Reading Notes of Pattern Classification and Machine Learning

Tianyi Cui

August 18, 2012

Contents

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

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

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 $p(x) \geqslant 0$ 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

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

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

$$\begin{aligned} \mathrm{cov}[x,y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned} \tag{1.17}$$

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

$$\begin{aligned} \mathrm{cov}[\mathbf{x},\mathbf{y}] &= \mathbb{E}_{x,y}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\left\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\right\}\right] \\ &= \mathbb{E}_{x,y}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}] \end{aligned} \tag{1.18}$$

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

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \tag{1.19}$$

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

$$\text{posterior} \propto \text{likelihood} \times \text{prior} \tag{1.20}$$

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

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{1.21}$$

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

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{1.22}$$

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

$$p\left(\mathbf{x}|\mu,\sigma^2\right) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2) \tag{1.23}$$

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

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi) \tag{1.24}$$

Maximizing (1.24) with respect to μ gives the maximum likelihood solution

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{1.25}$$

which is the *sample mean*. Similarly, Maximize (1.24) with respect to σ^2gives

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2 \tag{1.26}$$

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

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu \tag{1.27}$$

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2 \tag{1.28}$$

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

From (1.28) we see the following estimate for the variance parameter is unbiased

$$\widetilde{\sigma}^2 = \frac{N}{N-1}\sigma^2_{\mathrm{ML}} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2 \tag{1.29}$$

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

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right) \tag{1.30}$$

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

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

and its logarithm is

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) \tag{1.32}$$

Maximizing (1.32) 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 (1.32) with respect to β gives

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n, \mathrm{w}_{\mathrm{ML}}) - t_n\}^2 \tag{1.33}$$

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

$$p(t|x, \mathrm{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{ML}), \beta_{\mathrm{ML}}^{-1}\right) \tag{1.34}$$

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

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2}\exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\} \tag{1.35}$$

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

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha) \tag{1.36}$$

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 (1.36) and combining with (1.32) and (1.35), we find that the maximum of the posterior is given by the minimum of

$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \tag{1.37}$$

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

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t})\mathrm{d}\mathbf{w} \tag{1.38}$$

Here $p(t|x, \mathbf{w})$ and $p(\mathbf{w}|\mathbf{x}, \mathbf{t})$ are given by (1.30) and normalizing the right-hand side of (1.36).

The calculation and integration in (1.38) can be performed analytically with the result in a Gaussian distribution

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}\left(t|m(x), s^2(x)\right) \tag{1.39}$$

where the mean and variance are given by

$$m(x) = \beta\boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)t_n \tag{1.40}$$

$$s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\boldsymbol{\phi}(x) \tag{1.41}$$

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

$$\mathbf{S}^{-1} = \alpha\mathbf{I} + \beta\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x)^{\mathrm{T}} \tag{1.42}$$

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 (1.39) is dependent on x. The first term in (1.41) 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

$$\ln p(\mathcal{D}|\mathbf{w}_{\mathrm{ML}}) - M \tag{1.43}$$

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

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} \tag{1.44}$$

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

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\mathrm{d}\mathbf{x} \end{aligned} \tag{1.45}$$

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 (1.45). 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

$$\begin{aligned} p(\text{correct}) &= \sum_{k=1}^{K} p(\mathbf{x} \in \mathcal{R}_k, \mathcal{C}_k) \\ &= \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathbf{x}, \mathcal{C}_k)\mathrm{d}\mathbf{x} \end{aligned} \tag{1.46}$$

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

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) \mathrm{d}\mathbf{x} \tag{1.47}$$

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

$$\sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k) \tag{1.48}$$

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

$$p(\mathbf{x}_\mathrm{I}, \mathbf{x}_\mathrm{B}|\mathcal{C}_k) = p(\mathbf{x}_\mathrm{I}|\mathcal{C}_k)p(\mathbf{x}_\mathrm{B}|\mathcal{C}_k) \tag{1.49}$$

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

$$\begin{aligned} p(\mathcal{C}_k|\mathbf{x}_\mathrm{I}, \mathbf{x}_\mathrm{B}) &\propto p(\mathbf{x}_\mathrm{I}, \mathbf{x}_\mathrm{B}|\mathcal{C}_k)p(\mathcal{C}_k) \\ &\propto p(\mathbf{x}_\mathrm{I}|\mathcal{C}_k)p(\mathbf{x}_\mathrm{B}|\mathcal{C}_k)p(\mathcal{C}_k) \\ &\propto p(\mathcal{C}_k|\mathbf{x}_\mathrm{I})p(\mathcal{C}_k|\mathbf{x}_\mathrm{B}) \end{aligned} \tag{1.50}$$

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

$$\mathbb{E}[L] = \iint L(t, y(\mathbf{x}))p(\mathbf{x}, t)\mathrm{d}\mathbf{x}\mathrm{d}t \tag{1.51}$$

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

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\mathrm{d}\mathbf{x}\mathrm{d}t \tag{1.52}$$

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

$$\frac{\delta\mathbb{E}[L]}{\delta y(\mathbf{x})} = 2\int \{y(\mathbf{x}) - t\}\, p(\mathbf{x}, t)\mathrm{d}t = 0 \tag{1.53}$$

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

$$y(\mathbf{x}) = \frac{\int t\, p(\mathbf{x}, t)\mathrm{d}t}{p(\mathbf{x})} = \int t\, p(t|\mathbf{x})\mathrm{d}t = \mathbb{E}_t[t|\mathbf{x}] \tag{1.54}$$

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

$$\mathbb{E}[L_q] = \iint |y(\mathbf{x}) - t|^q\, p(\mathbf{x}, t)\mathrm{d}\mathbf{x}\mathrm{d}t \tag{1.55}$$

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 $q \to 0$.

1.6 Information Theory

Consider a discrete random variable x and we ask how much information is received when we observe a specific value for this variable. The amount of information can be viewed as the 'degree of surprise' on learning the value of x therefore will depend on $p(x)$. We should look for a quantity $h(x)$ that is a monotonic function of the probability $p(x)$ and that expresses the information content. The form of $h(\cdot)$ can be found by noting that if we have two events x and y that are unrelated, then the information gain from observing both of them should be the sum of the information gained from each of them separately, so that $h(x, y) = h(x) + h(y)$. Two unrelated events will be statistically independent and so $p(x, y) = p(x)p(y)$. From these two relationships, it is easily shown that $h(x)$ must be given by the logarithm of $p(x)$ and so we have

$$h(x) = -\log_2 p(x) \tag{1.56}$$

where the negative sign ensures that information is positive or zero. The choice of basis is arbitrary, and in the case of base of 2, the units of $h(x)$ are bits.

Suppose that a sender wishes to transmit the value of a random variable to a receiver. The average amount of information that they transmit is obtained by taking the expectation of (1.56) with respect to $p(x)$ and is given by

$$\mathrm{H}[x] = -\sum_x p(x) \log_2 p(x) \tag{1.57}$$

This important quantity is called the *entropy* of the random variable x. Note that $\lim_{p\to 0} p \ln p = 0$ so we shall take $p(x) \ln p(x) = 0$ whenever we encounter a value for x such that $p(x) = 0$.

The concept of entropy indeed possess useful properties. For example, the *noiseless coding theorem* states that the entropy is a lower bound on the number of bits needed to transmit the state of a random variable.

From now on, we shall switch to the use of natural logarithms in defining entropy, as this will provide a more convenient link with ideas elsewhere in this book. In this case, the entropy is measured in units of 'nats' instead of bits, which differ simply by a factor of $\ln 2$.

Actually, the concept of entropy has much earlier origins in physics through development in statical mechanics.

The minimum of entropy is 0 when one of the $p_i = 1$ and all other $p_{j\neq i} = 0$. Using the Lagrange multiplier method, we can see the maximum of entropy H is $\ln M$ when all of the $p(x_i) = 1/M$ where M is the total number of states x_i.

We can extend the definition of entropy to include distribution $p(x)$ over continuous variables x. First divide x into bins of width Δ. Then, assuming $p(x)$ is continuous, the *mean value theorem* tells us that, for each such bin, there must exist a value x_i such that

$$\int_{i\Delta}^{(i+1)\Delta} p(x)\mathrm{d}x = p(x_i)\Delta \tag{1.58}$$

We can now quantize the continuous variable x by assigning any value x to the value x_i whenever x falls in the i^{th} bin. This gives a discrete distribution for which the entropy takes the form

$$\mathrm{H}_\Delta = -\sum_i p(x_i)\Delta \ln(p(x_i)\Delta) = -\sum_i p(x_i)\Delta \ln p(x_i) - \ln \Delta \tag{1.59}$$

Omit the second term $-\ln \Delta$ on the right-hand side of (1.59) and consider the limit $\Delta \to 0$. The first term on the right-hand side of (1.59) will became

$$\lim_{\Delta \to 0} \left\{ \sum_i p(x_i)\Delta \ln p(x_i) \right\} = -\int p(x) \ln p(x) \mathrm{d}x \tag{1.60}$$

where the quantity on the right-hand side is called the *differential entropy*. We see that the discrete and continuous forms of the entropy differ by a quantity $\ln \Delta$, which diverges in the limit $\Delta \to 0$. This reflects the fact that to specify a continuous variable very precisely requires a large number of bits. For a density defined over multiple continuous variables, denoted collectively by the vector $\mathbf{x}$, the differential entropy is given by

$$\mathrm{H}[\mathbf{x}] = -\int p(\mathbf{x}) \ln p(\mathbf{x}) \mathrm{d}\mathbf{x} \tag{1.61}$$

Let us now consider the maximum entropy configuration for a continuous variable. In order for this maximum to be well defined, it will be necessary to constrain the first and second moments of $p(x)$ as well as preserving the normalization constraint. We therefore maximize the differential entropy with the three constraints

$$\int_{-\infty}^{\infty} p(x) \mathrm{d}x = 1 \tag{1.62}$$

$$\int_{-\infty}^{\infty} x p(x) \mathrm{d}x = \mu \tag{1.63}$$

$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x) \mathrm{d}x = \sigma^2 \tag{1.64}$$

The constrained maximization can be performed using Lagrange multipliers, leading finally to the result

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \tag{1.65}$$

and so the distribution that maximizes the differential entropy is the Gaussian. Note that we did not constrain the distribution to be nonnegative when we maximized the entropy. However, because the resulting distribution is indeed nonnegative, we see with hindsight that such a constraint is not necessary.

The differential entropy of the Gaussian is

$$\mathrm{H}[x] = \frac{1}{2}\left\{1 + \ln(2\pi\sigma^2)\right\} \tag{1.66}$$

Thus we see again that the entropy increases as the distribution becomes broader, i.e., as σ^2 increases. This result also shows that the differential entropy, unlike the discrete entropy, can be negative.

In a joint distribution $p(\mathbf{x}, \mathbf{y})$, if a value of $\mathbf{x}$ is already known, then the additional information needed to specify the corresponding value of $\mathbf{y}$ is given by $-\ln p(\mathbf{y}|\mathbf{x})$. Thus the average additional information needed to specify can be written as

$$\mathrm{H}[\mathbf{y}|\mathbf{x}] = -\iint p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} \tag{1.67}$$

which is called the *conditional entropy* of $\mathbf{y}$ given $\mathbf{x}$. It is easily seen, using the product rule, that the conditional satisfies the relation

$$\mathrm{H}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}] \tag{1.68}$$

Thus the information needed to describe $\mathbf{x}$ and $\mathbf{y}$ is given by the sum of the information needed to describe $\mathbf{x}$ alone plus the additional information required to specify $\mathbf{y}$ given $\mathbf{x}$.

1.6.1 Relative entropy and mutual information

Consider some unknown distribution $p(\mathbf{x})$, and suppose that we have modeled this using an approximating distribution $p(\mathbf{x})$. If we use $q(\mathbf{x})$ to construct a coding scheme for the purpose of transmitting values of $\mathbf{x}$ to a receiver, then the average *additional* amount of information (in nats) required to specify the value of $\mathbf{x}$ as a result of using $q(\mathbf{x})$ instead of the true distribution $p(\mathbf{x})$ is given by

$$\begin{aligned} \mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x}) \mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x}) \mathrm{d}\mathbf{x} \right) \\ &= -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} \mathrm{d}\mathbf{x} \end{aligned} \tag{1.69}$$

This is known as the *relative entropy* or *Kullback-Leibler divergence*, or *KL divergence*, between the distributions $p(\mathbf{x})$ and $q(\mathbf{x})$. Note that it is not a symmetrical quantity, that is to say $\mathrm{KL}(p\|q) \not\equiv \mathrm{KL}(q\|p)$.

The KL divergence satisfies $\mathrm{KL}(p\|q) \geqslant 0$ with equality if and only if $p(\mathbf{x}) = q(\mathbf{x})$. This can be proved by using *Jensen's inequality,* which is, a convex function $f(x)$ satisfies

$$f\left(\sum_{i=1}^{M} \lambda_i x_i \right) \leqslant \sum_{i=1}^{M} \lambda_i f(x_i) \tag{1.70}$$

where $\lambda_i \geqslant 0$ and $\sum_i \lambda_i = 1$, for any set of points $\{x_i\}$. If we interpret $\lambda_i = p(x_i)$ in a probability distribution over x, (1.70) can be written

$$f(\mathbb{E}[x]) \leqslant \mathbb{E}[f(x)] \tag{1.71}$$

where $\mathbb{E}[\cdot]$ denotes the expectation. For continuous variables, Jensen's inequality takes the form

$$f\left(\int \mathbf{x} p(\mathbf{x}) \mathrm{d}\mathbf{x} \right) \leqslant \int f(\mathbf{x}) p(\mathbf{x}) \mathrm{d}\mathbf{x} \tag{1.72}$$

Apply (1.72) to the Kullback-Leibler divergence (1.69) to give

$$\mathrm{KL}(p\|q) = -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} \mathrm{d}\mathbf{x} \geqslant -\ln \int q(\mathbf{x}) \mathrm{d}\mathbf{x} = 0 \tag{1.73}$$

where we have used the fact that $-\ln x$ is a convex function, together with the normalization condition $\int q(\mathbf{x})\mathrm{d}\mathbf{x} = 1$. In fact, $-\ln x$ is a strictly convex function, so the equality will hold if and only if $q(\mathbf{x}) = p(\mathbf{x})$ for all $\mathbf{x}$. Thus we can interpret the Kullback-Leibler divergence as a measure of the dissimilarity of the two distributions $p(\mathbf{x})$ and $q(\mathbf{x})$.

Suppose that data is being generated from an unknown distribution $p(\mathbf{x})$ that we wish to model. We can try to approximate this distribution using some parametric distribution $q(\mathbf{x}|\boldsymbol{\theta})$, governed by a set of adjustable parameters $\boldsymbol{\theta}$, for example a multivariate Gaussian. One way to determine $\boldsymbol{\theta}$ is to minimize the KL divergence between $p(\mathbf{x})$ and $q(\mathbf{x}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. We cannot do this directly because we don't know $p(\mathbf{x})$. Suppose, however, that we have observed a finite set of training points $\mathbf{x}_n$, for $n = 1, \ldots, N$, drawn from $p(\mathbf{x})$. Then the expectation with respect to $p(\mathbf{x})$ can be approximated by a finite sum over these points, using (1.14), so that

$$\mathrm{KL}(p\|q) \simeq \sum_{n=1}^{N} \{-\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n)\} \tag{1.74}$$

The second term on the right-hand side is independent of $\boldsymbol{\theta}$, and the first term is the negative log likelihood function for $\boldsymbol{\theta}$ under the distribution $q(\mathbf{x}|\boldsymbol{\theta})$ evaluated using the training set. Thus we see that minimizing the KL divergence is equivalent to maximizing the likelihood function.

If two sets of variables $\mathbf{x}$ and $\mathbf{y}$ given by $p(\mathbf{x}, \mathbf{y})$ are independent, then $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$. If the variables are not independent, we can gain some idea of whether they are 'close' to being independent by considering the KL divergence between $p(\mathbf{x}, \mathbf{y})$ and $p(\mathbf{x})p(\mathbf{y})$, given by

$$\begin{aligned} \mathrm{I}[\mathbf{x}, \mathbf{y}] &\equiv \mathrm{KL}(p(\mathbf{x}, \mathbf{y})\|p(\mathbf{x})p(\mathbf{y})) \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \end{aligned} \tag{1.75}$$

which is called the *mutual information* between the variables $\mathbf{x}$ and $\mathbf{y}$. $\mathrm{I}(\mathbf{x}, \mathbf{y}) \geqslant 0$ with equality if and only if $\mathbf{x}$ and $\mathbf{y}$ are independent. Use the sum and product rules of probability, we see that the mutual information is related to the conditional entropy through

$$\mathrm{I}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}] \tag{1.76}$$

Thus we can view the mutual information as the reduction in the uncertainty about $\mathbf{x}$ by virtue of being told the value of $\mathbf{y}$ (or vice versa). From a Bayesian perspective, we can view $p(\mathbf{x})$ as the prior distribution for $\mathbf{x}$ and $p(\mathbf{x}|\mathbf{y})$ as the posterior distribution after we have observed new data $\mathbf{y}$. The mutual information therefore represents the reduction in uncertainty about $\mathbf{x}$ as a consequence of the new observation $\mathbf{y}$.

2 Probability Distributions

In this chapter, we'll discuss different probability distributions. They're used as building blocks for more complex models, and to provide the opportunity to discuss some key statistical concepts.

The problem known as *density estimation* is to model the probability distribution $p(\mathbf{x})$ of a random variable $\mathbf{x}$, given a finite set $\mathbf{x}_1, \ldots, \mathbf{x}_N$ of observations. For the purpose of this chapter, we shall assume that the data points are i.i.d. It should be emphasized that the problem of density estimation is fundamentally ill-posed, because any distribution $p(\mathbf{x})$ that is nonzero at each of the data points is a potential candidate. The issue of choosing an appropriate distribution relates to the problem of model selection and is a central issue in pattern recognition.

To apply *parametric* distributions to density estimation, we need a procedure for determining suitable values for the parameters, given an observed data set. In a frequentist treatment, we choose specific values for the parameters by optimizing some criterion, such as the likelihood function. In a Bayesian treatment we introduce prior distributions over the parameters and then use Bayes theorem to compute the corresponding posterior distribution given the observed data.

Conjugate priors lead to posterior distributions having the same functional form as the priors, and therefore lead to a greatly simplified Bayesian analysis.

One limitation of the parametric approach is that it assumes a specific functional form for the distribution, which may turn out to be inappropriate for a particular application. An alternative approach is given by *nonparametric* density estimation methods in which the form of the distribution typically depends on the size of the data set. Such models still contain parameters, but these control the model complexity rather than the form of the distribution.

2.1 Binary Variables

The distribution of single binary variable $x \in \{0, 1\}$, with a single parameter μ given by $p(x = 1) = \mu$, can be written in the form

$$\text{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x} \tag{2.1}$$

where $0 \leqslant \mu \leqslant 1$, It is known as the *Bernoulli* distribution. Its mean and variance are given by

$$\mathbb{E}[x] = \mu \tag{2.2}$$

$$\text{var}[x] = \mu(1 - \mu) \tag{2.3}$$

Suppose we have a data set $\mathcal{D} = \{x_1, \ldots, x_N\}$ of observed values of x. By maximizing the likelihood function over μ, we obtain the maximum likelihood estimator

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{2.4}$$

which is also known as the *sample mean*. If we denote the number of observations of $x = 1$ within the data set by m, then we can write (2.4) in the form

$$\mu_{\mathrm{ML}} = \frac{m}{N} \tag{2.5}$$

so that the probability of landing heads is given, in this maximum likelihood framework, by the fraction of observations of heads in the data set.

If we flip a coin 3 times and happen to observe 3 heads. Then $N = m = 3$ and $\mu_{\mathrm{ML}} = 1$. In this case, the maximum likelihood result would predict that all future observations should give heads. This is an extreme example of the over-fitting associated with maximum likelihood. We shall see shortly how to arrive at more sensible conclusions through the introduction of a prior distribution over μ.

We can also work out the distribution of the number m of observations of $x = 1$, in a data set which has size N. This is called the *binomial* distribution

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m} \tag{2.6}$$

which has mean and variance

$$\mathbb{E}[m] = N_\mu \tag{2.7}$$

$$\mathrm{var}[m] = N\mu(1-\mu) \tag{2.8}$$

2.1.1 The beta distribution

Here we consider a form of prior distribution of the parameter μ in the Bernoulli distribution, which has a simple interpretation as well as some useful analytical properties. To motivate this prior, we note that the likelihood function takes the form of the product of factors of the form $\mu^x(1-\mu)^{1-x}$. If we choose a prior to be proportional to powers of μ and $(1-\mu)$, then the posterior distribution will have the same functional form as the prior. This property is called *conjugacy*. We therefore choose a prior, called the *beta* distribution, given by

$$\mathrm{Beta}[\mu|a,b] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{2.9}$$

where $\Gamma(x)$ is the gamma function defined by

$$\Gamma(x) \equiv \int_0^\infty u^{x-1}e^{-u}\mathrm{d}u \tag{2.10}$$

and the coefficients in (2.9) ensures that the beta distribution is normalized.

The mean and variance of the beta distribution are given by

$$\mathbb{E}[\mu] = \frac{a}{a+b} \tag{2.11}$$

$$\text{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)} \tag{2.12}$$

The parameters a and b are often called *hyperparameters* because they control the distribution of the parameter μ.

The posterior distribution of μ is now obtained by multiplying the beta prior (2.9) by the binomial likelihood function (2.6) and normalizing. Keep only the factors that depend on μ, we see that this posterior distribution has the form

$$p(\mu|m,l,a,b) \propto \mu^{m+a-1}(1-\mu)^{l+b-1} \tag{2.13}$$

where $l = N - m$. We see thats (2.13) has the same functional dependence on μ as the prior distribution, reflecting the conjugacy properties of the prior with respect to the likelihood function. Indeed, it is simply another beta distribution

$$p(\mu|m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1} \tag{2.14}$$

This allows us to provide a simple interpretation of the hyperparameters a and b in the prior as an *effective number of observations* of $x = 1$ and $x = 0$, respectively. Note that a and b need not be integers. Furthermore, the posterior distribution can act as the prior if we subsequently observe additional data. An additional observation of $x = 1$ simply corresponds to incrementing the value of a by 1, whereas for an observation of $x = 0$ we increment b by 1.

We see that this *sequential* approach to learning arises naturally when we adopt a Bayesian viewpoint. It is independent of the choice of prior and of the likelihood function and depends only on the assumption of i.i.d. data. Sequential methods make use of observations one at a time, or in small batches, and then discard them before the next observations are used. They can be used, for example, in real-time learning scenarios where a steady stream of data is arriving, and predictions must be made before all of the data is seen. Because they do not require the whole data set to be stored or loaded into memory, sequential methods are also useful for large data sets. Maximum likelihood methods can also be cast into a sequential framework.

Given the observed data set $\mathcal{D}$, we can predict the next x by the form

$$p(x = 1|\mathcal{D}) = \int_0^1 p(x = 1|\mu)p(\mu|\mathcal{D})\mathrm{d}\mu = \int_0^1 \mu\, p(\mu|\mathcal{D})\mathrm{d}\mu = \mathbb{E}[\mu|\mathcal{D}] \tag{2.15}$$

Using result (2.14) together with (2.11), we obtain

$$p(x = 1|\mathcal{D}) = \frac{m+a}{m+a+l+b} \tag{2.16}$$

which has a simple interpretation as the total fraction of observations (both real observations and fictitious prior observations).

As the number of observations increases, the posterior distribution becomes more sharply peaked, in which the variance goes to zero for $a \to \infty$ or $b \to \infty$. In fact, we might wonder whether it is a general property of Bayesian learning that, as we observe more and more data, the uncertainty represented by the posterior distribution will steadily decrease.

To address this, we can take a frequentist view of Bayesian learning and show that, on average, such a property does indeed hold. Consider a general Bayesian inference problem for a parameter θ for which we have observed a data set $\mathcal{D}$, described by the joint distribution $p(\theta, \mathcal{D})$. The following result

$$\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}] = \mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\mathcal{D}]] \tag{2.17}$$

where

$$\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}] \equiv \int p(\boldsymbol{\theta})\boldsymbol{\theta}\mathrm{d}\boldsymbol{\theta} \tag{2.18}$$

$$\mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\mathcal{D}]] \equiv \int \left\{ \int \boldsymbol{\theta}\, p(\boldsymbol{\theta}|\mathcal{D})\mathrm{d}\boldsymbol{\theta} \right\} p(\mathcal{D})\mathrm{d}\mathcal{D} \tag{2.19}$$

says that the posterior mean of $\boldsymbol{\theta}$, averaged over the distribution generating the data, is equal to the prior mean of $\boldsymbol{\theta}$. Similarly, we can show that

$$\mathrm{var}_{\boldsymbol{\theta}}[\boldsymbol{\theta}] = \mathbb{E}_{\mathcal{D}}[\mathrm{var}_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\mathcal{D}]] + \mathrm{var}_{\mathcal{D}}[\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\mathcal{D}]] \tag{2.20}$$

The term on the left-hand side of (2.20) is the prior variance of $\boldsymbol{\theta}$. On the right-hand side, the first term is the average posterior variance of $\boldsymbol{\theta}$, and the second term measures the variance in the posterior mean of $\boldsymbol{\theta}$. Because this variance is a positive quantity, this result shows that, on average, the posterior variance of $\boldsymbol{\theta}$ is smaller than the prior variance. The reduction in variance is greater if the variance in the posterior mean is greater. Note, however, that this result only holds on average, and that for a particular observed data set it is possible for the posterior variance to be larger than the prior variance.

2.2 Multinomial Variables

Often, we encounter discrete variables that can take on one of K possible mutually exclusive states. One particularly convenient representation to express such variables is the 1-of-K scheme, in which, the variable is represented by a K-dimensional vector $\mathbf{x}$ in which one of the elements x_k equals 1, and all remaining elements equal 0. Note that such vectors satisfy $\sum_{k=1}^{K} x_k = 1$. If we denote the probability of $x_k = 1$ by the parameter μ_k, then the distribution of $\mathbf{x}$ is given by

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \tag{2.21}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}}$, and the parameters μ_k are constrained to satisfy $\mu_k \geqslant 0$ and $\sum_k \mu_k = 1$. The distribution (2.21) can be regarded as a generalization of the Bernoulli distribution to more than two outcomes. Its mean is given by

$$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{x} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}} = \boldsymbol{\mu} \tag{2.22}$$

Now consider a data set $\mathcal{D}$ of N independent observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$. The corresponding likelihood function takes the form

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mu_k^{x_{nk}} = \prod_{k=1}^{K}\mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K}\mu_k^{m_k} \tag{2.23}$$

which depends on the N data points only through the K quantities

$$m_k = \sum_n x_{nk} \tag{2.24}$$

which represents the number of observations of $x_k = 1$. These are called the *sufficient statistics* for this distribution.

The maximum likelihood solution for $\boldsymbol{\mu}$ taking account of its constraint, which can be solved using Lagrange multiplier, is in the form

$$\mu_k^{\mathrm{ML}} = \frac{m_k}{N} \tag{2.25}$$

which is the fraction of the N observations for which $x_k = 1$.

We can consider the joint distribution of the quantities $m_1, \ldots, m_K$, conditioned on the parameters $\boldsymbol{\mu}$ and on the total number N of observations. From (2.23) this takes the form

$$\mathrm{Mult}(m_1, m_2, \ldots, m_K|\boldsymbol{\mu}, N) = \binom{N}{m_1\, m_2\, \ldots\, m_K}\prod_{k=1}^{K}\mu_k^{m_k} \tag{2.26}$$

which is known as the *multinomial* distribution. Note that the variables m_k are subject to the constraint

$$\sum_{k=1}^{K} m_k = N \tag{2.27}$$

2.2.1 The Dirichlet distribution

By inspecting the form of (2.26), we see that the conjugate prior is given by

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K}\mu_k^{\alpha_k - 1} \tag{2.28}$$

where $0 \leqslant \mu_k \leqslant 1$ and $\sum_k \mu_k = 1$. Here $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^{\mathrm{T}}$ are the parameters of the distribution. Note that, because of the summation constraint, the distribution over the space of the $\{\mu_k\}$ is confined to a *simplex* of dimensionality $K - 1$.

The normalized form for this distribution is by

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\dots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \tag{2.29}$$

which is called the *Dirichlet* distribution, here $\alpha_0 = \sum_{k=1}^{N} \alpha_K$.

Multiplying the prior (2.29) by the likelihood function (2.26), we obtain the posterior distribution for the parameters $\{\mu_k\}$ in the form

$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha} + \mathbf{m}) \tag{2.30}$$

where we have denoted $\mathbf{m} = (m_1, \dots, m_K)^{\mathrm{T}}$.