need a variable to quantify the risk inherent in a random variable. One such measure is the *variance* of a random variable.

Definition 1.4 (Variance) *We define the variance of a random variable X as*

$$\mathrm{Var}[X] := \mathbb{E}\left[(X - \mathbf{E}[X])^2\right]. \tag{1.9}$$

As before, if $f : \mathbb{R} \to \mathbb{R}$ is a function, then the variance of $f(X)$ is given by

$$\mathrm{Var}[f(X)] := \mathbb{E}\left[(f(X) - \mathbf{E}[f(X)])^2\right]. \tag{1.10}$$

The variance measures by how much on average $f(X)$ deviates from its expected value. As we shall see in Section 2.1, an upper bound on the variance can be used to give guarantees on the probability that $f(X)$ will be within ϵ of its expected value. This is one of the reasons why the variance is often associated with the risk of a random variable. Note that often one discusses properties of a random variable in terms of its *standard deviation*, which is defined as the square root of the variance.

1.2.4 Marginalization, Independence, Conditioning, and Bayes Rule

Given two random variables X and Y, one can write their joint density $p(x, y)$. Given the joint density, one can recover $p(x)$ by integrating out y. This operation is called marginalization:

$$p(x) = \int_y dp(x, y). \tag{1.11}$$

If Y is a discrete random variable, then we can replace the integration with a summation:

$$p(x) = \sum_y p(x, y). \tag{1.12}$$

We say that X and Y are independent, *i.e.,* the values that X takes does not depend on the values that Y takes whenever

$$p(x, y) = p(x)p(y). \tag{1.13}$$

Independence is useful when it comes to dealing with large numbers of random variables whose behavior we want to estimate jointly. For instance, whenever we perform repeated measurements of a quantity, such as when

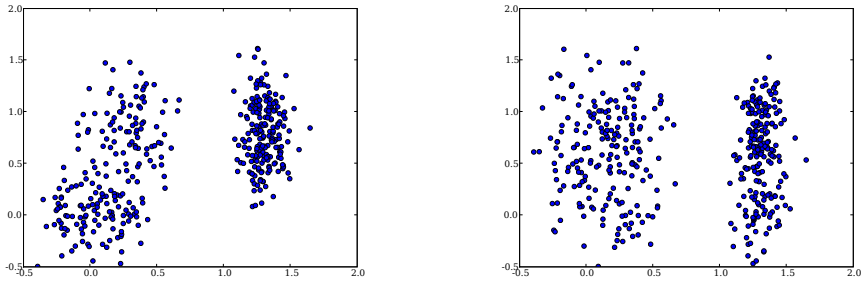

Fig. 1.12. Left: a sample from two dependent random variables. Knowing about first coordinate allows us to improve our guess about the second coordinate. Right: a sample drawn from two independent random variables, obtained by randomly permuting the dependent sample.

measuring the voltage of a device, we will typically assume that the individual measurements are drawn from the same distribution and that they are independent of each other. That is, having measured the voltage a number of times will not affect the value of the next measurement. We will call such random variables to be *independently and identically distributed*, or in short, *iid* random variables. See Figure 1.12 for an example of a pair of random variables drawn from dependent and independent distributions respectively.

Conversely, dependence can be vital in classification and regression problems. For instance, the traffic lights at an intersection are dependent of each other. This allows a driver to perform the inference that when the lights are green in his direction there will be no traffic crossing his path, i.e. the other lights will indeed be red. Likewise, whenever we are given a picture x of a digit, we hope that there will be dependence between x and its label y.

Especially in the case of dependent random variables, we are interested in conditional probabilities, *i.e.,* probability that X takes on a particular value given the value of Y. Clearly $Pr(X = rain|Y = cloudy)$ is higher than $Pr(X = rain|Y = sunny)$. In other words, knowledge about the value of Y significantly influences the distribution of X. This is captured via conditional probabilities:

$$p(x|y) := \frac{p(x,y)}{p(y)}. \tag{1.14}$$

Equation 1.14 leads to one of the key tools in statistical inference.

Theorem 1.5 (Bayes Rule) *Denote by X and Y random variables then*

the following holds

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}. \tag{1.15}$$

This follows from the fact that $p(x, y) = p(x|y)p(y) = p(y|x)p(x)$. The key consequence of (1.15) is that we may *reverse* the conditioning between a pair of random variables.

1.2.4.1 An Example

We illustrate our reasoning by means of a simple example — inference using an AIDS test. Assume that a patient would like to have such a test carried out on him. The physician recommends a test which is guaranteed to detect HIV-positive whenever a patient is infected. On the other hand, for healthy patients it has a 1% error rate. That is, with probability 0.01 it diagnoses a patient as HIV-positive even when he is, in fact, HIV-negative. Moreover, assume that 0.15% of the population is infected.

Now assume that the patient has the test carried out and the test returns 'HIV-negative'. In this case, logic implies that he is healthy, since the test has 100% detection rate. In the converse case things are not quite as straightforward. Denote by X and T the random variables associated with the health status of the patient and the outcome of the test respectively. We are interested in $p(X = \text{HIV+}|T = \text{HIV+})$. By Bayes rule we may write

$$p(X = \text{HIV+}|T = \text{HIV+}) = \frac{p(T = \text{HIV+}|X = \text{HIV+})p(X = \text{HIV+})}{p(T = \text{HIV+})}$$

While we know all terms in the numerator, $p(T = \text{HIV+})$ itself is unknown. That said, it can be computed via

$$\begin{aligned} p(T = \text{HIV+}) &= \sum_{x \in \{\text{HIV+},\text{HIV-}\}} p(T = \text{HIV+}, x) \\ &= \sum_{x \in \{\text{HIV+},\text{HIV-}\}} p(T = \text{HIV+}|x)p(x) \\ &= 1.0 \cdot 0.0015 + 0.01 \cdot 0.9985. \end{aligned}$$

Substituting back into the conditional expression yields

$$p(X = \text{HIV+}|T = \text{HIV+}) = \frac{1.0 \cdot 0.0015}{1.0 \cdot 0.0015 + 0.01 \cdot 0.9985} = 0.1306.$$

In other words, even though our test is quite reliable, there is such a low prior probability of having been infected with AIDS that there is not much evidence to accept the hypothesis even after this test.

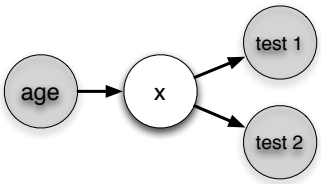

Fig. 1.13. A graphical description of our HIV testing scenario. Knowing the age of the patient influences our prior on whether the patient is HIV positive (the random variable X). The outcomes of the tests 1 and 2 are independent of each other given the status X. We observe the shaded random variables (age, test 1, test 2) and would like to infer the un-shaded random variable X. This is a special case of a graphical model which we will discuss in Chapter **??**.

Let us now think how we could improve the diagnosis. One way is to obtain further information about the patient and to use this in the diagnosis. For instance, information about his age is quite useful. Suppose the patient is 35 years old. In this case we would want to compute $p(X = \text{HIV+}|T = \text{HIV+}, A = 35)$ where the random variable A denotes the age. The corresponding expression yields:

$$\frac{p(T = \text{HIV+}|X = \text{HIV+}, A)p(X = \text{HIV+}|A)}{p(T = \text{HIV+}|A)}$$

Here we simply *conditioned* all random variables on A in order to take additional information into account. We may assume that the test is *independent* of the age of the patient, i.e.

$$p(t|x, a) = p(t|x).$$

What remains therefore is $p(X = \text{HIV+}|A)$. Recent US census data pegs this number at approximately 0.9%. Plugging all data back into the conditional expression yields $\frac{1 \cdot 0.009}{1 \cdot 0.009 + 0.01 \cdot 0.991} = 0.48$. What has happened here is that by including additional observed random variables our estimate has become more reliable. Combination of evidence is a powerful tool. In our case it helped us make the classification problem of whether the patient is HIV-positive or not more reliable.

A second tool in our arsenal is the use of multiple measurements. After the first test the physician is likely to carry out a second test to confirm the diagnosis. We denote by T_1 and T_2 (and t_1, t_2 respectively) the two tests. Obviously, what we want is that T_2 will give us an "independent" second opinion of the situation. In other words, we want to ensure that T_2 does not make the same mistakes as T_1. For instance, it is probably a bad idea to repeat T_1 without changes, since it might perform the same diagnostic

mistake as before. What we want is that the diagnosis of T_2 is independent of that of T_2 *given* the health status X of the patient. This is expressed as

$$p(t_1, t_2|x) = p(t_1|x)p(t_2|x). \tag{1.16}$$

See Figure 1.13 for a graphical illustration of the setting. Random variables satisfying the condition (1.16) are commonly referred to as *conditionally independent*. In shorthand we write $T_1, T_2 \perp\!\!\!\perp X$. For the sake of the argument we assume that the statistics for T_2 are given by

$p(t_2\|x)$	$x =$ HIV-	$x =$ HIV+
$t_2 =$ HIV-	0.95	0.01
$t_2 =$ HIV+	0.05	0.99

Clearly this test is less reliable than the first one. However, we may now combine both estimates to obtain a very reliable estimate based on the combination of both events. For instance, for $t_1 = t_2 =$ HIV+ we have

$$p(X = \text{HIV+}|T_1 = \text{HIV+}, T_2 = \text{HIV+}) = \frac{1.0 \cdot 0.99 \cdot 0.009}{1.0 \cdot 0.99 \cdot 0.009 + 0.01 \cdot 0.05 \cdot 0.991} = 0.95.$$

In other words, by combining two tests we can now confirm with very high confidence that the patient is indeed diseased. What we have carried out is a combination of evidence. Strong experimental evidence of two positive tests effectively overcame an initially very strong prior which suggested that the patient might be healthy.

Tests such as in the example we just discussed are fairly common. For instance, we might need to decide which manufacturing procedure is preferable, which choice of parameters will give better results in a regression estimator, or whether to administer a certain drug. Note that often our tests may not be conditionally independent and we would need to take this into account.

1.3 Basic Algorithms

We conclude our introduction to machine learning by discussing four simple algorithms, namely Naive Bayes, Nearest Neighbors, the Mean Classifier, and the Perceptron, which can be used to solve a binary classification problem such as that described in Figure 1.5. We will also introduce the K-means algorithm which can be employed when labeled data is not available. All these algorithms are readily usable and easily implemented from scratch in their most basic form.

For the sake of concreteness assume that we are interested in spam filtering. That is, we are given a set of m e-mails x_i, denoted by $\mathbf{X} := \{x_1, \ldots, x_m\}$

```
From: "LucindaParkison497072" <LucindaParkison497072@hotmail.com>
To: <kargr@earthlink.net>
Subject: we think ACGU is our next winner
Date: Mon, 25 Feb 2008 00:01:01 -0500
MIME-Version: 1.0
X-OriginalArrivalTime: 25 Feb 2008 05:01:01.0329 (UTC) FILETIME=[6A931810:01C8776B]
Return-Path: lucindaparkison497072@hotmail.com

(ACGU) .045 UP 104.5%

I do think that (ACGU) at it's current levels looks extremely attractive.

Asset Capital Group, Inc., (ACGU) announced that it is expanding the marketing of bio-remediation fluids and cleaning equipment. After
its recent acquisition of interest in American Bio-Clean Corporation and an 80

News is expected to be released next week on this growing company and could drive the price even higher. Buy (ACGU) Monday at open. I
believe those involved at this stage could enjoy a nice ride up.
```

Fig. 1.14. Example of a spam e-mail

x_1: `The quick brown fox jumped over the lazy dog.`
x_2: `The dog hunts a fox.`

	`the`	`quick`	`brown`	`fox`	`jumped`	`over`	`lazy`	`dog`	`hunts`	`a`
x_1	2	1	1	1	1	1	1	1	0	0
x_2	1	0	0	1	0	0	0	1	1	1

Fig. 1.15. Vector space representation of strings.

and associated labels y_i, denoted by $\mathbf{Y} := \{y_1, \ldots, y_m\}$. Here the labels satisfy $y_i \in \{\text{spam}, \text{ham}\}$. The key assumption we make here is that the pairs (x_i, y_i) are drawn jointly from some distribution $p(x, y)$ which represents the e-mail generating process for a user. Moreover, we assume that there is sufficiently strong dependence between x and y that we will be able to estimate y given x and a set of labeled instances $\mathbf{X}, \mathbf{Y}$.

Before we do so we need to address the fact that e-mails such as Figure 1.14 are *text*, whereas the three algorithms we present will require data to be represented in a *vectorial* fashion. One way of converting text into a vector is by using the so-called *bag of words* representation [Mar61, Lew98]. In its simplest version it works as follows: Assume we have a list of all possible words occurring in $\mathbf{X}$, that is a dictionary, then we are able to assign a unique number with each of those words (e.g. the position in the dictionary). Now we may simply count for each document x_i the number of times a given word j is occurring. This is then used as the value of the j-th coordinate of x_i. Figure 1.15 gives an example of such a representation. Once we have the latter it is easy to compute distances, similarities, and other statistics directly from the vectorial representation.