4 Basic probability theory

As outlined in Chapter 1, a major milestone for the modern approach to machine learning is to acknowledge our limited knowledge about the world and the unreliability of sensors and hence data in general. It is then only natural to consider quantities in our approaches as **random variables**. While a regular variable, once set, has only one specific value, a random variable will have different values every time we 'look' at it (draw an example from the distributions). For example, lets think about some data that we acquired by a light sensor. We might think that an ideal light sensor will give us only one reading while holding it to a specific surface. However, the characteristics of the internal electronic might change due to changing temperatures or fatigue in the sensor itself, or since we move the sensor unintentionally away from the surface, it is more than likely that we get different readings over time. Or think about image recognition; maybe we have a lighter in shape of a gun so that the functionality of the object is uncertain from its shape. Acknowledging uncertainty instead of denying it or simple trying to avoid it is an important paradigm shift in machine learning.

A common misconception about randomness is that one can not predict anything for a random variables. While we might not be able to predict one specific value, it is often the case that some values might be more likely than others. Indeed, we might be able to say something about how often a certain number will appear when drawing many examples. We might even be able to state how confident we are with this number, or, in other words, how variable these predictions are. The complete knowledge of a random variable, that is, how likely each value is for a random variable x, is captured by the **probability density function** $pdf(x)$. We discuss some specific examples of pdfs below. In these examples we assume that we know the pdf, but in many practical applications we must estimate this function. Indeed, estimation of pdfs is at in some sense the essence of machine learning. If we would know the 'world pdf', the probability function of all possible events in the world, then we could predict as much as possible in this world.

Many of the methods discussed later are **stochastic models** to capture the uncertainties in the world. Stochastic models are models with random variables, and it is therefore useful to remind ourselves about the properties of such variables. Even in the case our model does not have explicit random variables, the language of probability theory is often be used as the data in machine learning can be considered as an uncertain quantity. This chapter is a refresher on concepts in probability theory. Note that we are mainly interested in the language of probability theory rather than statistics. Statistics is more specifically about specific methods for hypothesis testing and related procedures. While machine learning methods can be viewed as some advanced statistical methods, our concern here is to learn about the language and tools of probability theory as a tool for developing out methods.

4.1 Random numbers and their probability (density) function

Probability theory is the theory of **random numbers**. We denoted such numbers by capital letters to distinguish them from regular numbers written in lower case. A random variable, X, is a quantity that can have different values each time the variable is inspected, such as in measurements in experiments. This is fundamentally different to a regular variable, x, which does not change its value once it is assigned. A random number is thus a new mathematical concept, not included in the regular mathematics of numbers. A specific value of a random number is still meaningful as it might influence specific processes in a deterministic way. However, since a random number can change every time it is inspected, it is also useful to describe more general properties when drawing examples many times. The frequency with which numbers can occur is then the most useful quantity to take into account. This frequency is captured by the mathematical construct of a **probability**. Note that there is often a debate if random numbers should be defined solely on the basis of a frequency measurement, or if they should be treated as a special kind of object with this inherent property. This philosophical debate between 'Frequentists' and 'Bayesians' is of minor importance for our applications. We do not ask where the uncertainty is coming from, we simply use the probability construct as a tool to describe uncertainty, and it is of minor importance to us if this is a inherent limitation or simply a lack of knowledge.

We can formalize the idea of expressing probabilities of drawing specific values for random variable with some compact notations. We speak of a **discrete random variable** in the case of discrete numbers for the possible values of a random number. A **continuous random variable** is a random variable that has possible values in a continuous set of numbers. There is, in principle, not much difference between these two kinds of random variables, except that the mathematical formulation has to be slightly different to be mathematically correct. For example, the **probability function**, also called **probability mass function** for discrete random numbers,

$$P(x) = P(X = x) \tag{4.1}$$

describes the frequency with which each possible value x of a discrete variable X occurs. Note that x is a regular variable, not a random variable. The value of $P(x)$ gives the fraction of the times we get a value x for the random variable X if we draw many examples of the random variable. Probabilities are sometimes written as a percentage, but we will stick to the fractional notation. From this definition it follows that the frequency of having any of the possible values is equal to one, which is an important normalization condition to be a probability function,

$$\sum_x P(x) = 1. \tag{4.2}$$

In the case of continuous random numbers we have an infinite number of possible values x so that the fraction for each number becomes formally infinitesimally small. It is hence necessary to write the probability distribution function as $P(x) = p(x)\mathrm{d}x$, where $p(x)$ is the **probability density function** (pdf). Note that we used here carefully upper case and lower case letters. The sum in eqn 4.2 then becomes an integral, and normalization condition for a continuous random variable is

$$\int_x p(x)\mathrm{d}x = 1. \tag{4.3}$$

A finite probability value makes then only sense for a certain range of numbers such as

$$P(a < x < b) = \int_{x=a}^{b} p(x)\mathrm{d}x. \tag{4.4}$$

We will formulate the rest of this section in terms of continuous random variables. The corresponding formulas for discrete random variables can easily be deduced by replacing the integrals over the pdf with sums over the probability function. It is also possible to use the δ-function, to write discrete random processes in a continuous form. The δ-function is a very convenient notation, which is formally a functional since it is only defined as an operation over a function. One can think of it as a density function that is zero except for its arguments for which it is infinite, and the integral over the δ-function is one; that is,

$$\int_{-\infty}^{\infty} \delta(x = x_1)\mathrm{d}x = 1. \tag{4.5}$$

It is usually used as an integration kernel with other functions as in

$$\int_{-\infty}^{\infty} \delta(x_1) f(x)\mathrm{d}x = f(x_1). \tag{4.6}$$

The delta function is useful for writing discrete events in a continuous form. For example, we could write the discrete density function for throwing a dice

$$P(x) = \frac{1}{6} \quad \text{for } x = \{1, 2, ..., 6\}, \tag{4.7}$$

as a density function

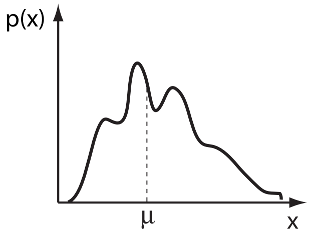
$$p(x) = \frac{1}{6}\delta(x = x_i) \quad \text{with } x_i = \{1, 2, ..., 6\}. \tag{4.8}$$

Note that we are here only playing with notations to introduce a concise language for our purposes.

4.2 Moments: mean, variance, etc.

In the following we only consider independent random values that are drawn from identical pdfs, often labeled as iid (independent and identically distributed) data. That is, we do not consider cases where with different probabilities when having given a specific value the random variable in a previous trial. The static probability density function describes all we can know about the corresponding random variable.

Let us consider the arbitrary pdf, $p(x)$, with the following graph:



Such a distribution is called **multimodal** because it has several peaks. Since this is a pdf, the area under this curve must be equal to one, as stated in eqn 4.3, to be a legitimate probability density function. It would be useful to have this function parameterized in an analytical format, and we will list some common parameterized density function below. Since we often don't know the probability density function of the quantities in interest in a machine learning setting, we will have to estimate pdfs. This approximation is the learning process in machine learning, and we will later outline specific methods to do this.

Finding a precise form of a pdf is difficult, and traditionally it is common to describe random variables with a small set of numbers that are meant to capture some properties of the probability density function. For example, we might ask what the most frequent value is when drawing many examples. This number is given by the largest peak value of the distribution.

$$p^{\max} = \mathbf{argmax}_x p(x). \tag{4.9}$$

Even more common is ask about the average value of the random sample when drawing many examples. A common quantity to know is thus the expected arithmetic average of those numbers, which is called the **mean**, **expected value**, or **expectation value** of the distribution, defined by

$$\mu = \int_{-\infty}^{\infty} x p(x) \mathrm{d}x. \tag{4.10}$$

This formula formalizes the calculation of adding all the different numbers together with their frequency.

A careful reader might have noticed a little oddity in our discussion. On the one hand we are saying that we want to characterize random variables through some simple measurements because we do not know the pdf, yet the last formula uses the pdf $p(x)$ that we usually don't know. To solve this apparent oddity we need to be more careful and talk about the **true underlying functions** and the **estimation** of such functions. If we would know the pdf that governs the random variable X, then equation 4.10 is the definition of the mean. However, in most applications we do not know the pdf, but we can define an approximation of the mean from measurements. For example, if we measure the frequency p_i of values in certain intervals around values x_i, then we can estimate the true mean μ by

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i p_i. \tag{4.11}$$

It is a common practice to denote an estimate of a quantity by adding a hat symbol to the quantity name. Also, note that we have use here a discretization procedure to approximate random variable that can be continuous in the most general case. Also note that we could enter here again the philosophical debate. Indeed, we have treated the pdf as fundamental and described the arithmetic average like an estimation of the mean. This might be viewed as *Bayesian*. However, we could also be pragmatic and say that we only have a collection of measurements so that the numbers are the 'real' thing, and that pdfs are only a mathematical construct. We will continue with a Bayesian description but note that this makes no difference at the end when using the formalism in specific applications.

The mean of a distribution is not the only interesting quantity that characterizes a distribution. For example, we might want to ask what the **median** value is. The median value is the value for the random variable for which it is equally likely to find a value lower or larger than this value,

$$\int_{-\infty}^{\mathrm{median}(x)} p(x)\mathrm{d}x = \int_{\mathrm{median}(x)}^{-\infty} p(x)\mathrm{d}x. \tag{4.12}$$

The median is equal to the mean in case of a symmetric distribution. Furthermore, the spread of the pdf around the mean is also very revealing as it gives us a sense of how spread the values are. This spread is often characterized by the standard deviation (std), or its square, which is called **variance**, σ^2, and is defined as

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\mathrm{d}x. \tag{4.13}$$

This is also called the second **moment** of the distribution whereas the mean would be the first moment. These two moments are generally not enough to characterize the probability function uniquely; this is only possible if we know all moments of a distribution, where the nth moment about the mean is defined as

$$m^n = \int_{-\infty}^{\infty} (x-\mu)^n f(x)\mathrm{d}x. \tag{4.14}$$

The **variance** is the second moment about the mean,

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\mathrm{d}x. \tag{4.15}$$

Higher moments specify further characteristics of distributions such as terms with third-order exponents (related to the quantity called skewness) or fourth-oder (such as a quantity called kurtosis). Knowing all moments of a distribution is equivalent to knowing the distribution precisely, and knowing a pdf is equivalent to knowing everything we could know about a random variable.

In case the distribution function is not given, moments have to be estimated from data. For example, the mean can be estimated from a sample of measurements by the **sample mean**,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{4.16}$$

and the variance from the **sample variance**,

$$s_1^2 = \frac{1}{n}\sum_{i=1}^{n}(\bar{x} - x_i)^2. \tag{4.17}$$

We will discuss later that these are the appropriate maximum likelihood estimates of these parameters. Note that the sample mean is an **unbiased estimate** while the sample variance is **biased**. A statistic is said to be biased if the mean of the sampling distribution is not equal to the parameter that is intended to be estimated. It can be shown that $E(s_1^2) = \frac{1}{n}\sigma^2$, and we can therefore adjust for the bias with a different normalization. It is hence common to use the **unbiased sample variance**

$$s_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{x} - x_i)^2, \tag{4.18}$$

as estimator of the variance. Of course, the difference is small for large sample sizes.

As mentioned above, knowing all moments uniquely specifies a pdf. This also implies that an incomplete list of moments does not uniquely define a pdf. Just extracting a list of estimated moments is thus of limited use for generalization without an explicit hypothesis of the underlying density function. For example, it is common to report the mean and variance of samples. This is specifically useful in case of an assumed Gaussian distributions as all higher moments are zero for this specific distribution. Thus, in practice it is mostly assumed, often without explicit mention, that the data are assumed to be Gaussian distributed. This is often done without any other consideration. In the age of computers with good plotting programs it is easy to a least make some checks. For example, plotting a histogram and seeing if this resembles somewhat a bell-shaped function is easy to do. Also, since we can now plot easily many data such as point clouds, the old way of summarizing distributions with moments should be seen as a more limited option. Machine learning can indeed be seen as a new approach to statistics that is currently making its way into many scientific areas as data analytics method.

4.3 Examples of probability (density) functions

There is an infinite number of possible pdfs. However, some specific forms have been very useful for describing some specific processes and have thus been given names. The following is a small list of examples with some discrete and several continuous distributions. The list is intended to give an overview of distributions that are often mentioned in scientific work, and some of them will be mentioned again in a later chapter. Such simple distributions are often a good starting assumptions. Most examples are discussed as one-dimensional distributions except the last example, which is a higher dimensional distribution. Again, we need to keep in mind that machine learning is mostly concerned with high dimensional cases so that these distributions are merely a starting point for illustrating some ideas.

4.3.1 Bernoulli distribution

A Bernoulli random variable is a variable from an experiment that has two possible outcomes: success with probability p; or failure, with probability $(1-p)$.

Probability function:
$P(\text{success}) = p \quad \Rightarrow \quad P(\text{failure}) = 1-p$
mean: p
variance: $p(1-p)$

4.3.2 Multinomial distribution

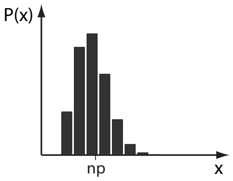
This is the distribution of outcomes in n trials that have k possible outcomes. The probability of each outcome is thereby p_i.

Probability function:
$P(x_i) = n! \prod_{i=1}^{k} (p_i^{x_i}/x_i!)$
mean: np_i
variance: $np_i(1-p_i)$

An important example is the Binomial distribution ($k=2$), which describes the the number of successes in n Bernoulli trials with probability of success p. Note that the binomial coefficient is defined as

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{4.19}$$

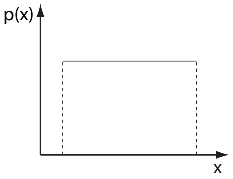
and is given by the Python function `itertools.permutations`.



Probability function:
$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$
mean: np
variance: $np(1-p)$

4.3.3 Uniform distribution

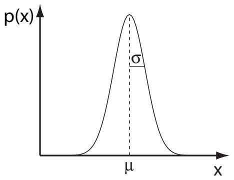
Equally distributed random numbers in the interval $a \leq x \leq b$. Pseudo-random variables with this distribution are often generated by routines in many programming languages.



Probability density function:
$p(x) = \frac{1}{b-a}$
mean: $(a+b)/2$
variance: $(b-a)^2/12$

4.3.4 Normal (Gaussian) distribution

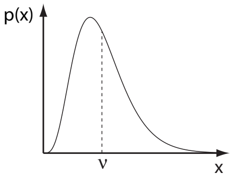
Limit of the binomial distribution for a large number of trials. Depends on two parameters, the mean μ and the standard deviation σ. The importance of the normal distribution stems from the central limit theorem outlined below.



Probability density function:
$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
mean: μ
variance: σ^2

4.3.5 Chi-square distribution

The sum of the squares of normally distributed random numbers is chi-square distributed and depends on a parameter ν that is equal to the mean. Γ is the gamma function included in Python as `scipy.stats.gamma`.



Probability density function:
$p(x) = \frac{x^{(\nu-2)/2}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}$
mean: ν
variance: 2ν

4.3.6 Multivariate Normal distribution

We will later consider density functions of a several random variables, $x_1, ..., x_n$. Such density functions are functions in higher dimensions. An important example is the multivariate Normal distribution (`scipy.stats.multivariate_normal` in Python) given by

$$p(x_1, ..., x_n) = p(\mathbf{x}) = \frac{1}{(\sqrt(2\pi))^n\sqrt(\det(\mathbf{\Sigma})}\exp(-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)). \tag{4.20}$$

This is a straight forward generalization of the one-dimensional Gaussian distribution mentioned before where the mean is now a vector, μ and the variance generalizes to a covariance matrix, $\mathbf{\Sigma} = [\mathrm{Cov}[X_i, X_j]]_{i=1,2,...,k;j=1,2,...,k}$ which must be symmetric and positive semi-definit. An example with mean $\mu = (1\ 2)^T$ and covariance $\Sigma = (1\ 0.5; 0.5\ 1)$ is shown in Fig,4.1.

4.4 Some advanced concepts

4.4.1 Cumulative probability (density) function and the Gaussian error function

We have mainly discussed probabilities of single values as specified by the probability (density) functions. However, in many cases we want to know the probabilities of

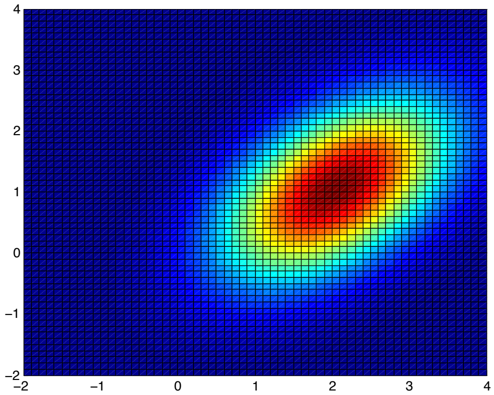


Fig. 4.1 Multivariate Gaussian with mean $\mu = (1\ 2)^T$ and covariance $\Sigma = (1\ 0.5; 0.5\ 1)$.

having values in a certain range. Indeed, the probability of a specific valuer of a continuous random variable is actually infinitesimally small (nearly zero), and only the probability of a range of values is finite and has a useful meaning of a probability. This integrated version of a probability density function is the probability of having a value x for the random variable X in the range of $x_1 \leq x \leq x_2$ and is given by

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x) \mathrm{d}x. \tag{4.21}$$

Note that we have shortened the notation by replacing the notation $P(x_1 \leq X \leq x_2)$ by $P(x_1 \leq X \leq x_2)$ to simplify the following expressions. In the main text we often need to calculate the probability that a normally (or Gaussian) distributed variable has values between $x_1 = 0$ and $x_2 = y$. The probability of eqn 4.21 then becomes a function of y. This defines the **Gaussian error function**

$$\frac{1}{\sqrt{2\pi}\sigma}\int_0^y \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x = \frac{1}{2}\mathrm{erf}(\frac{y-\mu}{\sqrt{2}\sigma}). \tag{4.22}$$

The name of this function comes from the fact that this integral often occurs when calculating confidence intervals with Gaussian noise and is often abbreviated as erf. This Gaussian error function for normally distributed variables (Gaussian distribution with mean $\mu = 0$ and variance $\sigma = 1$) is commonly tabulated in books on statistics. Programming libraries also frequently include routines that return the values for specific arguments. In Python this is implemented by the routine `scipy.special.erf`, and values for the inverse of the error function are returned by the routine `scipy.special.erfinv`.

Another special case of eqn 4.21 is when x_1 in the equation is equal to the lowest possible value of the random variable (usually $-\infty$). The integral in eqn 4.21 then corresponds to the probability that a random variable has a value smaller than a certain value, say y. This function of y is called the **cumulative density function** (cdf),[1]

$$P^{\mathrm{cum}}(x < y) = \int_{-\infty}^{y} p(x)\mathrm{d}x, \tag{4.23}$$

which we will utilize further below.

4.4.2 Functions of random variables and the central limit theorem

A function of a random variable X,

$$Y = f(X), \tag{4.24}$$

is also a random variable, Y, and we often need to know what the pdf of this new random variable is. Calculating with functions of random variables is a bit different to regular functions and some care has be given in such situations. Let us illustrate how to do this with an example. Say we have an equally distributed random variable X as commonly approximated with pseudo-random number generators on a computer. The probability density function of this variable is given by

$$p(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{4.25}$$

We are seeking the probability density function $p_Y(y)$ of the random variable

$$Y = \mathrm{e}^{-X^2}. \tag{4.26}$$

The random number Y is **not** Gaussian distributed as we might think naively. To calculate the probability density function we can employ the cumulative density function eqn 4.23 by noting that

[1] Note that this is a probability function, not a density function.

$$P(Y \leq y) = P(\mathrm{e}^{-X^2} \leq y) = P(X \geq \sqrt{-\ln y}). \tag{4.27}$$

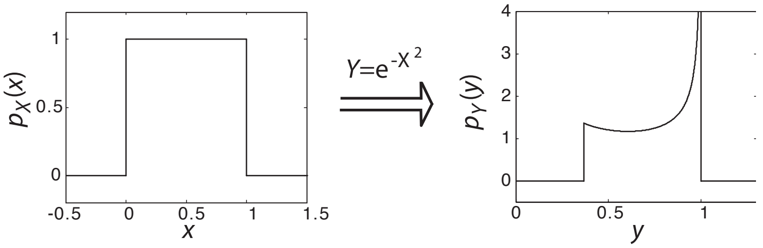
Thus, the cumulative probability function of Y can be calculated from the cumulative probability function of X,

$$P(X \geq \sqrt{-\ln y}) = \begin{cases} \int_{\sqrt{-\ln y}}^{1} p(x)\mathrm{d}y = 1 - \sqrt{-\ln y} & \text{for } \mathrm{e}^{-1} \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{4.28}$$

The probability density function of Y is the the derivative of this function,

$$p_Y(y) = \begin{cases} 1 - \sqrt{-\ln y} & \text{for } e^{-1} \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{4.29}$$

The probability density functions of X and Y are shown below.



A special function of random variables, which is of particular interest it can approximate many processes in nature, is the sum of many random variables. For example, such a sum occurs if we calculate averages from measured quantities, that is,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{4.30}$$

and we are interested in the probability density function of such random variables. This function depends, of course, on the specific density function of the random variables X_i. However, there is an important observation summarized in the **central limit theorem**. This theorem states that the average (normalized sum) of n random variables that are drawn from any distribution with mean μ and variance σ is approximately normally distributed with mean μ and variance σ/n for a sufficiently large sample size n. The approximation is, in practice, often very good also for small sample sizes. For example, the normalized sum of only seven uniformly distributed pseudo-random numbers is often used as a pseudo-random number for a normal distribution.

4.4.3 Measuring the difference between distributions

An important practical consideration is how to measure the similarity of difference between two density functions, say the density function p and the density function q. Note that such a measure is a matter of definition, similar to distance measures of real numbers or functions. However, there are some basic properties that we expect from a distance measure $d(a, b)$ between two item a and b. For example, a distance measure should be zero if the items to be compared are the same, that is $d(a,a) = 0$. Also, the value should be positive otherwise, $d(a,b) > 0$ for $a \neq b$, and a distance measure should be symmetrical, meaning that $d(a,b) = d(b,a)$.

To measure the difference between to distributions we could just plot them on top of each other and maybe measure the difference in area. Although not discussed here, it is common to define the information content in form of a logarithm, and the difference of logarithms is equal to the logarithm of the quotient. As common in probability theory, this measure should be weighted itself with the probability of the densities. A popular measure of similarity between two density functions is hence the so called **Kulbach–Leibler divergence** that is given by

$$d^{\mathrm{KL}}(p,q) = \int p(x)\log(\frac{p(x)}{q(x)})\mathrm{d}x \tag{4.31}$$

$$= \int p(x)\log(p(x))\mathrm{d}x - \int p(x)\log(q(x))\mathrm{d}x \tag{4.32}$$

This measure is zero if $p = q$ and always larger than zero if $p \neq q$. However, this measure is not symmetric, and this measure is therefore called a divergence instead of a distance. This measure is related to the information gain or relative entropy in information theory.

4.5 Density functions of multiple random variables

So far, we have discussed mainly probability (density) functions of single random variables. As mentioned before, we use random variables to describe data such as sensor readings in robots. Of course, we often have then more than one sensor and also other quantities that we describe by random variables at the same time. Thus, in many applications we consider multiple random variables. The quantities described by the random variables might be independent, but in many cases they are also related. Indeed, we will later talk about how to describe various types of relations. Thus, in order to talk about situations with multiple random variables, or multivariate statistics, it is useful to know basic rules. We start by illustrating these basic multivariate rules with two random variables since the generalization from there is usually quite obvious. But we will also talk about the generalization to more than two variables at the end of this section.

4.5.1 Basic definitions

We have seen that probability theory is quite handy to model data, and probability theory also considers multiple random variables. In the following we introduce the concepts and some definitions for two variables, although these concepts generalize directly to an arbitrary number of variables. Note that the number of random variable relates directly to the dimensionality of the problem in machine learning.

Let us start with some useful definitions. The total knowledge about the co-occurrence of specific values for two random variables X and Y is captured by the

$$\textbf{joined distribution:}\quad p(x,y) = p(X = x, Y = y). \tag{4.33}$$

This is a two dimensional functions. The two dimensions refers here to the number of variables, although a plot of this function would be a three dimensional plot. An example is shown in Fig.4.2. All the information we can have about a stochastic system is encapsulated in the joined pdf. The slice of this function, given the value of one variable, say y, is the

$$\textbf{conditional distribution}:\quad p(x|y) = p(X = x|Y = y). \tag{4.34}$$

A conditional pdf is also illustrated in Fig.4.2. If we sum over all realizations of y we get the

$$\textbf{marginal distribution}:\quad p(x) = \int p(x,y)dy. \tag{4.35}$$

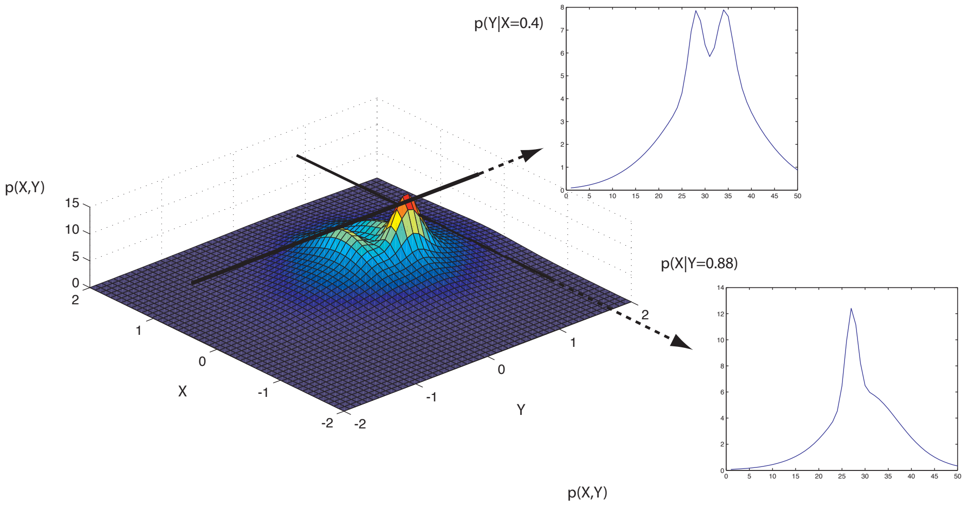


Fig. 4.2 Example of a two-dimensional probability density function (pdf) and some examples of conditional pdfs.

If we know some functional form of the density function or have a parameterized hypothesis of this function, than we can use common statistical methods, such as maximum likelihood estimation, to estimate the parameters as in the one dimensional cases. If we do not have a parameterized hypothesis we need to use other methods, such as treating the problem as discrete and building histograms to describe the density function of the system. Approximating a density function with histograms is a parameter-free method, although the bin size is a hyper-parameter of the method. The problem with this method is that it becomes very challenging or "data hungry" with increasing dimensions, Considering a simple histogram method to estimate the joined density function where we discretize the space along every dimension into n bins. This leads to n^2 bins for a two-dimensional histogram, and n^d for a d-dimensional problem. This exponential scaling is a major challenge in practice since we need also considerable data in each bin to sufficiently estimate the probability of each bin. This problem has been termed the **curse of dimensionality** by Richard Bellman.

4.5.2 The chain rule

As mentioned before, if we know the joined distribution of some random variables we can make the most predictions of these variables. However, in practice we have often to estimate these functions, and we can often only estimate conditional density functions. A very useful rule to know is therefore how a joined distribution can be decompose into the product of a conditional and a marginal distribution,

$$p(x,y) = p(x|y)p(y) = p(y|x)p(x), \tag{4.36}$$

which is an example of a **chain rule**. Note the two different ways in which we can decompose the joined distribution. This is easily generalized to n random variables by

$$\begin{aligned} p(x_1, x_2, ..., x_n) &= p(x_n|x_1, ...x_{n-1})p(x_1, ..., x_{n-1}) && (4.37)\\ &= p(x_n|x_1, ..., x_{n-1}) * ... * p(x_2|x_1) * p(x_1) && (4.38)\\ &= \Pi_{i=1}^{n} p(x_i|x_{i-1}, ...x_1) && (4.39) \end{aligned}$$

but note that there are also different decompositions possible. We will learn more about this and useful graphical representations in Chapter **??**.

Estimations of processes are greatly simplified when random variables are independent. A random variable X is independent of Y if

$$p(x|y) = p(x). \tag{4.40}$$

Using the chain rule eq.4.36, we can write this also as

$$p(x,y) = p(x)p(y), \tag{4.41}$$

that is, the joined distribution of two independent random variables is the product of their marginal distributions. Similar, we can also define conditional independence. For example, two random variables X and Y are conditionally independent of random variable Z if

$$p(x,y|z) = p(x|z)p(y|z). \tag{4.42}$$

Note that total independence does generally not imply conditionally independence and visa versa, although this might hold true for some examples.

4.5.3 How to combine prior knowledge with new evidence: Bayes rule

One of the most common tasks we will encounter in the following is the integration of prior knowledge with new evidence. For example, we could have an estimate of the location of an agent and get new (noisy) sensory data that adds some suggestions for different locations. A similar task is the fusion of data from different sensors. Or we have already a model that we build from previous data, and now we want to refine this with new data. The general question we have to solve is how to weight the different evidence in light of the reliability of this information. Solving this problem is easy in a probabilistic framework and is one of the main reasons that so much progress has been made in many application areas.

How prior knowledge should be combined with prior knowledge is an important question. Luckily, we basically already know how to do it best in a probabilistic sense. Namely, if we divide this chain rule eq. 4.36 by $p(x)$, which is possible as long as $p(x) > 0$, we get the identity

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}, \tag{4.43}$$

which is called **Bayes theorem** after the inventor Thomas Bayes. This theorem is important because it tells us how to combine a **prior** knowledge, such as the expected distribution over a random variable, $p(x)$, with some evidence called the likelihood $p(y|x)$. The likelihood can often be measured in some way, for example by measuring some sensors reading y when controlling the state x. The **posterior** distribution, $p(y|x)$ can then be calculated by multiplying the likelihood with the prior property for each x and normalizing this properly by the marginal distribution of y, $p(y)$. We will see that in practice knowing the marginal of y is difficult, but we will also see that the unnormalized version is useful in some applications such as classification as discussed later.

Bayes rule in conjunction with the chain rule and the rule of total probability are basically all you rules that are needed to do probabilistic inference. Probabilistic inference is to use these rules together with the known or estimated density functions to derive probabilistic statements. For example, let us calculate how likely it is to rain if a metrologist is predicting rain,

$$p(X = r|Y = r) = ?. \tag{4.44}$$

The random variable X stands for "actual condition", and r means rain, and the random variable Y stands for "predicted condition". Let us assume we know the following factors that we can easily measure. Let us assume that it rains in 30% of the days,

$$p(X = r) = 0.3, \tag{4.45}$$

Which we just calculated from past data by taking the ratio of days in which it rained. Since there are only two choices, it follows that the probability of no rain, which we write as $\not r$, is

$$p(X = \not r) = 1 - p(X = r) = 0.7. \tag{4.46}$$

Furthermore, from past predictions we know that the metrologist predicts correctly that it is raining 90%.

$$p(Y = r|X = r) = 0.9. \tag{4.47}$$

This number is again derived from previous data. From this we might conclude that it should be raining with a 90% probability, but we also need to take the prior knowledge into account which should bias our prediction downwards because it is less likely to rain than not to rain. To apply Bayes rule we also need to know how the metrologist does predicting no rain, and lets us assume that she is slightly better to predict when it is not raining as she gets this right 95% of the times. That is

$$p(Y = \not{r}|X = \not{r}) = 0.95. \tag{4.48}$$

The last equations also implies that she predicts rain in 5% of the cases when it does not rain,

$$p(Y = r|X = \not{r}) = 0.05. \tag{4.49}$$

We now have all the components we need to find a solution to the above question using Bayes theorem, namely

$$\begin{aligned} p(X = r|Y = r) &= \frac{p(Y = r|X = r)p(X = r)}{p(Y = r|X = r)p(X = r) + p(Y = r|X = \not{r})p(X = \not{r})} \\ &= \frac{0.90.3}{0.90.3 + 0.1 * 0.7} \\ &\approx 0.8 \end{aligned} \tag{4.50}$$

So we see that the actual probability that it is not raining if the metrologist predicts it is smaller than we might have thought.

Taking the prior into account is an essential part in Baysian reasoning. This will become clear in classification. For example, if we have a binary classification problem in which the positive is highly unlikely, say 99%, than always predicting the positive outcome would give us an accuracy of 0.99. Even though that seems good, a rue success of a prediction system should considerable outperform this naive prediction.

The above example is already an example of Bayesian modeling. We actually made a model of two Bernoulli random variables random

$$P(X = r) = p_{xr} \quad \text{and} \quad P(Y = r) = p_{yr} \tag{4.51}$$

and the conditional distributions

$$P(X = r|Y = r) = p_{xryr} \quad \text{and} \quad P(X = r|Y = \not{r}) = p_{xry\not{r}} \tag{4.52}$$

and estimated the parameters p_{xr}, p_{yr}, p_{xryr}, and $p_{xry\not{r}}$ from data. Estimating the parameters for data is the learning part, and we will derive below that the procedure of using the ratio of previous events is appropriate for this model. We then used this model with the rules of probability theory, specifically Bayes rule and the rule of total probability, the make predictions. This last step is sometimes called making a statistical inference. We will formalize below the process of how to learn the parameters of the model in the next chapter, and we will discuss the process of Bayesian modeling with associated tools in Chapter **??**.