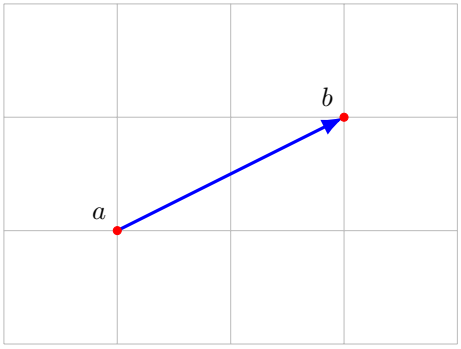

Figure 3.1 The norm of the displacement $b - a$ is the distance between the points with coordinates a and b.

value. The right-hand side is the inverse square of the ratio of a to $\mathbf{rms}(x)$. It says, for example, that no more than $1/25 = 4\%$ of the entries of a vector can exceed its RMS value by more than a factor of 5. The Chebyshev inequality partially justifies the idea that the RMS value of a vector gives an idea of the size of a typical entry: It states that not too many of the entries of a vector can be much bigger (in absolute value) than its RMS value. (A converse statement can also be made: At least one entry of a vector has absolute value as large as the RMS value of the vector; see exercise 3.8.)

3.2 Distance

Euclidean distance. We can use the norm to define the *Euclidean distance* between two vectors a and b as the norm of their difference:

$$\mathbf{dist}(a, b) = \|a - b\|.$$

For one, two, and three dimensions, this distance is exactly the usual distance between points with coordinates a and b, as illustrated in figure 3.1. But the Euclidean distance is defined for vectors of any dimension; we can refer to the distance between two vectors of dimension 100. Since we only use the Euclidean norm in this book, we will refer to the Euclidean distance between vectors as, simply, the distance between the vectors. If a and b are n-vectors, we refer to the RMS value of the difference, $\|a - b\|/\sqrt{n}$, as the *RMS deviation* between the two vectors.

When the distance between two n-vectors x and y is small, we say they are 'close' or 'nearby', and when the distance $\|x - y\|$ is large, we say they are 'far'. The particular numerical values of $\|x-y\|$ that correspond to 'close' or 'far' depend on the particular application.

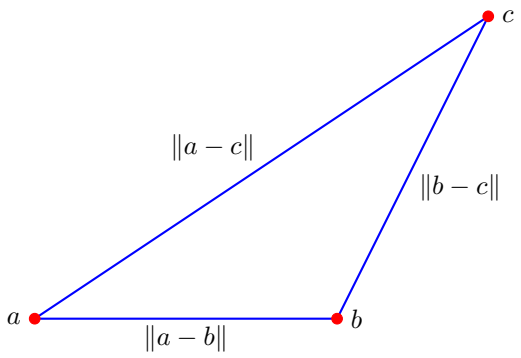

Figure 3.2 Triangle inequality.

As an example, consider the 4-vectors

$$u = \begin{bmatrix} 1.8 \\ 2.0 \\ -3.7 \\ 4.7 \end{bmatrix}, \qquad v = \begin{bmatrix} 0.6 \\ 2.1 \\ 1.9 \\ -1.4 \end{bmatrix}, \qquad w = \begin{bmatrix} 2.0 \\ 1.9 \\ -4.0 \\ 4.6 \end{bmatrix}.$$

The distances between pairs of them are

$$\|u - v\| = 8.368, \qquad \|u - w\| = 0.387, \qquad \|v - w\| = 8.533,$$

so we can say that u is much nearer (or closer) to w than it is to v. We can also say that w is much nearer to u than it is to v.

Triangle inequality. We can now explain where the triangle inequality gets its name. Consider a triangle in two or three dimensions, whose vertices have coordinates a, b, and c. The lengths of the sides are the distances between the vertices,

$$\mathbf{dist}(a, b) = \|a - b\|, \qquad \mathbf{dist}(b, c) = \|b - c\|, \qquad \mathbf{dist}(a, c) = \|a - c\|.$$

Geometric intuition tells us that the length of any side of a triangle cannot exceed the sum of the lengths of the other two sides. For example, we have

$$\|a - c\| \leq \|a - b\| + \|b - c\|. \tag{3.3}$$

This follows from the triangle inequality, since

$$\|a - c\| = \|(a - b) + (b - c)\| \leq \|a - b\| + \|b - c\|.$$

This is illustrated in figure 3.2.

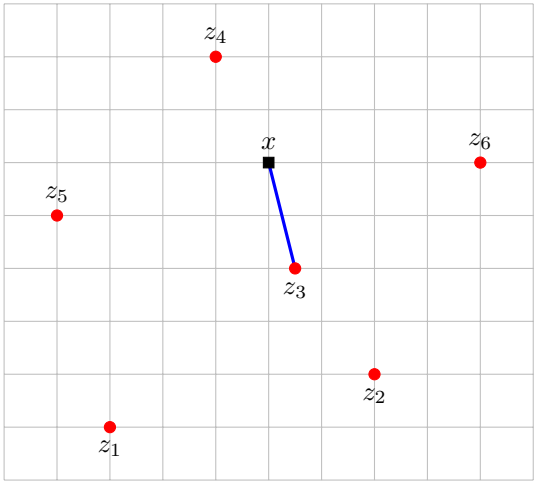

Figure 3.3 A point x, shown as a square, and six other points $z_1, \ldots, z_6$. The point z_3 is the nearest neighbor of x among the points $z_1, \ldots, z_6$.

Examples.

- *Feature distance.* If x and y represent vectors of n features of two objects, the quantity $\|x - y\|$ is called the *feature distance*, and gives a measure of how different the objects are (in terms of their feature values). Suppose for example the feature vectors are associated with patients in a hospital, with entries such as weight, age, presence of chest pain, difficulty breathing, and the results of tests. We can use feature vector distance to say that one patient case is near another one (at least in terms of their feature vectors).

- *RMS prediction error.* Suppose that the n-vector y represents a time series of some quantity, for example, hourly temperature at some location, and $\hat{y}$ is another n-vector that represents an estimate or prediction of the time series y, based on other information. The difference $y - \hat{y}$ is called the *prediction error*, and its RMS value $\mathbf{rms}(y - \hat{y})$ is called the *RMS prediction error*. If this value is small (say, compared to $\mathbf{rms}(y)$) the prediction is good.

- *Nearest neighbor.* Suppose $z_1, \ldots, z_m$ is a collection of m n-vectors, and that x is another n-vector. We say that z_j is the *nearest neighbor* of x (among $z_1, \ldots, z_m$) if
$$\|x - z_j\| \leq \|x - z_i\|, \quad i = 1, \ldots, m.$$
In words: z_j is the closest vector to x among the vectors $z_1, \ldots, z_m$. This is illustrated in figure 3.3. The idea of nearest neighbor, and generalizations such as the k-nearest neighbors, are used in many applications.

- *Document dissimilarity.* Suppose n-vectors x and y represent the histograms of word occurrences for two documents. Then $\|x - y\|$ represents a measure of the dissimilarity of the two documents. We might expect the dissimilarity

	Veterans Day	Memorial Day	Academy Awards	Golden Globe Awards	Super Bowl
Veterans Day	0	0.095	0.130	0.153	0.170
Memorial Day	0.095	0	0.122	0.147	0.164
Academy A.	0.130	0.122	0	0.108	0.164
Golden Globe A.	0.153	0.147	0.108	0	0.181
Super Bowl	0.170	0.164	0.164	0.181	0

Table 3.1 Pairwise word count histogram distances between five Wikipedia articles.

to be smaller when the two documents have the same genre, topic, or author; we would expect it to be larger when they are on different topics, or have different authors. As an example we form the word count histograms for the 5 Wikipedia articles with titles 'Veterans Day', 'Memorial Day', 'Academy Awards', 'Golden Globe Awards', and 'Super Bowl', using a dictionary of 4423 words. (More detail is given in §4.4.) The pairwise distances between the word count histograms are shown in table 3.1. We can see that pairs of related articles have smaller word count histogram distances than less related pairs of articles.

Units for heterogeneous vector entries. The square of the distance between two n-vectors x and y is given by

$$\|x - y\|^2 = (x_1 - y_1)^2 + \cdots + (x_n - y_n)^2,$$

the sum of the squares of the differences between their respective entries. Roughly speaking, the entries in the vectors all have equal status in determining the distance between them. For example, if x_2 and y_2 differ by one, the contribution to the square of the distance between them is the same as the contribution when x_3 and y_3 differ by one. This makes sense when the entries of the vectors x and y represent the same type of quantity, using the same units (say, at different times or locations), for example meters or dollars. For example if x and y are word count histograms, their entries are all word occurrence frequencies, and it makes sense to say they are close when their distance is small.

When the entries of a vector represent different types of quantities, for example when the vector entries represent different types of features associated with an object, we must be careful about choosing the units used to represent the numerical values of the entries. If we want the different entries to have approximately equal status in determining distance, their numerical values should be approximately of the same magnitude. For this reason units for different entries in vectors are often chosen in such a way that their typical numerical values are similar in magnitude, so that the different entries play similar roles in determining distance.

As an example suppose that the 2-vectors x, y, and z are the feature vectors for three houses that were sold, as in the example described on page 39. The first entry of each vector gives the house area and the second entry gives the number of

bedrooms. These are very different types of features, since the first one is a physical area, and the second one is a count, *i.e.*, an integer. In the example on page 39, we chose the unit used to represent the first feature, area, to be thousands of square feet. With this choice of unit used to represent house area, the numerical values of both of these features range from around 1 to 5; their values have roughly the same magnitude. When we determine the distance between feature vectors associated with two houses, the difference in the area (in thousands of square feet), and the difference in the number of bedrooms, play equal roles.

For example, consider three houses with feature vectors

$$x = (1.6, 2), \quad y = (1.5, 2), \quad z = (1.6, 4).$$

The first two are 'close' or 'similar' since $\|x - y\| = 0.1$ is small (compared to the norms of x and y, which are around 2.5). This matches our intuition that the first two houses are similar, since they both have two bedrooms and are close in area. The third house would be considered 'far' or 'different' from the first two houses, and rightly so since it has four bedrooms instead of two.

To appreciate the significance of our choice of units in this example, suppose we had chosen instead to represent house area directly in square feet, and not thousands of square feet. The three houses above would then be represented by feature vectors

$$\tilde{x} = (1600, 2), \quad \tilde{y} = (1500, 2), \quad \tilde{z} = (1600, 4).$$

The distance between the first and third houses is now 2, which is very small compared to the norms of the vectors (which are around 1600). The distance between the first and second houses is much larger. It seems strange to consider a two-bedroom house and a four-bedroom house as 'very close', while two houses with the same number of bedrooms and similar areas are much more dissimilar. The reason is simple: With our choice of square feet as the unit to measure house area, distances are very strongly influenced by differences in area, with number of bedrooms playing a much smaller (relative) role.

3.3 Standard deviation

For any vector x, the vector $\tilde{x} = x - \mathbf{avg}(x)\mathbf{1}$ is called the associated *de-meaned* vector, obtained by subtracting from each entry of x the mean value of the entries. (This is not standard notation; *i.e.*, $\tilde{x}$ is not generally used to denote the de-meaned vector.) The mean value of the entries of $\tilde{x}$ is zero, *i.e.*, $\mathbf{avg}(\tilde{x}) = 0$. This explains why $\tilde{x}$ is called the de-meaned version of x; it is x with its mean removed. The de-meaned vector is useful for understanding how the entries of a vector deviate from their mean value. It is zero if all the entries in the original vector x are the same.

The *standard deviation* of an n-vector x is defined as the RMS value of the de-meaned vector $x - \mathbf{avg}(x)\mathbf{1}$, *i.e.*,

$$\mathbf{std}(x) = \sqrt{\frac{(x_1 - \mathbf{avg}(x))^2 + \cdots + (x_n - \mathbf{avg}(x))^2}{n}}.$$