Gaussians

Hiroshi Shimodaira[*]

January-March 2020

In this chapter we introduce the basics of how to build probabilistic models of continuous-valued data, including the most important probability distribution for continuous data: the Gaussian, or Normal, distribution. We discuss both the univariate Gaussian (the Gaussian distribution for one-dimensional data) and the multivariate Gaussian distribution (the Gaussian distribution for multi-dimensional data).

8.1 Continuous random variables

First we review the concepts of the cumulative distribution function and the probability density function of a continuous random variable.

Many events that we want to model probabilistically are described by real numbers rather than discrete symbols or integers. In this case we must use *continuous random variables*. Some examples of continuous random variables include:

- The sum of two numbers drawn randomly from the interval $0 < x < 1$;
- The length of time for a bus to arrive at the bus-stop;
- The height of a member of a population.

8.1.1 Cumulative distribution function

We will develop the way we model continuous random variables using a simple example.

Consider waiting for a bus, which runs every 30 minutes. We shall make the idealistic assumption that the buses are always exactly on time, thus a bus arrives every 30 minutes. If you are waiting for a bus, but don't know the timetable, then the precise length of time you need to wait is unknown. Let the continuous random variable X denote the length of time you need to wait for a bus.

Given the above information we may assume that X never takes values above 30 or below 0. We can write this as:

$$\begin{aligned} P(X < 0) &= 0 \\ P(0 \le X \le 30) &= 1 \\ P(X > 30) &= 0 \end{aligned}$$

[*] ©2014-2020 University of Edinburgh. All rights reserved. This note is heavily based on notes inherited from Steve Renals and Iain Murray.

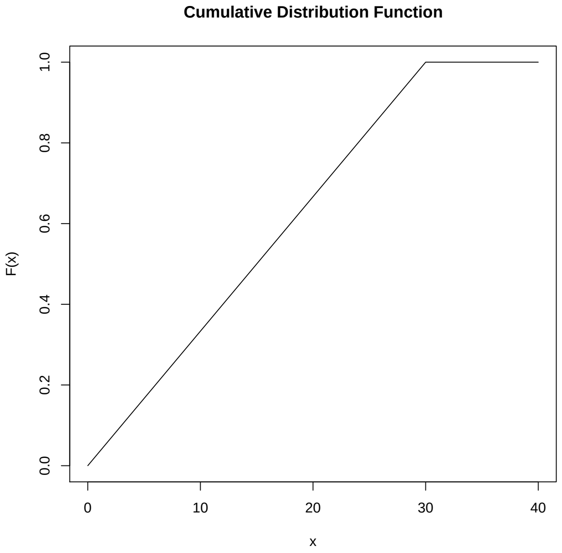


Figure 8.1: Cumulative distribution function of random variable X in the 'bus' example.

The probability distribution function for a random variable assigns a probability to each value that the variable may take. It is impossible to write down a probability distribution function for a continuous random variable X, since $P(X = x) = 0$ for all x. This is because X is continuous and can take infinitely many values (and $1/\infty = 0$). However we can write down a *cumulative distribution* $F(X)$, which gives the probability of X taking a value that is less than or equal to x. For the current example:

$$F(x) = P(X \le x) = \begin{cases} 0 & x < 0 \\ (x-0)/30 = x/30 & 0 \le x \le 30 \\ 1 & x > 30 \end{cases} \tag{8.1}$$

In writing down this cumulative distribution, we have made the (reasonable) assumption that the probability of a bus arriving increases in proportion to the interval of time waited (in the region 0–30 minutes). This cumulative density function is plotted in Figure 8.1.

Cumulative distribution functions have the following properties:

(i) $F(-\infty) = 0$;

(ii) $F(\infty) = 1$;

(iii) If $a \le b$ then $F(a) \le F(b)$.

To obtain the probability of falling in an interval we can do the following:

$$P(a < X \le b) = P(X \le b) - P(X \le a) = F(b) - F(a). \tag{8.2}$$

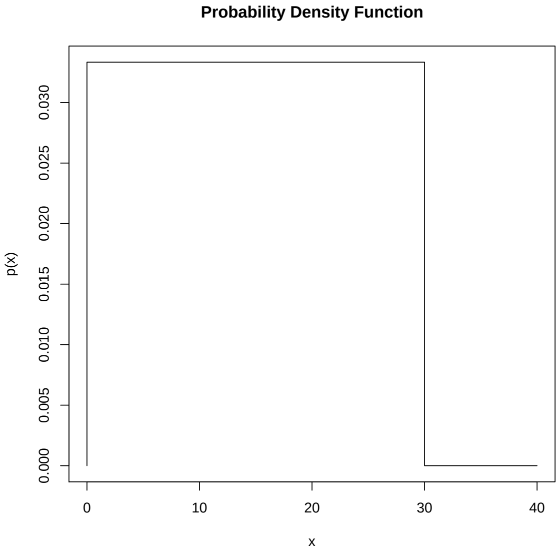


Figure 8.2: Probability density function of random variable X in the 'bus' example.

For the 'bus' example:

$$\begin{aligned} P(15 < X \leq 21) &= F(21) - F(15) \\ &= 0.7 - 0.5 = 0.2 \end{aligned}$$

8.1.2 Probability density function

Although we cannot define a probability distribution function for a continuous random variable, we can define a closely related function, called the *probability density function* (pdf), $p(x)$:

$$\begin{aligned} p(x) &= \frac{\mathrm{d}}{\mathrm{d}x} F(x) = F'(x) \\ F(x) &= \int_{-\infty}^{x} p(x)\, \mathrm{d}x\,. \end{aligned}$$

The pdf is the gradient of the cdf. Note that $p(x)$ is *not* the probability that X has value x. However, the pdf is proportional to the probability that X lies in a small interval centred on x. The pdf is the usual way of describing the probabilities associated with a continuous random variable X. We usually write probabilities using upper case P and probability densities using lower case p.

We can immediately write down the pdf for the 'bus' example:

$$p(x) = \begin{cases} 0 & x < 0 \\ 1/30 & 0 \leq x \leq 30 \\ 0 & x > 30 \end{cases}$$

Figure 8.2 shows a graph of this pdf. X is said to be *uniform* on the interval $(0, 30)$.

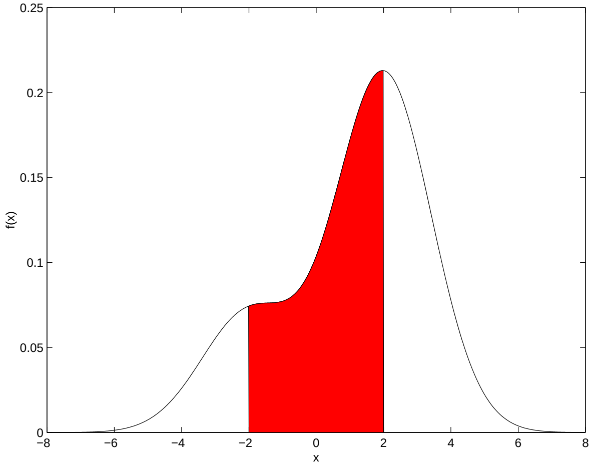


Figure 8.3: $P(-2 < X \leq 2)$ is the shaded area under the pdf.

The probability that the random variable lies in interval (a, b) is the area under the pdf between a and b:

$$\begin{aligned} P(a < X \leq b) &= F(b) - F(a) \\ &= \int_{-\infty}^{b} p(x)\, \mathrm{d}x - \int_{-\infty}^{a} p(x)\, \mathrm{d}x \\ &= \int_{a}^{b} p(x)\, \mathrm{d}x\,. \end{aligned}$$

This integral is illustrated in Figure 8.3. The total area under the pdf equals 1, the probability that x takes on some value between $-\infty$ and ∞. We can also confirm that $F(\infty) - F(-\infty) = 1 - 0 = 1$.

8.2 The Gaussian distribution

The *Gaussian* (or *Normal*) distribution is the most commonly encountered (and easily analysed) continuous distribution. It is also a reasonable model for many situations (the famous 'bell curve').

If a (scalar) variable has a Gaussian distribution, then it has a probability density function with this form:

$$p(x|\mu, \sigma^2) = N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right). \tag{8.3}$$

The Gaussian is described by two parameters:

- the mean μ (location)
- the variance σ^2 (dispersion)

We write $p(x|\mu, \sigma^2)$ because the pdf of x depends on the parameters. Sometimes (to slim down the notation) we simply write $p(x)$.

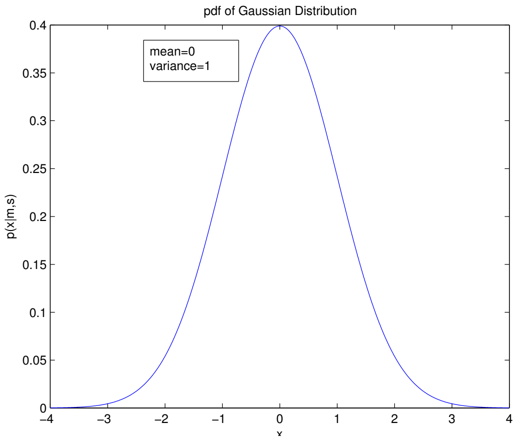


Figure 8.4: One dimensional Gaussian ($\mu = 0$, $\sigma^2 = 1$)

All Gaussians have the same shape, with the location controlled by the mean, and the dispersion (horizontal scaling) controlled by the variance.[1] Figure 8.4 shows a one-dimensional Gaussian with zero mean and unit variance ($\mu = 0$, $\sigma^2 = 1$.)

In Equation (8.3), the mean occurs in the exponential part of the pdf, $\exp(-0.5(x-\mu)^2/\sigma^2)$. This term will have a maximum ($\exp(0) = 1$) when $x = \mu$; thus the peak of the Gaussian corresponds to the mean, and we can think of it as the location parameter.

In one dimension, the variance can be thought of as controlling the width of the Gaussian pdf. Since the area under the pdf must equal 1, this means that the wide Gaussians have lower peaks than narrow Gaussians. This explains why the variance occurs twice in the formula for a Gaussian. In the exponential part $\exp(-0.5(x-\mu)^2/\sigma^2)$, the variance parameter controls the width: for larger values of σ^2, the value of the exponential decreases more slowly as x moves away from the mean. The term $1/\sqrt{2\pi\sigma^2}$ is the *normalisation* constant, which scales the whole pdf to ensure that it integrates to 1. This term decreases with σ^2: hence as σ^2 decreases so the pdf gets a taller (but narrower) peak. The behaviour of the pdf with respect to the variance parameter is shown in Figure 8.5.

[1] To be precise, the width of a distribution scales with its standard deviation, σ, i.e. the square root of the variance.

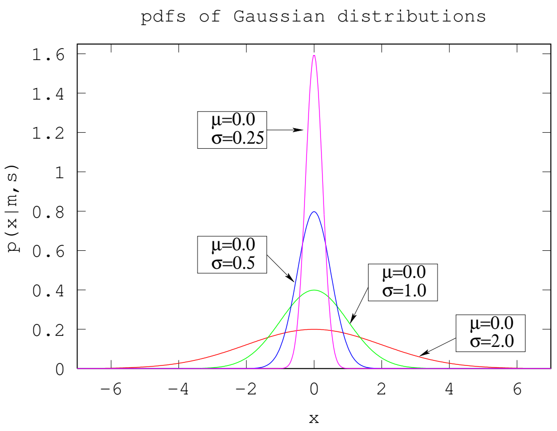


Figure 8.5: Four Gaussian pdfs with zero mean and different standard deviations.

8.3 Parameter estimation

A Gaussian distribution has two parameters the mean (μ) and the variance(σ^2). In machine learning or pattern recognition we are not given the parameters, we have to estimate them from data. As in the case of Naive Bayes we can choose the parameters such that they maximise the likelihood of the model generating the training data. In the case of the Gaussian distribution the *maximum likelihood estimate* (MLE) of the mean and the variance[2] results in:

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n\,, \tag{8.4}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2\,, \tag{8.5}$$

where x_n denotes the feature value of n'th sample, and N is the number of samples in total.

8.3.1 Maximum likelihood parameter estimation for Gaussian distribution

The two formulas Equation (8.4) and Equation (8.5) are very popular, but it is not a good practice that we just remember them without understanding how they are derived in the context of Gaussian distribution.

We here consider the parameter estimation as an optimisation problem:

$$\max_{\mu,\sigma^2} p(x_1, \ldots, x_N \,|\, \mu, \sigma^2)\,, \tag{8.6}$$

where we try to find such μ and σ^2 that maximise the likelihood. Note that this likelihood is a function of μ and σ^2, and not of the data, since the data are fixed - they are given and do not change. To

[2] The maximum likelihood estimate of the variance turns out to be a biased form of the sample variance that is normalised by $N-1$ rather than N.

make the optimisation problem tractable, we introduce an assumption that all the training samples are independent from each other, so that the optimisation problem is simplified to

$$\max_{\mu,\sigma^2} p(x_1|\mu,\sigma^2)\cdots p(x_N|\mu,\sigma^2) \tag{8.7}$$

Applying the natural log to the likelihood and letting it denoted by $LL(\mu,\sigma^2)$,

$$\begin{aligned} LL(\mu,\sigma^2) &= \ln\left(p(x_1|\mu,\sigma^2)\cdots p(x_N|\mu,\sigma^2)\right) & (8.8)\\ &= \sum_{n=1}^{N} \ln p(x_1|\mu,\sigma^2) & (8.9)\\ &= \sum_{n=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x_n-\mu)^2}{2\sigma^2}\right)\right) & (8.10)\\ &= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\left(\sigma^2\right) - \sum_{n=1}^{N}\frac{(x_n-\mu)^2}{2\sigma^2} & (8.11) \end{aligned}$$

As we studied in Section 5.5 in Note 5, we can find the optimal parameters of this unconstrained optimisation problem by solving the following system of equations:

$$\frac{\partial LL(\mu,\sigma^2)}{\partial\mu} = 0 \tag{8.12}$$

$$\frac{\partial LL(\mu,\sigma^2)}{\partial\sigma^2} = 0 \tag{8.13}$$

You can easily confirm that Equation (8.4) and Equation (8.5) are the solutions.

8.4 Example

A pattern recognition problem has two classes, S and T. Some observations are available for each class:

Class S:	10	8	10	10	11	11
Class T:	12	9	15	10	13	13

We assume that each class may be modelled by a Gaussian. Using the above data, estimate the parameters of the Gaussian pdf for each class, and sketch the pdf for each class.

The mean and variance of each pdf are estimated with MLE shown in Equation (8.4) and Equation (8.5).

$$\begin{aligned} \hat{\mu}_S &= \frac{(10+8+10+10+11+11)}{6} = 10\\ \hat{\sigma}^2_S &= \frac{(10-10)^2+(8-10)^2+(10-10)^2+(10-10)^2+(11-10)^2+(11-10)^2}{6} = 1\\ \hat{\mu}_T &= \frac{(12+9+15+10+13+13)}{6} = 12\\ \hat{\sigma}^2_T &= \frac{(12-12)^2+(9-12)^2+(15-12)^2+(10-12)^2+(13-12)^2+(13-12)^2}{6} = 4 \end{aligned}$$

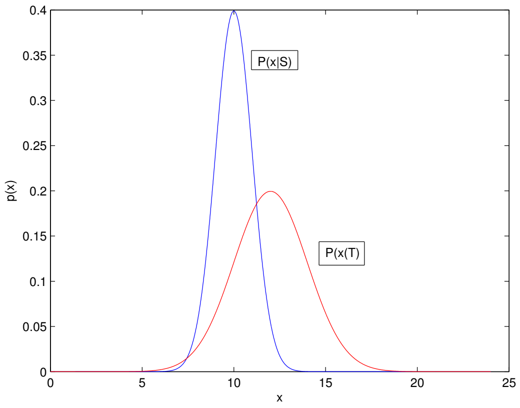


Figure 8.6: Estimated Gaussian pdfs for class S ($\hat{\mu}=10$, $\hat{\sigma}^2=1$) and class class T ($\hat{\mu}=12$, $\hat{\sigma}^2=4$)

The process of estimating the parameters from the training data is sometimes referred to as *fitting* the distribution to the data.

Figure 8.6 shows the pdfs for each class. The pdf for class T is twice the width of that for class S: the width of a distribution scales with its standard deviation, not its variance.

8.5 The multivariate Gaussian distribution and covariance

The univariate (one-dimensional) Gaussian may be extended to the multivariate (multi-dimensional) case. The D-dimensional Gaussian is parameterised by a mean vector, $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_D)^T$, and a *covariance matrix*[3], $\boldsymbol{\Sigma} = (\sigma_{ij})$, and has a probability density

$$p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{8.14}$$

The 1-dimensional Gaussian is a special case of this pdf. The covariance matrix gives the variance of each variable (dimension) along the leading diagonal, and the off-diagonal elements measure the correlations between the variables. The argument to the exponential $\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ is an example of a *quadratic form*. $|\boldsymbol{\Sigma}|$ is the determinant of the covariance matrix $\boldsymbol{\Sigma}$.

The mean vector $\boldsymbol{\mu}$ is the expectation of $\mathbf{x}$:

$$\boldsymbol{\mu} = E[\mathbf{x}]\,.$$

The covariance matrix $\boldsymbol{\Sigma}$ is the expectation of the deviation of $\mathbf{x}$ from the mean:

$$\boldsymbol{\Sigma} = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T]\,. \tag{8.15}$$

From Equation (8.15) it follows that $\boldsymbol{\Sigma} = (\sigma_{ij})$ is a $D\times D$ symmetric matrix; that is $\boldsymbol{\Sigma}^T = \boldsymbol{\Sigma}$:

$$\sigma_{ij} = E[(x_i-\mu_i)(x_j-\mu_j)] = E[(x_j-\mu_j)(x_i-\mu_i)] = \sigma_{ji}\,.$$

[3] $\boldsymbol{\Sigma}$ is a D-by-D square matrix, and σ_{ij} or Σ_{ij} denotes its element at i'th row and j'th column.

Note that σ_{ij} denotes not the standard deviation but the *covariance* between i'th and j'th elements of $\mathbf{x}$. For example, in the 1-dimensional case, $\sigma_{11} = \sigma^2$.

It is helpful to consider how the covariance matrix may be interpreted. The sign of the covariance, σ_{ij}, helps to determine the relationship between two components:

- If x_j is large when x_i is large, then $(x_j - \mu_j)(x_i - \mu_i)$ will tend to be positive;[4]
- If x_j is small when x_i is large, then $(x_j - \mu_j)(x_i - \mu_i)$ will tend to be negative.

If variables are highly correlated (large covariance) then this may indicate that one does not give much extra information if the other is known. If two components of the input vector, x_i and x_j, are statistically independent then the covariance between them is zero, $\sigma_{ij} = 0$.

Correlation coefficient The values of the elements of the covariance matrix depend on the unit of measurement: consider the case when x is measured in metres, compared when x is measured in millimetres. It is useful to define a measure of dispersion that is independent of the unit of measurement. To do this we may define the *correlation coefficient*[5] between features x_i and x_j, $\rho(x_i, x_j)$:

$$\rho(x_i, x_j) = \rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}\,. \tag{8.16}$$

The correlation coefficient ρ_{ij} is obtained by normalising the covariance σ_{ij} by the square root of the product of the variances σ_{ii} and σ_{jj}, and satisfies $-1 \le \rho_{ij} \le 1$:

$$\begin{aligned} \rho(x,y) &= +1 \qquad \text{if } y = ax + b \quad a > 0 \\ \rho(x,y) &= -1 \qquad \text{if } y = ax + b \quad a < 0\,. \end{aligned}$$

The correlation coefficient is both scale-invariant and location(or shift)-invariant, i.e.:

$$\rho(x_i, x_j) = \rho(ax_i + b, cx_j + d)\,. \tag{8.17}$$

where $a > 0$, $c > 0$, and c and d are arbitrary constants.

8.6 The 2-dimensional Gaussian distribution

Let's look at a two dimensional case, with the following *inverse* covariance matrix[6]:

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21}} \begin{pmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}.$$

[4] Note that x_i in this section denotes the i'th element of a vector $\mathbf{x}$ (which is a vector of D random variables) rather than the i'th sample in a data set $\{x_1, \ldots, x_N\}$.

[5] This is normally referred as 'Pearson's correlation coefficient', whose another version for sampled data was discussed in Note 2.

[6] The inverse covariance matrix is sometimes called the *precision matrix*.

(Remember the covariance matrix is symmetric so $a_{12} = a_{21}$.) To avoid clutter, assume that $\boldsymbol{\mu} = (0,0)^T$, then the quadratic form is:

$$\begin{aligned} \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} &= \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{12}x_1 + a_{22}x_2 \end{pmatrix} \\ &= a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2\,. \end{aligned}$$

Thus we see that the argument to the exponential expands as a quadratic of D variables ($D = 2$ in this case).[7]

In the case of a diagonal covariance matrix:

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{\sigma_{11}} & 0 \\ 0 & \frac{1}{\sigma_{22}} \end{pmatrix} = \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix},$$

and the quadratic form has no cross terms:

$$\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} = a_{11}x_1^2 + a_{22}x_2^2\,.$$

In the multidimensional case the normalisation term in front of the exponential is $\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}$. Recall that the determinant of a matrix can be regarded as a measure of its size. And the dependence on the dimension reflects the fact that the volume increases with dimension.

Consider a two-dimensional Gaussian with the following mean vector and covariance matrix:

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

We refer to this as a *spherical* Gaussian since the probability distribution has spherical (circular) symmetry. The covariance matrix is diagonal (so the off-diagonal correlations are 0), and the variances are equal (1). This pdf is illustrated in the plots of this pdf in Figure 8.7a.

Now consider a two-dimensional Gaussian with the following mean vector and covariance matrix [8]:

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$$

In this case the covariance matrix is again diagonal, but the variances are not equal. Thus the resulting pdf has an elliptical shape, illustrated in Figure 8.7b.

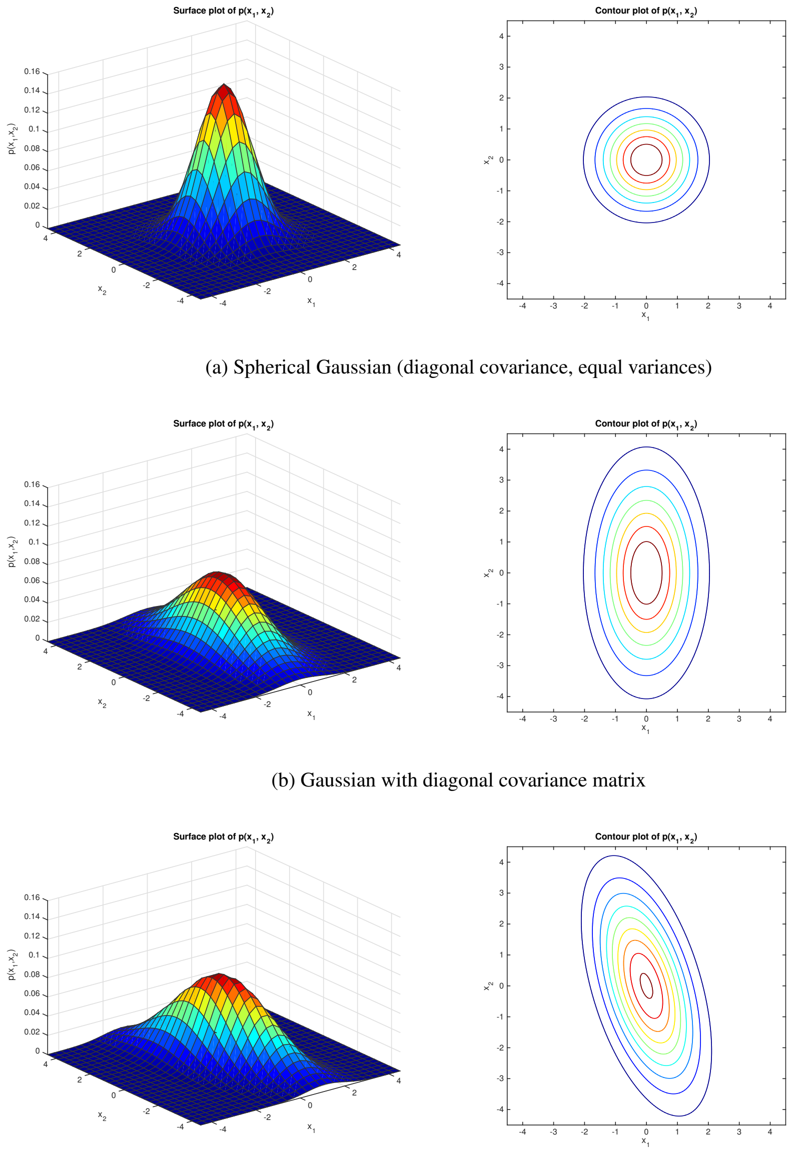
Now consider the following two-dimensional Gaussian:

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix}$$

In this case we have a full covariance matrix (off-diagonal terms are non-zero). The resultant pdf is shown in Figure 8.7c.

[7] Any covariance matrix is *positive semi-definite*, meaning $\mathbf{x}^T\boldsymbol{\Sigma}\,\mathbf{x} \ge 0$ for any real-valued vector $\mathbf{x}$. The inverse of covariance matrix, if it exists, is also positive semi-definite, i.e., $\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} \ge 0$.

[8] Like the covariance shown here, a covariance matrix whose off-diagonal elements are all zeros is called '*diagonal covariance matrix*', as opposed to '*full covariance matrix*' that has non-zero off-diagonal elements.



(a) Spherical Gaussian (diagonal covariance, equal variances)

(b) Gaussian with diagonal covariance matrix

(c) Gaussian with full covariance matrix

Figure 8.7: Surface and contour plots of 2–dimensional Gaussian with different covariance structures

8.7 Parameter estimation

It is possible to show that the mean vector and covariance matrix that maximise the likelihood of the Gaussian generating the training data are given by: [9]

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n \tag{8.18}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{x}_n - \hat{\boldsymbol{\mu}})\,(\mathbf{x}_n - \hat{\boldsymbol{\mu}})^T \,. \tag{8.19}$$

Alternatively, in a scalar representation:

$$\hat{\mu}_i = \frac{1}{N}\sum_{n=1}^{N} x_{ni}, \qquad \text{for } i = 1, \ldots, D \tag{8.20}$$

$$\hat{\sigma}_{ij} = \frac{1}{N}\sum_{n=1}^{N} (x_{ni} - \hat{\mu}_i)\,(x_{nj} - \hat{\mu}_j) \qquad \text{for } i, j = 1, \ldots, D \,. \tag{8.21}$$

As an example consider the data points displayed in Figure 8.8a. To fit a Gaussian to these samples we compute the mean and variance with MLE. The resulting Gaussian is superimposed as a contour map on the training data in Figure 8.8b.

8.8 Bayes' theorem and Gaussians

Many of the rules for combining probabilities that were outlined at the start of the course, are similar for probability density functions. For example, if x and y are continuous random variables, with probability density functions (pdfs) $p(x)$, and $p(y)$:

$$p(x, y) = p(x\,|\,y)\, p(y) \tag{8.22}$$

$$p(x) = \int p(x, y)\, \mathrm{d}y\,, \tag{8.23}$$

where $p(x\,|\,y)$ is the pdf of x given y.

Indeed we may mix probabilities of discrete variables and probability densities of continuous variables, for example if x is continuous and z is discrete:

$$p(x, z) = p(x\,|\,z)\, P(z)\,. \tag{8.24}$$

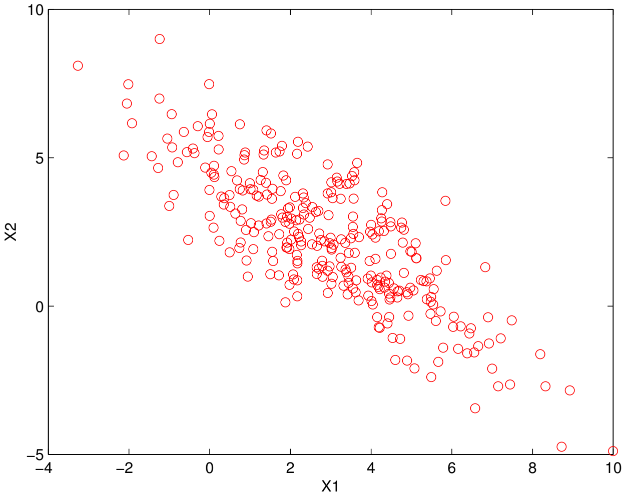
Proving that this is so requires a branch of mathematics called measure theory.

We can thus write Bayes' theorem for continuous data x and discrete class k as:

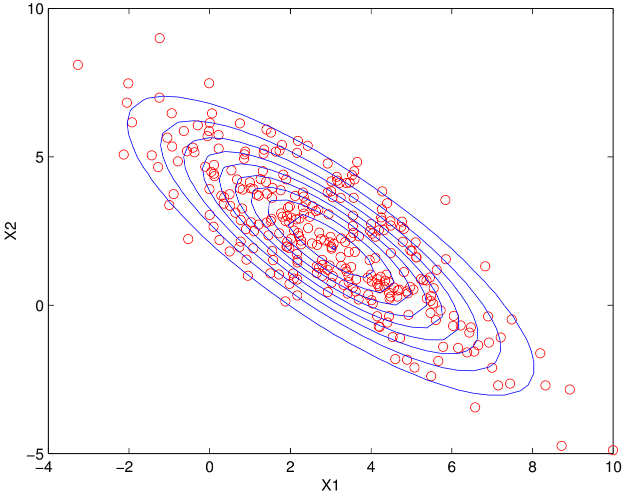
$$\begin{aligned} P(C_k\,|\,x) &= \frac{p(x\,|\,C_k)\, P(C_k)}{p(x)} \\ &= \frac{p(x\,|\,C_k)\, P(C_k)}{\sum_{\ell=1}^{K} p(x\,|\,C_\ell)\, P(C_\ell)} \end{aligned} \tag{8.25}$$

$$P(C_k\,|\,x) \propto p(x\,|\,C_k)\, P(C_k) \tag{8.26}$$

[9] Again the estimated covariance matrix with MLE is a biased estimator, rather than the unbiased estimator that is normalised by $N-1$.



(a) Training data



(b) Estimated Gaussian

Figure 8.8: Fitting a Gaussian to a set of two-dimensional data samples

The posterior probability of the class given the data is proportional to the probability density of the data times the prior probability of the class.

We can thus use Bayes' theorem for pattern recognition with continuous random variables.

If the pdf of continuous random variable x given class k is represented as a Gaussian with mean μ_k and variance σ_k^2, then we can write: [10]

$$\begin{aligned} P(C_k \mid x) &\propto p(x \mid C_k)\, P(C_k) \\ &\propto N(x\,;\,\mu_k, \sigma_k^2)\, P(C_k) \\ &\propto \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(\frac{-(x-\mu_k)^2}{2\,\sigma_k^2}\right) P(C_k)\,. \end{aligned} \tag{8.27}$$

We call $p(x \mid C_k)$ the likelihood of class k (given the observation x).

8.9 Appendix: Plotting Gaussians with Matlab

`plotgauss1D` is a function to plot a one-dimensional Gaussian with mean `mu` and variance `sigma2`:

```
function plotgauss1D(mu, sigma2)
% plot 1 dimension Gaussian with mean mu and variance sigma2
sd = sqrt(sigma2);  % std deviation
x = mu-3*sd:0.02:mu+3*sd; % location of points at which x is calculated
g = 1/(sqrt(2*pi)*sd)*exp(-0.5*(x-mu).^2/sigma2);
plot(x,g);
```

Recall that the standard deviation (SD) is the square root of the variance. It is a fact that about 0.68 of the probability mass of a Gaussian is within 1 SD (either side) of the mean, about 0.95 is within 2 SDs of the mean, and over 0.99 is within 3 SDs of the mean. Thus plotting a Gaussian for x ranging from $\mu - 3\sigma$ to $\mu + 3\sigma$ captures over 99% of the probability mass, and we take these as the ranges for the plot.

The following Matlab function plots two-dimensional Gaussians as a surface or a contour plot (and was used for the plots in the previous section). We could easily write it to take a (2-dimensional) mean vector and 2x2 covariance matrix, but it can be convenient to write the covariance matrix in terms of variances σ_{jj} and correlation coefficient, ρ_{jk}. Recall that:

$$\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}}\,. \tag{8.28}$$

Then we may write:

$$\sigma_{jk} = \rho_{jk}\,\sqrt{\sigma_{jj}}\,\sqrt{\sigma_{kk}} \tag{8.29}$$

where $\sqrt{\sigma_{jj}}$ is the standard deviation of dimension j. The following code does the job:

```
function plotgauss2D(xmu, ymu, xvar, yvar, rho)

% make a contour plot  and a surface plot of a 2D Gaussian
% xmu, ymu - mean of x, y
```

[10] The suffix k in μ_k and σ_k denotes the class number rather than the k' th element of vector.

```
% xvar, yvar - variance of x, y
% rho correlation coefficient between x and y

xsd = sqrt(xvar);  % std deviation on x axis
ysd = sqrt(yvar);  % std deviation on y axis

if (abs(rho) >= 1.0)
  disp(’error: rho must lie between -1 and 1’);
  return
end
covxy = rho*xsd*ysd;  % calculation of the covariance

C = [xvar covxy; covxy yvar];  % the covariance matrix
A = inv(C);    % the inverse covariance matrix

% plot between +-2SDs along each dimension
maxsd = max(xsd,ysd);
x = xmu-2*maxsd:0.1:xmu+2*maxsd; % location of points at which x is calculated
y = ymu-2*maxsd:0.1:ymu+2*maxsd; % location of points at which y is calculated

[X, Y] = meshgrid(x,y); % matrices used for plotting

% Compute value of Gaussian pdf at each point in the grid
z = 1/(2*pi*sqrt(det(C))) *
exp(-0.5 * (A(1,1)*(X-xmu).ˆ2 + 2*A(1,2)*(X-xmu).*(Y-ymu) + A(2,2)*(Y-ymu).ˆ2));

surf(x,y,z);
figure;
contour(x,y,z);
```

The above code computes the vectors x and y over which the function will be plotted. meshgrid takes these vectors and forms the set of all pairs ([X, Y]) over which the pdf is to be estimated. surf plots the function as a surface, and contour plots it as a contour map, or plan. You can use the Matlab help to find out more about plotting surfaces.

In the equation for the Gaussian pdf in plotgauss2D, because we are evaluating over points in a grid, we write out the quadratic form fully. More generally, if we want to evaluate a D-dimensional Gaussian pdf for a data point x, we can use a Matlab function like the following:

```
function y=evalgauss1(mu, covar, x)
% EVALGAUSS1 - evauate a Gaussian pdf

% y=EVALGAUSS1(MU, COVAR, X) evaluates a multivariate Gaussian with
% mean MU and covariance COVAR for a data point X


[d b] = size(covar);

% Check that the covariance matrix is square
if (d ˜= b)
```

```
  error(’Covariance matrix should be square’);
end

% force MU and X into column vectors
mu = reshape(mu, d, 1);
x = reshape(x, d, 1);

% subtract the mean from the data point
x = x-mu;

invcovar = inv(covar);

y = 1/sqrt((2*pi)ˆd*det(covar)) * exp (-0.5*x’*invcovar*x);
```

However, for efficiency it is usually better to estimate the Gaussian pdfs for a set of data points together. The following function, from the Netlab toolbox, takes an $n \times d$ matrix x, where each row corresponds to a data point.

```
function y = gauss(mu, covar, x)
% Y = GAUSS(MU, COVAR, X) evaluates a multi-variate Gaussian  density
% in D-dimensions at a set of points given by the rows of the matrix X.
% The Gaussian density has mean vector MU and covariance matrix COVAR.
%
% Copyright (c) Ian T Nabney (1996-2001)

[n, d] = size(x);
[j, k] = size(covar);

% Check that the covariance matrix is the correct dimension
if ((j ˜= d) | (k ˜=d))
  error(’Dimension of the covariance matrix and data should match’);
end

invcov = inv(covar);
mu = reshape(mu, 1, d);    % Ensure that mu is a row vector

x = x - ones(n, 1)*mu;    % Replicate mu and subtract from each data point
fact = sum(((x*invcov).*x), 2);

y = exp(-0.5*fact);

y = y./sqrt((2*pi)ˆd*det(covar));
```

Check that you understand how this function works. Note that sum(a,2) sums along rows of matrix a to return a column vector of the row sums. (sum(a,1) sums down columns to return a row vector.)

Exercises

1. Draw a one-dimensional Gaussian distribution by hand as accurate as possible when $\mu = 3.0$, $\sigma^2 = 1.0$. (You may use a calculator)

2. Using a calculator, find the height (i.e. maximum value) of a one-dimensional Gaussian distribution for $\sigma^2 = 10, 1.0, 0.1, 0.01, 0.001$. What the height will be, when $\sigma^2 \to 0$?

3. By solving the system of equations (8.12) and (8.13), confirm that the MLE for a a Gaussian distribution is given as (8.4) and (8.5).

4. Confirm that the correlation coefficient defined in Equation (8.16) is the same as the Pearson's correlation coefficient in Note 2.

5. Prove that the correlation coefficient is scale-invariant and location-invariant as is shown in Equation (8.17).

6. Consider a 2-dimensional Gaussian distribution with a mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ and a diagonal covariance matrix, i.e., $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{pmatrix}$, show that its pdf can be simplified to the product of two pdfs, each of which corresponds to a one-dimensional Gaussian distribution.

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(x_1|\mu_1, \sigma_{11})\, p(x_2|\mu_2, \sigma_{22})$$

7. For each of the Gaussian distributions shown in Figure 8.7, which type of correlation do x_1 and x_2 have, (i) a positive correlation, (ii) a negative correlation, or (iii) no correlation?