Reading Notes of Pattern Classification and Machine Learning

Tianyi Cui

August 25, 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

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

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

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

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 $\mathbf{x}$ 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

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

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

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

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 $\mathbf{m} = (m_1,\ldots,m_K)^{\mathrm{T}}$.

2.3. The Gaussian Distribution

The Gaussian, also known as the normal distribution, is a widely used model for the distribution of continuous variables. In the case of a single variable x, the Gaussian distribution can be written in the form

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{2.31}$$

where μ is the mean and σ^2 is the variance. For a D-dimensional vector $\mathbf{x}$, the multivariate Gaussian distribution takes the form

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{2.32}$$

where $\boldsymbol{\mu}$ is a D-dimensional mean vector, $\boldsymbol{\Sigma}$ is a $D\times D$ covariance matrix, and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

The Gaussian distribution arises in many different contexts and can be motivated from a variety of different perspectives. For example, we have already seen that for a single real variable, the distribution that maximizes the entropy is the Gaussian. This property applies also to the multivariate Gaussian.

Another situation in which the Gaussian distribution arises is when we consider the sum of multiple random variables. The *central limit theorem* tells us that, subject to certain mild conditions, the sum of a set of random variables, which is of course itself a random variable, has a distribution that becomes increasingly Gaussian as the number of terms in the sum increases.

The Gaussian distribution has many important analytical properties, and we shall consider several of these in detail. We begin by considering the geometrical form of the Gaussian distribution. The functional dependence of the Gaussian on $\mathbf{x}$ is through the quadratic form

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \tag{2.33}$$

which appears in the exponent. The quantity Δ is called the *Mahalanobis distance* from $\boldsymbol{\mu}$ to $\mathbf{x}$ and reduces to the Euclidian distance when $\boldsymbol{\Sigma}$ is the identity matrix. The Gaussian distribution will be constant on surfaces in $\mathbf{x}$-space for which this quadratic form is constant.

First of all, we note that the matrix $\boldsymbol{\Sigma}$ can be taken to be symmetric, without loss of generality, because any antisymmetric component would disappear from the exponent. Now consider the eigenvector equation for the covariance matrix

$$\boldsymbol{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i \tag{2.34}$$

where $i = 1, \ldots, D$. Because $\boldsymbol{\Sigma}$ is a real, symmetric matrix its eigenvalues will be real, and its eigenvectors can be chosen to form an orthonormal set, so that

$$\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = I_{ij} \tag{2.35}$$

The covariance matrix $\boldsymbol{\Sigma}$ can be expressed as an expansion in terms of its eigenvectors in the form

$$\boldsymbol{\Sigma} = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}} \tag{2.36}$$

and similarly the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ can be expressed as

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u_i}^{\mathrm{T}} \tag{2.37}$$

Substituting (2.37) into (2.33), the quadratic form becomes

$$\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i} \tag{2.38}$$

where we have defined

$$y_i = \mathbf{u}_i^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}) \tag{2.39}$$

We can interpret $\{y_i\}$ as a new coordinate system defined by the orthonormal vectors $\mathbf{u}_i$ that are shifted and rotate with respect to the original x_i coordinates. Forming the factor $\mathbf{y} = (y_1, \ldots, y_D)^{\mathrm{T}}$, we have

$$\mathbf{y} = \mathbf{U}(\mathbf{x} - \boldsymbol{\mu}) \tag{2.40}$$

where $\mathbf{U}$ is a matrix whose rows are given by $\mathbf{u}_i^{\mathrm{T}}$. From (2.35) $\mathbf{U}$ is an *orthogonal matrix*, i.e. $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$.

The quadratic form, and hence the Gaussian density, will be constant on surfaces for which (2.39) is constant. If all of the eigenvalues λ_i are positive, then these surfaces represent ellipsoids, with there centers at $\boldsymbol{\mu}$ and their axes oriented along $\mathbf{u}_i$, and with scaling factors in the directions of the axes given by $\lambda_i^{1/2}$.

For the Gaussian distribution to be well defined, it is necessary for all of the eigenvalues λ_i of the covariance matrix to be strictly positive, i.e. $\mathbf{\Sigma}$ to be positive definite, otherwise the distribution cannot be properly normalized. In Chapter 12, we will encounter Gaussian distributions for which one or more of the eigenvalues are zero, in which case the distribution is singular and is confined to a subspace of lower dimensionality.

Now consider the form of the Gaussian distribution in the new coordinate system defined by the y_i. In going from the $\mathbf{x}$ to the $\mathbf{y}$ coordinate system, we have a Jacobian matrix $\mathbf{J}$ with elements given by

$$J_{ij} = \frac{\partial x_i}{\partial y_j} = U_{ji} \tag{2.41}$$

where U_{ji} are the elements of the matrix $\mathbf{U}^{\mathrm{T}}$. Using the orthonormality property of the matrix $\mathbf{U}$, we see that the square of the determinant of the Jacobian matrix is

$$|\mathbf{J}|^2 = |\mathbf{U}^{\mathrm{T}}|^2 = |\mathbf{U}^{\mathrm{T}}||\mathbf{U}| = |\mathbf{U}^{\mathrm{T}}\mathbf{U}| = |\mathbf{I}| = 1 \tag{2.42}$$

and hence $|\mathbf{J}| = 1$. Also, the determinant $|\mathbf{\Sigma}|$ of the covariance matrix can be written as the product of its eigenvalues, and hence

$$|\mathbf{\Sigma}|^{1/2} = \prod_{j=1}^{D} \lambda_j^{1/2} \tag{2.43}$$

Thus in the y_j coordinate system, the Gaussian distribution takes the form

$$p(\mathbf{y}) = p(\mathbf{x})|\mathbf{J}| = \prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\left\{-\frac{y_j^2}{2\lambda_j}\right\} \tag{2.44}$$

which is the product of D independent univariate Gaussian distributions. The eigenvectors therefore define a new set of shifted and rotated coordinates with respect to which the joint probability distribution factorizes into a product of independent distributions.

It can be proved

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} \tag{2.45}$$

and

$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \mathbf{\Sigma} \tag{2.46}$$

Therefore

$$\begin{aligned} \mathrm{cov}[\mathbf{x}] &= \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}\right] \\ &= \mathbf{\Sigma} \end{aligned} \tag{2.47}$$

Although the Gaussian distribution is widely used as a density model, it suffers from some significant limitations. Its number of parameters, $D(D+3)/2$, grows quadratically with D. With large D, the computational task of manipulating and inverting large matrices can become prohibitive. One way to address this problem is to use restricted

forms of the covariance matrix. If we restrict $\mathbf{\Sigma} = \text{diag}(\sigma_i^2)$, then the total of independent parameters is $2D$ and the corresponding contours of constant density are given by axis-aligned ellipsoids. We could further restrict $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, known as an *isotropic* covariance, giving $D + 1$ independent parameters and spherical surfaces of constant density. Unfortunately,whereas such approaches limit the number of degrees of freedom in the distribution and make inversion of the covariance matrix a much faster operation, they also greatly restrict the form of the probability density and limit its ability to capture interesting correlations in the data.

A further limitation of the Gaussian distribution is that it is intrinsically unimodal (i.e., has a single maximum) and so is unable to provide a good approximation to multimodal distributions. Thus the Gaussian distribution can be both too flexible, in the sense of having too many parameters, while also being too limited in the range of distributions that it can adequately represent. We will see later that the introduction of *latent* variables, also called *hidden* variables or *unobserved* variables, allows both of these problems to be addressed.

2.3.1. Conditional Gaussian distributions

Given a joint Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma})$ with $\mathbf{\Lambda} \equiv \mathbf{\Sigma}^{-1}$ and

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \tag{2.48}$$

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab} \\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb} \end{pmatrix}, \quad \mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_{aa} & \mathbf{\Lambda}_{ab} \\ \mathbf{\Lambda}_{ba} & \mathbf{\Lambda}_{bb} \end{pmatrix} \tag{2.49}$$

The conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ is given by

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{a|b}, \mathbf{\Lambda}_{aa}^{-1}) \tag{2.50}$$

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \mathbf{\Lambda}_{aa}^{-1}\mathbf{\Lambda}_{ab}(\mathbf{x}_a - \mathbf{x}_b) \tag{2.51}$$

2.3.2. Marginal Gaussian distributions

Same conditions as above, the marginal distribution $p(\mathbf{x}_a)$ is given by

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \mathbf{\Sigma}_{aa}) \tag{2.52}$$

2.3.3. Bayes theorem for Gaussian variables

Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}^{-1}) \tag{2.53}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{2.54}$$

the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \tag{2.55}$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \tag{2.56}$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1} \tag{2.57}$$

2.3.4. Maximum likelihood for the Gaussian

Given a data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^{\mathrm{T}}$ in which the observations $\{\mathbf{x}_n\}$ are assumed to be drawn independently from a multivariate Gaussian distribution. The log likelihood function is given by

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}) \tag{2.58}$$

It depends on the data set only through the two quantities

$$\sum_{n=1}^{N}\mathbf{x}_n, \qquad \sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} \tag{2.59}$$

These are known as the *sufficient statistics* for the Gaussian distributions. The derivative of the log likelihood with respect to $\boldsymbol{\mu}$ is given by

$$\frac{\partial}{\partial\boldsymbol{\mu}}\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}) \tag{2.60}$$

and setting this derivative to zero, we obtain the solution

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \tag{2.61}$$

The maximization of (2.58) is more involved and the result is given by

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}} \tag{2.62}$$

If we evaluate the expectations of the maximum likelihood solutions under the true distribution, we obtain the following results

$$\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}] = \boldsymbol{\mu}$$

$$\mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] = \frac{N-1}{N}\boldsymbol{\Sigma}$$

Hence the maximum likelihood estimate for the covariance is biased. We can correct this bias by defining a different estimator $\widetilde{\boldsymbol{\Sigma}}$ given by

$$\widetilde{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}} \tag{2.63}$$

2.3.5. Sequential estimation

Our discussion of the maximum likelihood solution for the parameters of a Gaussian distribution provides a convenient opportunity to give a more general discussion of the topic of sequential estimation for maximum likelihood. Sequential methods allow data points to be processed one at a time and then discarded and are important for on-line applications, and also where large data sets are involved so that batch processing of all data points at once is infeasible.

If we denote the maximum likelihood estimator of the mean $\boldsymbol{\mu}_{\mathrm{ML}}^{(N)}$ when it is based on N observations, it can be simply shown

$$\boldsymbol{\mu}_{\mathrm{ML}}^{(N)} = \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} + \frac{1}{N}(\mathbf{x}_N - \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)}) \tag{2.64}$$

This result has a nice interpretation. After observing $N-1$ data points we have estimate $\boldsymbol{\mu}$ by $\boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)}$. We now observe data point $\mathbf{x}_N$, and we obtain our revised estimate $\boldsymbol{\mu}_{\mathrm{ML}}^{(N)}$ by moving the old estimate a small amount, proportional to $1/N$, in the direction of the 'error signal' $(\mathbf{x}_N - \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)})$. Note that, as N increases, so the contribution from successive data points gets smaller.

The *Robbins-Monro* algorithm is a more general formulation of sequential leading. Consider a pair of random variable θ and z governed by a joint distribution $p(z, \theta)$. The conditional expectation of z given θ defines a deterministic function $f(\theta)$ that is given by

$$f(\theta) \equiv \mathbb{E}[z|\theta] = \int z p(z|\theta) \mathrm{d}z \tag{2.65}$$

The Robbins-Monro can find the root $\theta^\star$ at which $f(\theta^\star) = 0$ sequentially.

2.3.6. Beyesian inference for the Gaussian

Now we develop a Bayesian treatment by introducing prior distributions over $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Suppose the covariance is known and mean is unknown, the conjugate prior of the mean is simply another Gaussian. In fact, the Bayesian paradigm leads very naturally to a sequential view of the inference problem.

$$p(\boldsymbol{\mu}|D) = \left[p(\boldsymbol{\mu}) \prod_{n=1}^{N-1} (p(\mathbf{x}_n|\boldsymbol{\mu}) \right] p(\mathbf{x}_N|\boldsymbol{\mu}) \tag{2.66}$$

The term in square brackets is (up to a normalization coefficient) just the posterior distribution after observing $N-1$ data points. We see that this can be viewed as a prior distribution, which is combined using Bayes theorem with the likelihood function associated with data point $\mathbf{x}_N$ to arrive at the posterior distribution after observing N data points. This sequential view of Bayesian inference is very general and applies to any problem in which the observed data are assumed to be independent and identically distributed.

Assume the mean is known and we wish to infer the variance. It's more convenient to work with the precision $\lambda \equiv 1/\sigma^2$. The likelihood function for λ takes the form

$$p(\mathbf{X}|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\} \tag{2.67}$$

The corresponding conjugate prior should therefore be proportional to the product of a power of λ and the exponential of a linear function of λ. This corresponds to the *gamma* distribution which is defined by

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \tag{2.68}$$

The gamma distribution has a finite integral if $a > 0$, and the distribution itself is finite if $a \geqslant 1$. The mean and variance are given by

$$\mathbb{E}[\lambda] = \frac{a}{b} \tag{2.69}$$

$$\mathrm{var}[\lambda] = \frac{a}{b^2} \tag{2.70}$$

Consider a prior distribution $\mathrm{Gam}(\lambda|a_0, b_0)$. If we multiply by the likelihood function (2.67), then we obtain a posterior distribution

$$p(\lambda|\mathbf{X}) \propto \lambda^{a_0-1}\lambda^{N/2} \exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\} \tag{2.71}$$

we can find $p(\mu|\lambda)$ and $p(\lambda)$ by inspection. In particular, we see that $p($which we recognize as a gamma distribution of the form $\mathrm{Gam}(\lambda|a_N, b_N)$ where

$$a_N = a_0 + \frac{N}{2} \tag{2.72}$$

$$b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 = b_0 + \frac{N}{2}\sigma^2_{\mathrm{ML}} \tag{2.73}$$

where σ^2_{ML} is the maximum likelihood estimator of the variance.

From (2.72), we see that the effect of observing N data points is to increase the value of the coefficient a by $N/2$. Thus we can interpret the parameter a_0 in the prior in terms of $2a_0$ 'effective' prior observations. Similarly, from (2.73) we can interpret the parameter b_0 in the prior as arising from the $2a_0$ 'effective' prior observations having variance b_0/a_0. Recall that we made an analogous interpretation for the Dirichlet prior. These distributions are examples of the exponential family, and we shall see that the interpretation of a conjugate prior in terms of effective fictitious data points is a general one for the exponential family of distributions.

Instead of working with the precision, we can consider the variance itself. The conjugate prior in this case is called the *inverse gamma* distribution.

Now suppose that both the mean and the precision are unknown. To find a conjugate prior, we consider the dependence of the likelihood function on μ and λ

$$\begin{aligned} p(\mathbf{X}|\mu,\lambda) &= \prod_{n=1}^{N}\left(\frac{\lambda}{2\pi}\right)^{1/2}\exp\left\{-\frac{\lambda}{2}(x_n-\mu)^2\right\} \\ &\propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu\sum_{n=1}^{N}x_n-\frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right\} \end{aligned} \tag{2.74}$$

We now wish to identify a prior distribution $p(\mu,\lambda)$ that has the same functional dependence on μ and λ as the likelihood function and that should therefore take the form

$$\begin{aligned} p(\mu,\lambda) &\propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^\beta \exp\{c\lambda\mu-d\lambda\} \\ &= \exp\left\{-\frac{\beta\lambda}{2}(\mu-c/\beta)^2\right\}\lambda^{\beta/2}\exp\left\{-\left(d-\frac{c^2}{2\beta}\right)\lambda\right\} \end{aligned}$$

where c, d, and β are constants. Since we can always write $p(\mu,\lambda)=p(\mu|\lambda)p(\lambda)$, we can find $p(\mu|\lambda)$ and $p(\lambda)$ by inspection, and we already know $p(\mu|\lambda)$ is a Gaussian whose precision is a linear function of λ and that $p(\lambda)$ is a gamma distribution, so that the normalized prior takes the form

$$p(\mu,\lambda)=\mathcal{N}(\mu|\mu_0,(\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a,b) \tag{2.75}$$

where we have defined new constants given by $\mu_0=c/\beta$, $a=1+\beta/2$, $b=d-c^2/2\beta$. The distribution (2.75) is called the *normal-gamma* or *Gaussian-gamma* distribution. Note that this is not simply the product of an independent Gaussian prior over μ and a gamma prior over λ, because the precision of μ is a linear function of λ. Even if we chose a prior in which μ and λ were independent, the posterior distribution would exhibit a coupling between them.

In the case of the multivariate Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1})$ for a D-dimensional variable $\mathbf{x}$, the conjugate prior distribution for the mean $\boldsymbol{\mu}$, assuming the precision is known, is again a Gaussian. For known mean and unknown precision matrix $\boldsymbol{\Lambda}$, the conjugate prior is the *Wishart* distribution given by

$$\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W},\nu)=B\,|\boldsymbol{\Lambda}|^{(\nu-D-1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right) \tag{2.76}$$

where ν is called the number of *degree of freedom* of the distribution, $\mathbf{W}$ is a $D\times D$ scale matrix, and the normalization constant B is given by

$$B(\mathbf{W},\nu)=|\mathbf{W}|^{\nu/2}\left(2^{\nu D/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1} \tag{2.77}$$

Again, it is also possible to define a conjugate prior over the covariance matrix itself, rather than over the precision matrix, which leads to the *inverse Wishart* distribution, although we shall not discuss this further. If both the mean and the precision are unknown, then, following a similar line of reasoning to the univariate case, the conjugate prior is given by

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{\mu}_0, \beta, \mathbf{W}, \nu) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\beta\boldsymbol{\Lambda})^{-1})\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, \nu) \tag{2.78}$$

which is known as the *normal-Wishart* or *Gaussian-Wishart* distribution.

2.3.7. Student's t-distribution

We have seen that the conjugate prior for the precision of a Gaussian is given by a gamma distribution. If we have a univariate Gaussian $\mathcal{N}(x|\mu, \tau^{-1})$ together with a Gamma prior $\mathrm{Gam}(\tau|a, b)$ and we integrate out the precision, we obtain the marginal distribution of x in the form

$$\begin{aligned} p(x|\mu, a, b) &= \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1})\mathrm{Gam}(\tau|a, b)\mathrm{d}\tau \\ &= \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\left[b + \frac{(x-\mu)^2}{2}\right]^{-a-1/2}\Gamma(a+1/2) \end{aligned} \tag{2.79}$$

By convention we define new parameters given by $\nu = 2a$ and $\lambda = a/b$, in terms of which the distribution $p(x|\mu, a, b)$ takes the form

$$\mathrm{St}(x|\mu, \lambda, \nu) = \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]^{-\nu/2-1/2} \tag{2.80}$$

which is known as *Student's t-distribution*. The parameter λ is sometimes called the *precision* of the t-distribution, even though it is not in general equal to the inverse of the variance. The parameter ν is called the *degrees of freedom*. For the particular case of $\nu = 1$, the t-distribution reduces to the *Cauchy* distribution, while in the limit $\nu \to \infty$ the t-distribution $\mathrm{St}(x|\mu, \lambda, \nu)$ becomes a Gaussian $\mathcal{N}(x|\mu, \lambda^{-1})$.

From (2.79), we see that Student's t-distribution is obtained by adding up an infinite number of Gaussian distributions having the same mean but different precisions. This can be interpreted as an infinite mixture of Gaussians. he result is a distribution that in general has longer 'tails' than a Gaussian. This gives the t-distribution an important property called *robustness*, which means that it is much less sensitive than the Gaussian to the presence of a few data points which are *outliers*. Note that the maximum likelihood solution for the t-distribution can be found using the expectation-maximization (EM) algorithm. Outliers can arise in practical applications either because the process that generates the data corresponds to a distribution having a heavy tail or simply through mislabeled data. Robustness is also an important property for regression problems. Unsurprisingly, the least squares approach to regression does not exhibit robustness, because it corresponds to maximum likelihood under a (conditional) Gaussian distribution. By

basing a regression model on a heavy-tailed distribution such as a t-distribution, we obtain a more robust model.

If we go back to (2.79) and substitute the alternative parameters $\nu = 2a$, $\lambda = a/b$, and $\eta = \tau b/a$, we see that the t-distribution can be written in the form

$$\mathrm{St}(x|\mu, \lambda, \nu) = \int_0^\infty \mathcal{N}(x|\mu, (\eta\lambda)^{-1})\mathrm{Gam}(\eta|\nu/2, \nu/2)\mathrm{d}\eta \tag{2.81}$$

We can generalize this to a multivariate Gaussian $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda})$ to obtain the corresponding multivariate Student's t-distribution in the form

$$\mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1})\mathrm{Gam}(\eta|\nu/2, \nu/2)\mathrm{d}\eta \tag{2.82}$$

We can evaluate this integral to give

$$\mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \frac{\Gamma(D/2 + \nu/2)}{\Gamma(\nu/2)} \frac{|\boldsymbol{\Lambda}|^{1/2}}{(\pi\nu)^{D/2}} \left[1 + \frac{\Delta^2}{\nu}\right]^{-D/2-\nu/2} \tag{2.83}$$

where D is the dimensionality of $\mathbf{x}$, and Δ^2 is the squared Mahalanobis distance defined by

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu}) \tag{2.84}$$

The multivariate form of Student's t-distribution satisfies the following properties

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} \qquad \text{if} \quad \nu > 1 \tag{2.85}$$

$$\mathrm{cov}[\mathbf{x}] = \frac{\nu}{(\nu - 2)} \boldsymbol{\Lambda}^{-1} \qquad \text{if} \quad \nu > 2 \tag{2.86}$$

$$\mathrm{mode}[\mathbf{x}] = \boldsymbol{\mu} \tag{2.87}$$

with corresponding results for the univariate case.

2.3.8. Periodic variables

Gaussian distributions are not always applicable for continuous variables, e.g., for periodic variables, which can conveniently be represented using an angular (polar) coordinate $0 \leqslant \theta < 2\pi$.

We might be tempted to treat periodic variables by choosing some direction as the origin and then applying a conventional distribution such as the Gaussian. Such an approach, however, would give results that were strongly dependent on the arbitrary choice of origin. For example, for two observations at $\theta_1 = 1°$ and $\theta_2 = 359°$, if we choose the origin at $0°$ then the sample mean will be $180°$ with the standard deviation $179°$, whereas if we choose the origin at $180°$, then the mean will be $0°$ and the standard deviation will be $1°$. We clearly need to develop a special approach for the treatment of periodic variables.

For a set of observations $\mathcal{D} = \{\theta_1, \ldots, \theta_N\}$, to find an invariant measure of the mean, we note that the observations can be viewed as points on the unit circle and can therefore

be described instead by two-dimensional unit vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ where $\|\mathbf{x}_n\| = 1$ for $n = 1, \ldots, N$. We can average the vectors $\{\mathbf{x}_n\}$ to give

$$\overline{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \tag{2.88}$$

and then find the corresponding angle $\overline{\theta}$ of this average. We can solve for $\overline{\theta}$ to give

$$\overline{\theta} = \tan^{-1} \left\{ \frac{\sum_n \sin \theta_n}{\sum_n \cos \theta_n} \right\} \tag{2.89}$$

Shortly, we shall see how this result arises naturally as the maximum likelihood estimator for an appropriately defined distribution over a periodic variable.

We now consider a periodic generalization of the Gaussian called the *von Mises* distribution. Here we shall limit our attention to univariate distributions, although periodic distributions can also be found over hyperspheres of arbitrary dimension.

By convention, we will consider $p(\theta)$ that have period 2π which must satisfy the three conditions

$$p(\theta) \geqslant 0 \tag{2.90}$$

$$\int_0^{2\pi} p(\theta) \mathrm{d}\theta = 1 \tag{2.91}$$

$$p(\theta + 2\pi) = p(\theta) \tag{2.92}$$

We can easily obtain a Gaussian-like distribution that satisfies these three properties as follows. Consider a Gaussian distribution over two variables $\mathbf{x} = (x_1, x_2)$ having mean $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and a covariance matrix $\boldsymbol{\Sigma} = \sigma \mathbf{I}_2$, so that

$$p(x_1, x_2) = \frac{1}{2\pi\sigma^2} \exp \left\{ -\frac{(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2}{2\sigma^2} \right\} \tag{2.93}$$

By transforming from Cartesian coordinates (x_1, x_2) to polar coordinates (r, θ) and (μ_1, μ_2) to (r_0, θ_0), then condition on the unit circle $r = 1$, we obtain our final expression for the distribution of $p(\theta)$ along the unit circle $r = 1$ in the form

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp \{m \cos(\theta - \theta_0)\} \tag{2.94}$$

which is called the *von Mises* distribution, or the *circular normal*. Here θ_0 corresponds to the mean of the distribution, while m, which is known as the *concentration* parameter, is analogous to the precision for the Gaussian. The normalization coefficient for (2.94) is expressed in terms of $I_0(m)$, which is the zeroth-order Bessel function of the first kind and is defined by

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp\{m \cos \theta\} \mathrm{d}\theta \tag{2.95}$$

For large m, the distribution becomes approximately Gaussian.

Now consider the maximum likelihood estimators for the parameters θ_0 and m. The log likelihood functions is given by

$$\ln p(\mathcal{D}|\theta_0, m) = -N \ln(2\pi) - N \ln I_0(m) + m \sum_{n=1}^{N} \cos(\theta_n - \theta_0) \tag{2.96}$$

Setting the derivative with respect to θ_0 equal to zero gives

$$\sum_{n=1}^{N} \sin(\theta_n - \theta_0) = 0 \tag{2.97}$$

from which we obtain

$$\theta_0^{\mathrm{ML}} = \tan^{-1} \left\{ \frac{\sum_n \sin \theta_n}{\sum_n \cos \theta_n} \right\} \tag{2.98}$$

which we recognized as the result (2.89).

Similarly, maximizing (2.96) with respect to m, and make use of $I_0'(m) = I_1(m)$, we have

$$A(m) = \frac{1}{N} \sum_{n=1}^{N} \cos(\theta_n - \theta_0^{\mathrm{ML}}) \tag{2.99}$$

where we have defined

$$A(m) = \frac{I_1(m)}{I_0(m)} \tag{2.100}$$

We can write (2.99) in the form

$$A(m_{\mathrm{ML}}) = \left(\frac{1}{N} \sum_{n=1}^{N} \cos \theta_n \right) \cos \theta_0^{\mathrm{ML}} - \left(\frac{1}{N} \sum_{n=1}^{N} \sin \theta_n \right) \sin \theta_0^{\mathrm{ML}} \tag{2.101}$$

The right hand side of (2.101) is easily evaluated, and the function $A(m)$ can be inverted numerically.

One limitation of the von Mises distribution is that it is unimodal. By forming *mixtures* of von Mises distributions, we obtain a flexible framework for modeling periodic variables that can handle multimodality.

2.3.9. Mixtures of Gaussians

The Gaussian distribution suffers from significant limitations when it comes to modeling real data sets, for example, the 'Old Faithful' data set, which forms two dominant clumps, whereas a linear superposition of two Gaussians gives a better characterization of the data set.

Such superpositions, formed by taking linear combinations of more basic distributions such as Gaussians, can be formulated as probabilistic models known as *mixture distributions*. By using a sufficient number of Gaussians, and by adjusting their means and covariances as well as the coefficients in the linear combination, almost any continuous density can be approximated to arbitrary accuracy.

We therefore consider a superposition of K Gaussian densities of the form

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2.102}$$

which is called a *mixture of Gaussians*. Each Gaussian density $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is called a *component* of the mixture and has its own mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$.

In this section we shall consider Gaussian components to illustrate the framework of mixture models. More generally, mixture models can comprise linear combinations of other distributions.

The parameters π_k in (2.102) are called *mixing coefficients*. It's easy to verify they should satisfy $\sum_{k=1}^{K} \pi_k = 1$ and $0 \leqslant p(\mathbf{x}) \leqslant 1$ for $p(\mathbf{x})$ to be probability.

From the sum and product rules, the marginal density is given by

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k)p(\mathbf{x}|k) \tag{2.103}$$

which is equivalent to (2.102) in which we can view $\pi_k = p(k)$ as the prior probability of picking the k^{th} component, and the density $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = p(\mathbf{x}|k)$ as the probability of $\mathbf{x}$ conditioned on k. An important role is played by the posterior probabilities $p(k|\mathbf{x})$, which are also known as *responsibilities*. From Bayes' theorem these are given by

$$\begin{aligned} \gamma_k(\mathbf{x}) &\equiv p(k|\mathbf{x}) \\ &= \frac{p(k)p(\mathbf{x}|k)}{\sum_l p(l)p(\mathbf{x}|l)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_l \pi_l \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \end{aligned} \tag{2.104}$$

We shall discuss the probabilistic interpretation of the mixture distribution in greater detail in Chapter 9.

The form of the Gaussian mixture distribution is governed by the parameters $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, where we have used the notation $\boldsymbol{\pi} \equiv \{\pi_1, \ldots, \pi_K\}$, $\boldsymbol{\mu} \equiv \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$ and $\boldsymbol{\Sigma} \equiv \{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}$. One way to set the values of these parameters is to use maximum likelihood. From (2.102) the log of the likelihood function is given by

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \tag{2.105}$$

where $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. We immediately see that the situation is now much more complex than with a single Gaussian, due to the presence of the summation over k inside the logarithm. As a result, the maximum likelihood solution for the parameters no longer has a closed-form analytical solution. One approach to maximizing the likelihood function is to use iterative numerical optimization techniques. Alternatively we can employ a powerful framework called *expectation maximization*, which will be discussed at length in Chapter 9.

2.4. The Exponential Family

The probability distributions that we have studied so far in this chapter (with the exception of the Gaussian mixture) are specific examples of a broad class of distributions called the *exponential family*. Members of the exponential family have many important properties in common, and it is illuminating to discuss these properties in some generality.

The exponential family of distributions over $\mathbf{x}$, given parameters $\boldsymbol{\eta}$, is defined to be the set of distributions of the form

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \tag{2.106}$$

where $\mathbf{x}$ may be scalar or vector, and may be discrete or continuous. Here $\boldsymbol{\eta}$ are called the *natural parameters* of the distribution, and $\mathbf{u}(\mathbf{x})$ is some function of $\mathbf{x}$. The function $g(\boldsymbol{\eta})$ can be interpreted as the coefficient that ensures that the distribution is normalized and therefore satisfies

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathrm{d}\mathbf{x} = 1 \tag{2.107}$$

Consider first the Bernoulli distribution

$$p(x|\mu) = \mathrm{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \tag{2.108}$$

Expressing the right-hand side as the exponential of the logarithm we have

$$\begin{aligned} p(x|\mu) &= \exp\{x \ln \mu + (1-x)\ln(1-\mu)\} \\ &= (1-\mu)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right)x\right\} \end{aligned} \tag{2.109}$$

Comparison with (2.106) allows us to identify

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right) \tag{2.110}$$

which can solve for μ to give $\mu = \sigma(\eta)$, where

$$\sigma(\eta) = \frac{1}{1+\exp(-\eta)} \tag{2.111}$$

is called the *logistic sigmoid* function. Thus we can write the Bernoulli distribution using the standard representation (2.106) in the form

$$p(x|\eta) = \sigma(-\eta)\exp(\eta x) \tag{2.112}$$

Comparison with (2.106) shows that

$$u(x) = x \tag{2.113}$$

$$h(x) = 1 \tag{2.114}$$

$$g(\eta) = \sigma(-\eta) \tag{2.115}$$

Next consider the multinomial distribution that, for a single observation $\mathbf{x}$, takes the form

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} \tag{2.116}$$

where $\mathbf{x} = (x_1, \ldots, x_N)^{\mathrm{T}}$. We can write this in the standard representation so that

$$p(\mathbf{x}|\boldsymbol{\eta}) = \exp(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{x}) \tag{2.117}$$

where $\eta_k = \ln \mu_k$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_M)^{\mathrm{T}}$. Again, comparing with (2.106) we have

$$\mathbf{u}(\mathbf{x}) = \mathbf{x} \tag{2.118}$$
$$h(\mathbf{x}) = 1 \tag{2.119}$$
$$g(\boldsymbol{\eta}) = 1 \tag{2.120}$$

Note that the parameters η_k are not independent because the parameters μ_k are subject to the constraint

$$\sum_{k=1}^{M} \mu_k = 1 \tag{2.121}$$

In some circumstances, it will be convenient to remove this constraint by expressing the distribution in terms of only $M - 1$ parameters. Making use of the constraint (2.121) to give

$$\mu_k = \frac{\exp(\eta_k)}{1 + \sum_j \exp(\eta_j)} \tag{2.122}$$

This is called the *softmax* function, or the *normalized exponential*. In this representation, the multinomial distribution therefore takes the form

$$p(\mathbf{x}|\boldsymbol{\eta}) = \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k)\right)^{-1} \exp(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{x}) \tag{2.123}$$

This is the standard form of the exponential family, with parameter vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{M-1})^{\mathrm{T}}$ in which

$$\mathbf{u}(\mathbf{x}) = \mathbf{x} \tag{2.124}$$
$$h(\mathbf{x}) = 1 \tag{2.125}$$
$$g(\boldsymbol{\eta}) = \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k)\right)^{-1} \tag{2.126}$$

Finally, let us consider the Gaussian distribution. For the univariate Gaussian, we have

$$\begin{aligned} p(x|\mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\right\} \end{aligned} \tag{2.127}$$

which, after some rearrangement, can be cast in the standard exponential family form (2.106) with

$$\boldsymbol{\eta} = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} \tag{2.128}$$

$$\mathbf{u}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \tag{2.129}$$

$$h(\mathbf{x}) = (2\pi)^{-1/2} \tag{2.130}$$

$$g(\boldsymbol{\eta}) = (-2\eta_2)^{1/2} \exp\left(\frac{\eta_1^2}{4\eta_2}\right) \tag{2.131}$$

2.4.1. Maximum likelihood and sufficient statistics

Let us now consider the problem of estimating the parameter vector $\boldsymbol{\eta}$ in the general exponential family distribution (2.106) using the technique of maximum likelihood. Taking the gradient of both sides of (2.107) with respect to $\boldsymbol{\eta}$, and use (2.106), we can obtain the result

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})] \tag{2.132}$$

Note that the covariance of $\mathbf{u}(\mathbf{x})$ can be expressed in terms of the second derivatives of $g(\boldsymbol{\eta})$, and similarly for higher order moments. Thus, provided we can normalize a distribution from the exponential family, we can always find its moments by simple differentiation.

Now consider a set of i.i.d. data denoted by $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, for which the likelihood function is given by

$$p(\mathbf{X}|\boldsymbol{\eta}) = \left(\prod_{n=1}^{N} h(\mathbf{x}_n)\right) g(\boldsymbol{\eta})^N \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)\right\} \tag{2.133}$$

Setting the gradient of $\ln p(\mathbf{X}|\boldsymbol{\eta})$ with respect to $\boldsymbol{\eta}$ to zero, we get the following condition to be satisfied by the maximum likelihood estimator $\boldsymbol{\eta}_{\mathrm{ML}}$

$$-\nabla \ln g(\boldsymbol{\eta}_{\mathrm{ML}}) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n) \tag{2.134}$$

which can in principle be solved to obtain $\boldsymbol{\eta}_{\mathrm{ML}}$. We see that the *sufficient statistic* of the distribution (2.106) is $\sum_n \mathbf{u}(\mathbf{x}_n)$. For example, for the Gaussian $\mathbf{u}(x) = (x, x^2)^{\mathrm{T}}$, so we should keep both the sum of $\{x_n\}$ and the sum of $\{x_n^2\}$.

2.4.2. Conjugate priors

In general, for a given probability distribution $p(\mathbf{x}|\boldsymbol{\eta})$, we can seek a prior $p(\boldsymbol{\eta})$ that is conjugate to the likelihood function, so that the posterior distribution has the same

functional form as the prior. For any member of the exponential family (2.106), there exists a conjugate prior that can be written in the form

$$p(\boldsymbol{\eta}|\boldsymbol{\chi},\nu) = f(\boldsymbol{\chi},\nu)g(\boldsymbol{\eta})^{\nu}\exp\left\{\nu\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\chi}\right\} \tag{2.135}$$

where $f(\boldsymbol{\chi},\nu)$ is a normalization coefficient, and $g(\boldsymbol{\eta})$ is the same function as appears in (2.106). To multiply the prior (2.135) by the likelihood function (2.133) will obtain the posterior distribution, up to a normalization coefficient, in the form

$$p(\boldsymbol{\eta}|\mathbf{X},\boldsymbol{\chi},\nu) \propto g(\boldsymbol{\eta})^{\nu+N}\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\left(\sum_{n=1}^{N}\mathbf{u}(\mathbf{x}_n)+\nu\boldsymbol{\chi}\right)\right\} \tag{2.136}$$

This again takes the same functional form as the prior (2.135), confirming conjugacy. Furthermore, we see that the parameter ν can be interpreted as an effective number of pseudo-observations in the prior, each of which has a value for the sufficient statistic $\mathbf{u}(\mathbf{x})$ given by $\boldsymbol{\chi}$.

2.4.3. Noninformative priors

In many cases, we may have little idea of what form the distribution should like. We may then seek a form of prior distribution, called a *noninformative prior*, which is intended to have as little influence on the posterior distribution as possible. This is sometimes referred to as 'letting the data speak for themselves'.

If we have a distribution $p(x|\lambda)$ governed by a parameter λ, we might be tempted to propose a prior distribution $p(\lambda) = \text{const}$. If λ is a discrete variable with K states, simply set $p(\lambda) = 1/K$. In the case of continuous parameters, there are two difficulties. First, if the domain of λ is unbounded, the prior distribution cannot be correctly normalized. Such priors are called *improper*. In practice, improper priors can often be used provided the corresponding posterior distribution is *proper*, i.e., that it can be correctly normalized.

A second difficulty arises from the transformation behavior of a probability density under a nonlinear change of variables, given by (1.10). If a function $h(\lambda)$ is constant and $\lambda = g(\eta)$, then $\widehat{h}(\eta) = h(g(\eta))$ will also be constant. However, if we choose the density $p_\lambda(\lambda)$ to be constant, then the density of η will be given, from (1.10), by

$$p_\eta(\eta) = p_\lambda(\lambda)\left|\frac{\mathrm{d}\lambda}{\mathrm{d}\eta}\right| = p_\lambda(g(\eta))\left|g'(\eta)\right| \propto \left|g'(\eta)\right| \tag{2.137}$$

and so the density over η will not be constant. This issue does not arise when we use maximum likelihood, because the likelihood function $p(x|\lambda)$ is a simple function of λ and so we are free to use any convenient parameterization. If, however, we are to choose a prior distribution that is constant, we must take care to use an appropriate representation for the parameters.

Here we consider two simple examples of noninformative priors. First of all, if a density takes the form

$$p(x|\mu) = f(x-\mu) \tag{2.138}$$

then the parameter μ is known as a *location parameter*. This family of densities exhibits *translation invariance* between if we shift x by a constant to give $\widehat{x} = x + c$, then

$$p(\widehat{x}|\widehat{\mu}) = f(\widehat{x} - \widehat{\mu}) \tag{2.139}$$

where we have defined $\widehat{\mu} = \mu + c$. Thus the density takes the same form in the new variable as in the original one, and so the density is independent of the choice of origin. We would like to choose a prior distribution that reflects this translation invariance property, which can be proved $p(\mu)$ is constant. An example of a location parameter would be the mean μ of a Gaussian distribution. As we have seen, the conjugate prior distribution for μ in this case is a Gaussian $p(\mu|\mu_0, \sigma_0^2) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$, and we obtain a noninformative prior by taking the limit $\sigma_0^2 \to \infty$.

As a second example, consider a density of the form

$$p(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right) \tag{2.140}$$

where $\sigma > 0$. The parameter σ is known as a *scale parameter*, and the density exhibits *scale invariance* because if we scale x by a constant to give $\widehat{x} = cx$, then

$$p(\widehat{x}|\widehat{\sigma}) = \frac{1}{\widehat{\sigma}} f\left(\frac{\widehat{x}}{\widehat{\sigma}}\right) \tag{2.141}$$

where we have defined $\widehat{\sigma} = c\sigma$. We can use this property to prove $p(\sigma) \propto 1/\sigma$. Note that again this is an improper prior because its integral is divergent. It is sometimes also convenient to think of the prior distribution for a scale parameter in terms of the density of the log of the parameter, so $p(\ln \sigma) = \text{const}$. An example of a scale parameter would be the standard deviation σ of a Gaussian distribution.

2.5. Nonparametric Methods

Throughout this chapter, we have focussed on the use of probability distributions having specific functional forms governed by a small number of parameters whose values are to be determined from a data set. This is called the *parametric* approach to density modeling. An important limitation of this approach is that the chosen density might be a poor model of the distribution that generates the data, which can result in poor predictive performance. For instance, if the process that generates the data is multimodal, then this aspect of the distribution can never be captured by a Gaussian, which is necessarily unimodal.

In this section, we consider some *nonparametric* approaches to density estimation that make few assumptions about the form of the distribution. Here we shall focus mainly on simple frequentist methods. The reader should be aware, however, that nonparametric Bayesian methods are attracting increasing interest.

The histogram method simply partition random variable x into distinct bins of width Δ_i and then count the number n_i of observations of x falling in bin i. The probability

values for each bin is simply given by

$$p_i = \frac{n_i}{N\Delta_i} \tag{2.142}$$

This gives a model for the density $p(x)$ that is constant over the width of each bin, and often the bins are chosen to have the same width $\Delta_i = \Delta$. The result of histogram density model is highly dependent on choosing the best value of Δ, not too large or too small.

In practice, the histogram technique can be useful for obtaining a quick visualization of data in one or two dimensions but is unsuited to most density estimation applications.

The histogram approach to density estimation does, however, teach us two important lessons. First, to estimate the probability density at a particular location, we should consider the data points that lie within some local neighborhood of that point. For histograms, this neighborhood property was defined by the bins, and there is a natural 'smoothing' parameter describing the spatial extent of the local region, in this case the bin width. Second, the value of the smoothing parameter should be neither too large nor too small in order to obtain good results.

2.5.1. Kernel density estimators

Let us suppose observations are being drawn from some unknown probability density $p(\mathbf{x})$ in Euclidean D-dimensional space, and we wish to estimate the value of $p(\mathbf{x})$. From our earlier discussion of locality, let us consider some small region $\mathcal{R}$ containing $\mathbf{x}$. The probability mass associated with this region is given by

$$P = \int_{\mathcal{R}} p(\mathbf{x})\mathrm{d}\mathbf{x} \tag{2.143}$$

Now suppose that we have collected a data set comprising N observations drawn from $p(\mathbf{x})$. Because each data point has a probability P of falling within $\mathcal{R}$, the total number K of points that lie inside $\mathcal{R}$ will be distributed according to the binomial distribution

$$\mathrm{Bin}(K|N,P) = \frac{N!}{K!(N-K)!}P^K(1-P)^{1-K} \tag{2.144}$$

Using (2.7), we see that the mean fraction of points falling inside the region is $\mathbb{E}[K/N] = P$, and similarly using (2.8) the variance around this mean is $\mathrm{var}[K/N] = P(1-P)/N$. For large N, this distribution will be sharply peaked around the mean and so

$$K \simeq NP \tag{2.145}$$

If, however, we also assume that the region $\mathcal{R}$ is sufficiently small that the probability density $p(\mathbf{x})$ is roughly constant over the region, then we have

$$P \simeq p(\mathbf{x})V \tag{2.146}$$

when V is the volume of $\mathcal{R}$. Combining (2.145) and (2.146), we obtain density estimate in the form

$$p(\mathbf{x}) = \frac{K}{NV} \tag{2.147}$$

Note that the validity of (2.147) depends on two contradictory assumptions, namely that the region $\mathcal{R}$ be sufficiently small that the density is approximately constantly over the region and yet sufficiently large (in relation to the value of that density) that the number K of points falling inside the region is sufficient for the binomial distribution to be sharply peaked.

We can exploit the result (2.147) in two different ways. Either we can fix K and determine the value of V from the data, which gives rise to the Knearest-neighbor technique discussed shortly, or we can fix V and determine K from the data, giving rise to the kernel approach.

We begin by discussing the kernel method in detail, and to start with we take the region $\mathcal{R}$ to be a small hypercube centered on the point $\mathbf{x}$ at which we wish to determine the probability density. In order to count the number K of points falling within the region, it is convenient to define the following function

$$k(\mathbf{u}) = \begin{cases} 1, & |u_i| \leqslant 1/2, \quad i = 1, \ldots, D \\ 0, & \text{otherwise} \end{cases} \tag{2.148}$$

which represents a unit cube centered on the origin. The function $k(\mathbf{u})$ is an example of a *kernel function*, and in this context is also called a *Parzen window*. From (2.148), the total number of data points lying inside this cube of side h will be

$$K = \sum_{n=1}^{N} k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right) \tag{2.149}$$

Substituting this expression into (2.147) then gives the following result for the estimated density at $\mathbf{x}$

$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{h^D} k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right) \tag{2.150}$$

where we have used $V = h^D$. Using the symmetry of the function $k(\mathbf{u})$, we can now re-interpret this equation, not as a single cube centered on $\mathbf{x}$ but as the sum over N cubes centered on the N data points $\mathbf{x}_n$.

The kernel density estimator (2.150) will suffer from one of the same problems that the histogram method suffered from, namely the presence of artificial discontinuities, in this case at the boundaries of the cubes. We can obtain a smoother density model if we choose a smoother kernel function, and a common choice is the Gaussian, which gives rise to the following kernel density model

$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{1/2}} \exp\left\{-\frac{\|\mathbf{x} - \mathbf{x}_n\|^2}{2h^2}\right\} \tag{2.151}$$

where h represents the standard deviation of the Gaussian components. Thus our density model is obtained by placing a Gaussian over each data point and then adding up the contributions over the whole data set, and then dividing by N so that the density is correctly normalized.

We can choose any other other kernel function $k(\mathbf{u})$ in (2.150) subject to the conditions

$$k(\mathbf{u}) \geqslant 0 \tag{2.152}$$

$$\int k(\mathbf{u})\mathrm{d}\mathbf{u} = 1 \tag{2.153}$$

which ensure that the resulting probability distribution is nonnegative everywhere and integrates to one. The class of density model given by (2.150) is called a kernel density estimator, or *Parzen* estimator. It has a great merit that there is no computation involved in the 'training' phase because this simply requires storage of the training set. However, this is also one of its great weaknesses because the computational cost of evaluating the density grows linearly with the size of the data set.

2.5.2. Nearest-neighbor methods

One of the difficulties with the kernel approach to density estimation is that the parameter h governing the kernel width is fixed for all kernels. In regions of high data density, a large value of h may lead to over-smoothing and a washing out of structure that might otherwise be extracted from the data. However, reducing h may lead to noisy estimates elsewhere in data space where the density is smaller. Thus the optimal choice for h may be dependent on location within the data space. This issue is addressed by nearest-neighbor methods for density estimation.

We therefore return to our general result (2.147) for local density estimation, and instead of fixing V and determining the value of K from the data, we consider a fixed value of K and use the data to find an appropriate value of V. To do this, we consider a small sphere centered on the point $\mathbf{x}$ at which we wish to estimate the density $p(\mathbf{x})$, and we allow the radius of the sphere to grow until it contains precisely K data points. The estimate of the density $p(\mathbf{x})$ is then given by (2.147) with V set to the volume of the resulting sphere. This technique is known as K *nearest neighbors*. We see that the value of K now governs the degree of smoothing and that again there is an optimum choice for K that is neither too large nor too small. Note that the model produced by K nearest neighbors is not a true density model because the integral over all space diverges.

The K-nearest-neighbor technique for density estimation can be extended to the problem of classification. To do this, we apply the K-nearest-neighbor density estimation technique to each class separately and then make use of Bayes theorem. Suppose that we have a data set comprising N_k point in class $\mathcal{C}_k$ with N points in total. If we wish to classify a new point $\mathbf{x}$, we draw a sphere centered on $\mathbf{x}$ containing precisely K points irrespective of their class. Suppose this sphere has volume V and contains K_k points from class $\mathcal{C}_k$. Then (2.147) provides an estimate of the density associated with each class

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{K_k}{N_k V} \tag{2.154}$$

Combining with the unconditional density (2.147), and the class priors are given by

$$p(\mathcal{C}_k) = \frac{N_k}{N} \tag{2.155}$$

we can use Bayes' theorem to obtain the posterior probability of class membership

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} = \frac{K_k}{K} \tag{2.156}$$

K controls the degree of smoothing, so that small K produces many small regions of each class, whereas large K leads to fewer larger regions. The particular case of $K = 1$ is called the nearest-neighbor rule, because a test point is simply assigned to the same class as the nearest point from the training set.

As discussed so far, both the K-nearest-neighbor method, and the kernel density estimator, require the entire training data set to be stored, leading to expensive computation if the data set is large. This effect can be offset, at the expense of some additional one-off computation, by constructing tree-based search structures to allow (approximate) near neighbors to be found efficiently without doing an exhaustive search of the data set. Nevertheless, these nonparametric methods are still severely limited. On the other hand, we have seen that simple parametric models are very restricted in terms of the forms of distribution that they can represent. We therefore need to find density models that are very flexible and yet for which the complexity of the models can be controlled independently of the size of the training set, and we shall see in subsequent chapters how to achieve this.

A. Data Sets

TODO

B. Probability Distributions

TODO

C. Properties of Matrices

Basic Matrix Identities

A matrix $\mathbf{A}$ has elements A_{ij} where i indexes the rows, and j indexes the columns. We use $\mathbf{I}_N$ to denote the $N \times N$ identity matrix (also called the unit matrix), and where there is no ambiguity over dimensionality we simply use $\mathbf{I}$. The transposes matrix $\mathbf{A}^{\mathrm{T}}$ has elements $(\mathbf{A}^{\mathrm{T}})_{ij} = A_{ji}$. From the definition of transpose, we have

$$(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} \tag{C.1}$$

The inverse of $\mathbf{A}$, denoted $\mathbf{A}^{-1}$, satisfies $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. Because $\mathbf{ABB}^{-1}\mathbf{A}^{-1} = \mathbf{I}$, we have

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} \tag{C.2}$$

Also we have

$$(\mathbf{A}^{\mathrm{T}})^{-1} = (\mathbf{A}^{-1})^{\mathrm{T}} \tag{C.3}$$

A useful identity involving matrix inverse is the following

$$(\mathbf{P}^{-1} + \mathbf{B}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{R}^{-1} = \mathbf{P}\mathbf{B}^{\mathrm{T}}(\mathbf{B}\mathbf{P}\mathbf{B}^{\mathrm{T}} + \mathbf{R})^{-1} \tag{C.4}$$

which is easily verified by right multiplying both sides by $(\mathbf{BPB}^{\mathrm{T}} + \mathbf{R})$. Suppose that $\mathbf{P}$ has dimensionality $N \times N$ while $\mathbf{R}$ has dimensionality $M \times M$, so that $\mathbf{B}$ is $M \times N$. Then if $M \ll N$, it will be much cheaper to evaluate the right-hand side of (C.4) than the left-hand side.

A set of vectors $\{\mathbf{a}_1, \ldots, \mathbf{a}_N\}$ is said to be *linearly independent* if the relation $\sum_n \alpha_n \mathbf{a}_n = \mathbf{0}$ holds only if all $\alpha_n = 0$. This implies that none of the vectors can be expressed as a linear combination of the remainder. The rank of a matrix is the maximum number of linearly independent rows (or equivalently the maximum number of linearly independent columns).

Traces and Determinants

Trace and determinant apply to square matrixes. The trace $\mathrm{Tr}(\mathbf{A})$ of a matrix $\mathbf{A}$is defined as the sum of the elements on the leading diagonal. By writing out the indices, we see that

$$\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA}) \tag{C.5}$$

By applying this formula multiple times to the product of three matrices, we see that

$$\mathrm{Tr}(\mathbf{ABC}) = \mathrm{Tr}(\mathbf{CAB}) = \mathrm{Tr}(\mathbf{BCA}) \tag{C.6}$$

which is known as the *cyclic* property of the trace operator and which clearly extends to the product of any number of matrices.

The determinant $|\mathbf{A}|$ of an $N \times N$ matrix $\mathbf{A}$ is defined by

$$|\mathbf{A}| = \sum (\pm 1) A_{1i_1} A_{2i_2} \cdots A_{Ni_N} \tag{C.7}$$

in which the sum is taken over all products consisting of precisely one element from each row and one element from each column, with a coefficient $+1$ or -1 according to whether the permutation $i_1 i_2 \ldots i_N$ is even or odd, respectively. Note that $|\mathbf{I}| = 1$.

The determinant of a product of two matrices is given by

$$|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}| \tag{C.8}$$

Also, the determinant of an inverse matrix is given by

$$|\mathbf{A}|^{-1} = \frac{1}{|\mathbf{A}|} \tag{C.9}$$

If $\mathbf{A}$ and $\mathbf{B}$ are matrices of size $N \times M$, then

$$\left|\mathbf{I}_N + \mathbf{A}\mathbf{B}^{\mathrm{T}}\right| = \left|\mathbf{I}_M + \mathbf{A}^{\mathrm{T}}\mathbf{B}\right| \tag{C.10}$$

A useful special case is

$$\left|\mathbf{I}_N + \mathbf{a}\mathbf{b}^{\mathrm{T}}\right| = 1 + \mathbf{a}^{\mathrm{T}}\mathbf{b} \tag{C.11}$$

where $\mathbf{a}$ and $\mathbf{b}$ are N-dimensional column vectors.

Matrix Derivatives

The derivative of a vector $\mathbf{a}$ with respect to a scalar x is itself a vector whose components are given by

$$\left(\frac{\partial \mathbf{a}}{\partial x}\right)_i = \frac{\partial a_i}{\partial x} \tag{C.12}$$

with an analogous definition for the derivative of a matrix. Derivatives with respect to vectors and matrices can also be defined, for instance

$$\left(\frac{\partial x}{\partial \mathbf{a}}\right)_i = \frac{\partial x}{\partial a_i} \tag{C.13}$$

and similarly

$$\left(\frac{\partial \mathbf{a}}{\partial \mathbf{b}}\right)_{ij} = \frac{\partial a_i}{\partial a_j} \tag{C.14}$$

The following is easily proven by writing out the components

$$\frac{\partial}{\partial \mathbf{x}}\left(\mathbf{x}^{\mathrm{T}}\mathbf{a}\right) = \frac{\partial}{\partial \mathbf{x}}\left(\mathbf{a}^{\mathrm{T}}\mathbf{x}\right) = \mathbf{a} \tag{C.15}$$

Similarly

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{AB}) = \frac{\partial \mathbf{A}}{\partial \mathbf{x}}\mathbf{B} + \mathbf{A}\frac{\partial \mathbf{B}}{\partial \mathbf{x}} \tag{C.16}$$

The derivative of the inverse of a matrix can be expressed as

$$\frac{\partial}{\partial x}\left(\mathbf{A}^{-1}\right) = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\mathbf{A}^{-1} \tag{C.17}$$

Also

$$\frac{\partial}{\partial x}\ln|\mathbf{A}| = \mathrm{Tr}\left(\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\right) \tag{C.18}$$

If we choose x to be one of the elements of $\mathbf{A}$, we have

$$\frac{\partial}{\partial A_{ij}}\mathrm{Tr}(\mathbf{AB}) = B_{ji} \tag{C.19}$$

We can write this result more compactly in the form

$$\frac{\partial}{\partial \mathbf{A}}\mathrm{Tr}(\mathbf{AB}) = \mathbf{B}^{\mathrm{T}} \tag{C.20}$$

With this notation, we have the following properties

$$\frac{\partial}{\partial \mathbf{A}}\mathrm{Tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{B}\right) = \mathbf{B} \tag{C.21}$$

$$\frac{\partial}{\partial \mathbf{A}}\mathrm{Tr}(\mathbf{A}) = \mathbf{I} \tag{C.22}$$

$$\frac{\partial}{\partial \mathbf{A}}\mathrm{Tr}(\mathbf{ABA}^{\mathrm{T}}) = \mathbf{A}(\mathbf{B} + \mathbf{B}^{\mathrm{T}}) \tag{C.23}$$

We also have

$$\frac{\partial}{\partial \mathbf{A}}\ln|\mathbf{A}| = \left(\mathbf{A}^{-1}\right)^{\mathrm{T}} \tag{C.24}$$

Eigenvector Equation

For a square matrix $\mathbf{A}$ with size $M \times M$, the eigenvector equation is defined by

$$\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i \tag{C.25}$$

for $i = 1, \ldots, M$, where $\mathbf{u}_i$ is an *eigenvector* and λ_i is the corresponding *eigenvalue*. This can be viewed as a set of M simultaneous homogeneous linear equations, and the condition for a solution is that

$$|\mathbf{A} - \lambda_i \mathbf{I}| = 0 \tag{C.26}$$

which is known as the *characteristic equation*. Because this is a polynomial of order M in λ_i, it must have M solutions (through these need not all be distinct). the rank of $\mathbf{A}$ is equal to the number of nonzero eigenvalues.

Of particular interest are symmetric matrices, which arise as covariance matrices, kernel matrices, and Hessians. In general, the eigenvalues of a matrix are complex numbers, but for symmetric matrices the eigenvalues λ_i are real. The eigenvectors $\mathbf{u}_i$ of a real symmetric matrix can be chosen to be orthonormal (i.e., orthogonal and of unit length) so that

$$\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = I_{ij} \tag{C.27}$$

where I_{ij} are the elements of the identity matrix $\mathbf{I}$.

We can take the eigenvectors $\mathbf{u}_i$ to be the columns of an $M \times M$ matrix $\mathbf{U}$, which from orthonormality satisfies

$$\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I} \tag{C.28}$$

Such a matrix is said to be *orthogonal*. Interestingly, the rows of this matrix are also orthogonal, so that $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}$, it also follows that $\mathbf{U}^{-1} = \mathbf{U}^{\mathrm{T}}$ and $|\mathbf{U}| = 1$.

The eigenvector equation (C.25) can be expressed in terms of $\mathbf{U}$ in the form

$$\mathbf{A}\mathbf{U} = \mathbf{U}\mathbf{\Lambda} \tag{C.29}$$

where $\mathbf{\Lambda}$ is an $M \times M$ diagonal matrix whose diagonal elements are given by the eigenvalues λ_i.

If we consider a column vector $\mathbf{x}$ that is transformed by an orthogonal matrix $\mathbf{U}$ to give a new vector

$$\widetilde{\mathbf{x}} = \mathbf{U}\mathbf{x} \tag{C.30}$$

then the length of the vector is preserved because

$$\widetilde{\mathbf{x}}^{\mathrm{T}}\widetilde{\mathbf{x}} = \mathbf{x}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{x} \tag{C.31}$$

and similarly the angle between any two such vectors is preserved because

$$\widetilde{\mathbf{x}}^{\mathrm{T}}\widetilde{\mathbf{y}} = \mathbf{x}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{y} = \mathbf{x}^{\mathrm{T}}\mathbf{y} \tag{C.32}$$

Thus, multiplication by $\mathbf{U}$ can be interpreted as a rigid rotation of the coordinate system.

From (C.29), it follows that

$$\mathbf{U}^{\mathrm{T}}\mathbf{A}\mathbf{U} = \mathbf{\Lambda} \tag{C.33}$$

and because $\mathbf{\Lambda}$ is a diagonal matrix, we say that the matrix $\mathbf{A}$ is *diagonalized* by the matrix $\mathbf{U}$. If we left multiply by $\mathbf{U}$ and right multiply by $\mathbf{U}^{\mathrm{T}}$, we obtain

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}} \tag{C.34}$$

Taking the inverse of this equation, and use $\mathbf{U}^{-1} = \mathbf{U}^{\mathrm{T}}$, we have

$$\mathbf{A}^{-1} = \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^{\mathrm{T}} \tag{C.35}$$

These last two equations can be written in the form

$$\mathbf{A} = \sum_{i=1}^{M} \lambda_i \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}} \tag{C.36}$$

$$\mathbf{A}^{-1} = \sum_{i=1}^{M} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}} \tag{C.37}$$

If take the determinant of (C.34), we obtain

$$|\mathbf{A}| = \prod_{i=1}^{M} \lambda_i \tag{C.38}$$

Similarly, taking the trace of (C.34), we have

$$\mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^{M} \lambda_i \tag{C.39}$$

A matrix $\mathbf{A}$ is said to be *positive definite*, denoted by $\mathbf{A} \succ 0$, is $\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w} > 0$ for all values of the vector $\mathbf{w}$. Equivalently, a positive definite matrix has $\lambda_i > 0$ for all of its eigenvalues (as can be seen by setting $\mathbf{w}$ to each of the eigenvectors in turn, and by noting that an arbitrary vector can be expanded as a linear combination of the eigenvectors). Note that positive definite is note the same as all the elements being positive. A matrix is said to be *positive semidefinite* if $\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w} \geqslant 0$ holds for all values of $\mathbf{w}$, which is denoted $\mathbf{A} \succeq 0$, and is equivalent to $\lambda_i \geqslant 0$.

D. Calculus of Variations

TODO

E. Lagrange Multipliers

TODO