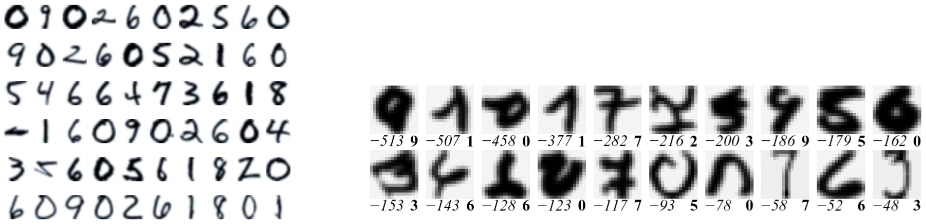

Fig. 1.8. Left: typical digits contained in the database of the US Postal Service. Right: unusual digits found by a novelty detection algorithm [SPST+01] (for a description of the algorithm see Section 7.4). The score below the digits indicates the degree of novelty. The numbers on the lower right indicate the class associated with the digit.

as to how novel it is. Readers familiar with density estimation might contend that the latter would be a reasonable solution. However, we neither need a score which sums up to 1 on the entire domain, nor do we care particularly much about novelty scores for *typical* observations. We will later see how this somewhat easier goal can be achieved directly. Figure 1.8 has an example of novelty detection when applied to an optical character recognition database.

1.2 Probability Theory

In order to deal with the instances of where machine learning can be used, we need to develop an adequate language which is able to describe the problems concisely. Below we begin with a fairly informal overview over probability theory. For more details and a very gentle and detailed discussion see the excellent book of [BT03].

1.2.1 Random Variables

Assume that we cast a dice and we would like to know our chances whether we would see 1 rather than another digit. If the dice is fair all six outcomes $\mathcal{X} = \{1, \ldots, 6\}$ are equally likely to occur, hence we would see a 1 in roughly 1 out of 6 cases. Probability theory allows us to model uncertainty in the outcome of such experiments. Formally we state that 1 occurs with probability $\frac{1}{6}$.

In many experiments, such as the roll of a dice, the outcomes are of a numerical nature and we can handle them easily. In other cases, the outcomes may not be numerical, *e.g.,* if we toss a coin and observe heads or tails. In these cases, it is useful to associate numerical values to the outcomes. This is done via a random variable. For instance, we can let a random variable

X take on a value $+1$ whenever the coin lands heads and a value of -1 otherwise. Our notational convention will be to use uppercase letters, *e.g.,* X, Y etc to denote random variables and lower case letters, *e.g.,* x, y etc to denote the values they take.

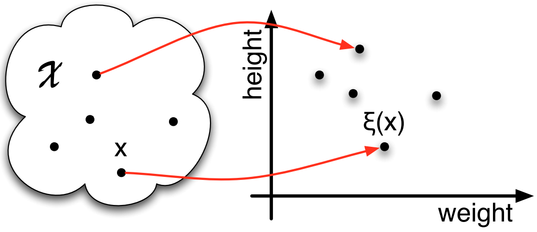

Fig. 1.9. The random variable ξ maps from the set of outcomes of an experiment (denoted here by $\mathcal{X}$) to real numbers. As an illustration here $\mathcal{X}$ consists of the patients a physician might encounter, and they are mapped via ξ to their weight and height.

1.2.2 Distributions

Perhaps the most important way to characterize a random variable is to associate probabilities with the values it can take. If the random variable is discrete, *i.e.,* it takes on a finite number of values, then this assignment of probabilities is called a *probability mass function* or PMF for short. A PMF must be, by definition, non-negative and must sum to one. For instance, if the coin is fair, *i.e.,* heads and tails are equally likely, then the random variable X described above takes on values of $+1$ and -1 with probability 0.5. This can be written as

$$Pr(X = +1) = 0.5 \text{ and } Pr(X = -1) = 0.5. \tag{1.1}$$

When there is no danger of confusion we will use the slightly informal notation $p(x) := Pr(X = x)$.

In case of a continuous random variable the assignment of probabilities results in a *probability density function* or PDF for short. With some abuse of terminology, but keeping in line with convention, we will often use density or distribution instead of probability density function. As in the case of the PMF, a PDF must also be non-negative and integrate to one. Figure 1.10 shows two distributions: the uniform distribution

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise,} \end{cases} \tag{1.2}$$

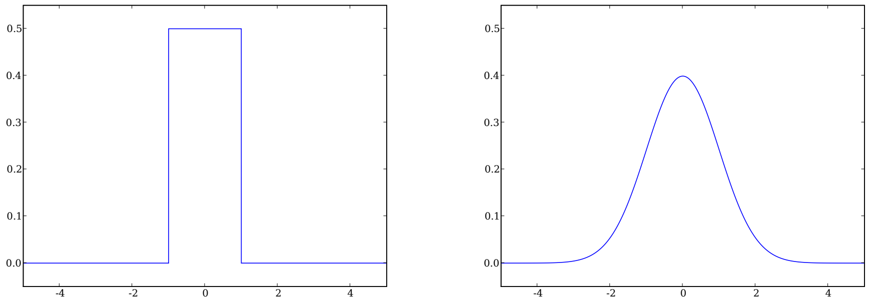

Fig. 1.10. Two common densities. Left: uniform distribution over the interval $[-1, 1]$. Right: Normal distribution with zero mean and unit variance.

and the Gaussian distribution (also called normal distribution)

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{1.3}$$

Closely associated with a PDF is the indefinite integral over p. It is commonly referred to as the cumulative distribution function (CDF).

Definition 1.1 (Cumulative Distribution Function) *For a real valued random variable X with PDF p the associated Cumulative Distribution Function F is given by*

$$F(x') := \Pr\left\{X \leq x'\right\} = \int_{-\infty}^{x'} dp(x). \tag{1.4}$$

The CDF $F(x')$ allows us to perform range queries on p efficiently. For instance, by integral calculus we obtain

$$\Pr(a \leq X \leq b) = \int_a^b dp(x) = F(b) - F(a). \tag{1.5}$$

The values of x' for which $F(x')$ assumes a specific value, such as 0.1 or 0.5 have a special name. They are called the *quantiles* of the distribution p.

Definition 1.2 (Quantiles) *Let $q \in (0, 1)$. Then the value of x' for which $\Pr(X < x') \leq q$ and $\Pr(X > x') \leq 1 - q$ is the q-quantile of the distribution p. Moreover, the value x' associated with $q = 0.5$ is called the median.*

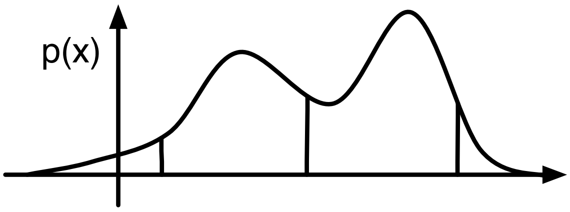

Fig. 1.11. Quantiles of a distribution correspond to the area under the integral of the density $p(x)$ for which the integral takes on a pre-specified value. Illustrated are the 0.1, 0.5 and 0.9 quantiles respectively.

1.2.3 Mean and Variance

A common question to ask about a random variable is what its expected value might be. For instance, when measuring the voltage of a device, we might ask what its typical values might be. When deciding whether to administer a growth hormone to a child a doctor might ask what a sensible range of height should be. For those purposes we need to define expectations and related quantities of distributions.

Definition 1.3 (Mean) *We define the mean of a random variable X as*

$$\mathbb{E}[X] := \int x dp(x) \tag{1.6}$$

More generally, if $f : \mathbb{R} \to \mathbb{R}$ is a function, then $f(X)$ is also a random variable. Its mean is mean given by

$$\mathbb{E}[f(X)] := \int f(x) dp(x). \tag{1.7}$$

Whenever X is a discrete random variable the integral in (1.6) can be replaced by a summation:

$$\mathbb{E}[X] = \sum_x x p(x). \tag{1.8}$$

For instance, in the case of a dice we have equal probabilities of 1/6 for all 6 possible outcomes. It is easy to see that this translates into a mean of $(1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$.

The mean of a random variable is useful in assessing expected losses and benefits. For instance, as a stock broker we might be interested in the expected value of our investment in a year's time. In addition to that, however, we also might want to investigate the *risk* of our investment. That is, how likely it is that the value of the investment might deviate from its expectation since this might be more relevant for our decisions. This means that we 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