such a rule will divide the input space into regions $\mathcal{R}_k$ called *decision regions*, one for each class. The boundaries between decision regions are called *decision boundaries* or *decision surfaces.*

In the case of two classes, the probability of misclassification is

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\mathrm{d}\mathbf{x} \end{aligned} \tag{1.46}$$

Clearly to minimize $p(\text{mistake})$ we should arrange that each $\mathbf{x}$ is assigned to whichever class has the smaller value of the integrand in (1.46). Since $p(\mathbf{x}, \mathcal{C}_k) = p(\mathcal{C}_k|\mathbf{x})p(\mathbf{x})$, it's equivalent to assign $\mathbf{x}$ to the class for which the posterior probability $p(\mathcal{C}_k|\mathbf{x})$ is largest.

For the more general case of K classes, it's slightly easier to maximize the probability of being correct, which is given by

$$\begin{aligned} p(\text{correct}) &= \sum_{k=1}^{K} p(\mathbf{x} \in \mathcal{R}_k, \mathcal{C}_k) \\ &= \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathbf{x}, \mathcal{C}_k)\mathrm{d}\mathbf{x} \end{aligned} \tag{1.47}$$

which is maximized when the regions $\mathcal{R}_k$ are chosen such that each x is assigned to the class for which $p(\mathbf{x}, \mathcal{C}_k)$ or $p(\mathcal{C}_k|\mathbf{x})$ is the largest.

1.5.2 Minimizing the expected loss

For many applications, different kinds of misclassifications lead to different penalty, which can be formalized through a *loss function*, also called a *cost function*, which is a

single, overall measure of loss incurred in taking any of the available decisions or actions. Our goal is then to minimize the total loss incurred. Suppose L_{kj} represents the loss when the true class is $\mathcal{C}_k$ and we assign the input to class $\mathcal{C}_j$, L is called a *loss matrix*.

The optimal solution is the one which minimizes the loss function. However, the loss function depends on the true class, which is unknown. So we seek instead of minimize the average loss respect to the distribution $p(\mathbf{x}, \mathcal{C}_k)$, which is given by

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) \mathrm{d}\mathbf{x} \tag{1.48}$$

Each $\mathbf{x}$ can be assigned to one of $\mathcal{R}_j$, which implies that for each $\mathbf{x}$ we should minimize $\sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k)$. As before we can use the product rule $p(\mathbf{x}, \mathcal{C}_k) = p(\mathcal{C}_k|\mathbf{x})p(\mathbf{x})$ to eliminate the common factor of $p(\mathbf{x})$. Thus the decision rule that minimizes the expected loss is the one that assigns each new $\mathbf{x}$ to the class j for which the quantity

$$\sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k) \tag{1.49}$$

is a minimum.

1.5.3 The reject option

The classification errors arise from the regions of input space where the largest of the posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$ is significantly less than unity, or equivalently where the joint distributions $p(\mathbf{x}, \mathcal{C}_k)$ have comparable values. These are the regions where we are relatively uncertain about class membership. In some applications, it will be appropriate to avoid making decisions on the difficult cases in anticipation of a lower error rate on those examples for which a classification decision is made. This is known as the reject option. We can achieve this by introducing a threshold θ and rejecting those inputs $\mathbf{x}$ for which the largest of the posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$ is less than or equal to θ.

We can easily extend the reject criterion to minimize the expected loss, when a loss matrix include the loss incurred when a reject decision is made.

1.5.4 Inference and decision

We have broken the classification problem down into two separate stages, the *inference stage* in which we use training data to learn a model for $p(\mathcal{C}_k|\mathbf{x})$, and the subsequent *decision stage* in which we use these posterior probabilities to make optimal class assignments. In fact, we can identity three distinct approaches to solving decision problems.

(a) First solve the inference problem of determining the class-conditional densities $p(\mathbf{x}|\mathcal{C}_k)$. Also separately infer the prior class probabilities $p(\mathcal{C}_k)$. Then use Bayes' theorem to find the posterior class probabilities $p(\mathcal{C}_k|\mathbf{x})$. Equivalently, we can model the joint distribution $p(\mathbf{x}, \mathcal{C}_k)$ directly and then normalize to obtain the posterior probabilities. Then we use decision theory to determine class membership. Approaches that explicitly or implicitly model the distribution of inputs as well

as outputs are known as *generative models*, because by sampling from them it is possible to generate synthetic data points in the data space.

(b) First solve the inference problem of determining the posterior class probabilities $p(\mathcal{C}_k|\mathbf{x})$, and then use decision theory to assign each new $\mathbf{x}$ to one of the classes. Approaches that model the posterior probabilities directly are called *discriminative models*.

(c) Find a function $f(\mathbf{x})$, called a *discriminant function*, which maps each input $\mathbf{x}$ directly onto a class label. In this case, probabilities play no role.

Approach (a) is the most demanding, because for many applications $\mathbf{x}$ will have high dimensionality, and consequently we need a large training set in order to determine the class-conditional densities $p(\mathbf{x}|\mathcal{C}_k)$ or the joint distribution $p(\mathbf{x}, \mathcal{C}_k)$ to reasonable accuracy. However, one advantage is it can also determine $p(\mathbf{x})$. This can be useful for detecting new data points that have low probability under the model and for which the predictions may be of low accuracy, which is known as *outlier detection* or *novelty detection*.

The class-conditional densities may contain a lot of structure that has little effect on the posterior probabilities, so in approach (b) we find the posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$ directly.

Approach (c) is even simpler, in which we combine the inference and decision stages into a simple learning problem.

There are many powerful reasons for wanting to compute the posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$ before making decisions:

- The loss matrix may be subjected to revision.
- The possibility of reject option.
- Compensating for class priors. Consider the medical X-ray problem, since cancer is rare, only 0.1% of our samples are in the cancer class. A classifier that assigned every point to the normal class would already achieve 99.9% accuracy and it would be difficult to avoid this trivial solution. Also, the learning algorithm will not be exposed to a broad range of examples in the cancer class and hence is not likely to generalize well. A balanced data set in which we have selected equal numbers of examples from each of the classes would allow us to find a more accurate model. However, we must compensate for the effects of our modifications to the training data. We can simply take the posterior probabilities obtained from our artificially balanced data set and first divide by the class fractions in that data set and then multiply by the class fractions in the population to which we wish to apply the model. Finally, we need to normalize to ensure that the new posterior probabilities sum to one. Note that this procedure cannot be applied if we have learned a discriminant function directly instead of determining posterior probabilities.
- Combining models. For complex applications, we can break the problem into a number of smaller subproblems each of which can be tackled by a separate model.

For example in the medical X-ray problem, we may assume that the distribution of inputs for X-ray images $\mathbf{x}_\mathrm{I}$ and the blood data $\mathbf{x}_\mathrm{B}$ are independently, so that

$$p(\mathbf{x}_\mathrm{I}, \mathbf{x}_\mathrm{B}|\mathcal{C}_k) = p(\mathbf{x}_\mathrm{I}|\mathcal{C}_k)p(\mathbf{x}_\mathrm{B}|\mathcal{C}_k) \tag{1.50}$$

This is an example of *conditional independence* property. Then the posterior probability given both the data is given by

$$\begin{aligned} p(\mathcal{C}_k|\mathbf{x}_\mathrm{I}, \mathbf{x}_\mathrm{B}) &\propto p(\mathbf{x}_\mathrm{I}, \mathbf{x}_\mathrm{B}|\mathcal{C}_k)p(\mathcal{C}_k) \\ &\propto p(\mathbf{x}_\mathrm{I}|\mathcal{C}_k)p(\mathbf{x}_\mathrm{B}|\mathcal{C}_k)p(\mathcal{C}_k) \\ &\propto p(\mathcal{C}_k|\mathbf{x}_\mathrm{I})p(\mathcal{C}_k|\mathbf{x}_\mathrm{B}) \end{aligned} \tag{1.51}$$

1.5.5 Loss functions for regression

In regression problems, the decision stage consists of choosing a specific estimate $y(\mathbf{x})$ of the value of t for each input $\mathbf{x}$. Suppose that in doing so, we incur a loss $L(t, y(\mathbf{x}))$. The expected loss is given by

$$\mathbb{E}[L] = \iint L(t, y(\mathbf{x}))p(\mathbf{x}, t)\mathrm{d}\mathbf{x}\mathrm{d}t \tag{1.52}$$

A common choice of the loss function is the squared loss $L(t, y(\mathbf{x})) = \{y(\mathbf{x}) - t\}^2$. In this case

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\mathrm{d}\mathbf{x}\mathrm{d}t \tag{1.53}$$

Our goal is to choose $y(\mathbf{x})$ so as to minimize $\mathbb{E}[L]$. If we assume a completely flexible function $y(\mathbf{x})$, we can do this formally using the calculus of variations to give

$$\frac{\delta\mathbb{E}[L]}{\delta y(\mathbf{x})} = 2\int \{y(\mathbf{x}) - t\} p(\mathbf{x}, t)\mathrm{d}t = 0 \tag{1.54}$$

Solving for $y(\mathbf{x})$, and using the sum and product rules of probability, we obtain

$$y(\mathbf{x}) = \frac{\int tp(\mathbf{x}, t)\mathrm{d}t}{p(\mathbf{x})} = \int tp(t|\mathbf{x})\mathrm{d}t = \mathbb{E}_t[t|\mathbf{x}] \tag{1.55}$$

which is the conditional average of t conditioned on $\mathbf{x}$ and is known as the *regression function*. It can readily be extended to multiple variables represented by the vector $\mathbf{t}$, in which case the optimal solution is the conditional average $\mathbf{y}(\mathbf{x}) = \mathbb{E}_\mathbf{t}[\mathbf{t}|\mathbf{x}]$.

The squared loss is not the only possible choice of loss function for regression. Indeed, there are situations in which squared loss can lead to very poor results and where we need to develop more sophisticated approaches. An important example concerns situations in which the conditional distribution $p(t|\mathbf{x})$ is multimodal, as often arises in the solution of inverse problems. One simple generalization of the squared loss is the *Minkowski* loss, whose expectation is given by

$$\mathbb{E}[L_q] = \iint |y(\mathbf{x}) - t|^q p(\mathbf{x}, t)\mathrm{d}\mathbf{x}\mathrm{d}t \tag{1.56}$$

The minimum of $\mathbb{E}[L_q]$ is given by the conditional mean for $q = 2$, the conditional media for $q = 1$, and the conditional mode for $q \to 0$.