Statistical Pattern Recognition, Second Edition. Andrew R. Webb

ISBNs: 0-470-84513-9 (HB); 0-470-84514-7 (PB)

A

Measures of dissimilarity

A.1 Measures of dissimilarity

Patterns or objects analysed using the techniques described in this book are usually represented by a vector of measurements. Many of the techniques require some measure of dissimilarity or distance between two pattern vectors, although sometimes data can arise directly in the form of a dissimilarity matrix.

A particular class of dissimilarity functions called *dissimilarity coefficients* are required to satisfy the following conditions. If d_{rs} is the dissimilarity of object s from object r, then

$$\begin{aligned} d_{rs} &\geq 0 \quad && \text{for every } r, s \\ d_{rr} &= 0 \quad && \text{for every } r \\ d_{rs} &= d_{sr} \quad && \text{for every } r, s \end{aligned}$$

The symmetry condition is not always satisfied by some dissimilarity functions. If the dissimilarity between two places in a city centre is the distance travelled by road between them, then because of one-way systems the distance may be longer in one direction than the other. Measures of dissimilarity can be transformed to similarity measures using various transformations, for example, $s_{ij} = 1/(1 + d_{ij})$ or $s_{ij} = c - d_{ij}$ for some constant c, where s_{ij} is the similarity between object i and object j.

If, in addition to the three conditions above, the dissimilarity measure satisfies the triangle inequality

$$d_{rt} + d_{ts} \geq d_{rs} \quad \text{for every } r, s, t \tag{A.1}$$

then the dissimilarity measure is a *metric* and the term *distance* is usually used.

A.1.1 Numeric variables

Many dissimilarity measures have been proposed for numeric variables. Table A.1 gives some of the more common measures. The choice of a particular metric depends on the application. Computational considerations aside, for feature selection and extraction purposes you would choose the metric that gives the best performance (perhaps in terms of classification error on a validation set).

Table A.1 Dissimilarity measures for numeric variables (between $\boldsymbol{x}$ and $\boldsymbol{y}$)

Dissimilarity measure	Mathematical form
Euclidean distance	$d_e = \left\{ \sum_{i=1}^{p} (x_i - y_i)^2 \right\}^{\frac{1}{2}}$
City-block distance	$d_{cb} = \sum_{i=1}^{p} \lvert x_i - y_i \rvert$
Chebyshev distance	$d_{ch} = \max_i \lvert x_i - y_i \rvert$
Minkowski distance of order m	$d_M = \left\{ \sum_{i=1}^{p} (x_i - y_i)^m \right\}^{\frac{1}{m}}$
Quadratic distance	$d_q = \sum_{i=1}^{p} \sum_{j=1}^{p} (x_i - y_i) Q_{ij} (x_j - y_j)$, Q positive definite
Canberra distance	$d_{ca} = \sum_{i=1}^{p} \frac{\lvert x_i - y_i \rvert}{x_i + y_i}$
Nonlinear distance	$d_n = \begin{cases} H & d_e > D \\ 0 & d_e \leq D \end{cases}$
Angular separation	$\frac{\sum_{i=1}^{p} x_i y_i}{\left[\sum_{i=1}^{p} x_i^2 \sum_{i=1}^{p} y_i^2 \right]^{1/2}}$

Euclidean distance

$$d_e = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

The contours of equal Euclidean distance from a point are hyperspheres (circles in two dimensions). It has the (perhaps undesirable) property of giving greater emphasis to larger differences on a single variable.

Although we may wish to use a dissimilarity measure that is a metric, some of the methods do not require the metric condition (A.1) above. Therefore in some cases a monotonic function of the Euclidean metric, which will still be a dissimilarity coefficient but not necessarily a metric, will suffice. For example, squared Euclidean distance is a dissimilarity coefficient but not a metric.

City-block distance

$$d_{cb} = \sum_{i=1}^{p} \lvert x_i - y_i \rvert$$

Also known as the *Manhattan* or *box-car* or *absolute value* distance, this metric uses a distance calculation which would be suitable for finding the distances between points in a city consisting of a grid of intersecting thoroughfares (hence the names used). The contours of equal distance from a point for the city-block metric are diamonds in two dimensions. The city-block metric is a little cheaper to compute than the Euclidean distance so it may be used if the speed of a particular application is important.

Chebyshev distance

$$d_{ch} = \max_i |x_i - y_i|$$

The Chebyshev or *maximum value* distance is often used in cases where the execution speed is so critical that the time involved in calculating the Euclidean distance is unacceptable. The Chebyshev distance, like the city-block distance, examines the absolute magnitude of the elementwise differences in the pair of vectors. The contour lines of equal Chebyshev distance from a point are squares in two dimensions. Figure A.1 plots the contours of equal distance in $\mathbb{R}^2$ for the Euclidean, city-block and Chebyshev metrics.

If the user needs an approximation to Euclidean distance but with a cheaper computational load then the first line of approach is to use either the Chebyshev or the city-block metrics. A better approximation can be gained by using a combination of these two distances:

$$d = \max(\tfrac{2}{3}d_{cb}, d_{ch})$$

In two dimensions the contours of equal distance form octagons.

Minkowski distance The Minkowski distance is a more general form of the Euclidean and city-block distances. The Minkowski distance of order m is

$$d_M = \left(\sum_{i=1}^{p} |x_i - y_i|^m \right)^{1/m}$$

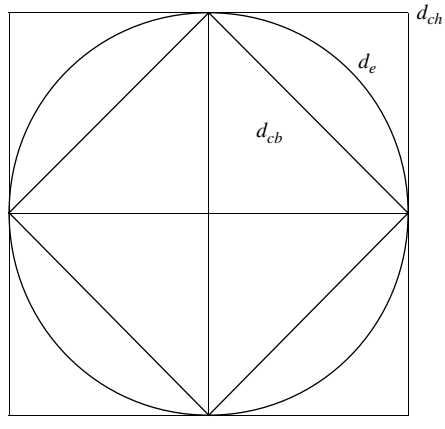

Figure A.1 Contours of equal distance

The Minkowski distance of the first order is the same as the city-block metric and the Minkowski distance of the second order is the Euclidean distance. The contours of equal distance for such metrics form squared-off circles which gradually obtain more abrupt vertices as m increases. The choice of an appropriate value for m depends on the amount of emphasis you would like to give to the larger differences: larger values of m give progressively more emphasis to the larger differences $|x_i - y_i|$, and as m tends to infinity the metric tends to the Chebyshev distance (and square contours).

Quadratic distance

$$d_q^2 = (\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{y})$$

A choice for $\boldsymbol{Q}$ is the inverse of the within-group covariance matrix. This is sometimes referred to as the Mahalanobis distance, because of the similarity to the distance measure between two distributions (see below).

Canberra metric

$$d_{ca} = \sum_{i=1}^{p} \frac{|x_i - y_i|}{x_i + y_i}$$

The Canberra metric is a sum of a series of fractions and is suitable for variables taking non-negative values. If both x_i and y_i are zero the ratio of the difference to the sum is taken to be zero. If only one value is zero, the term is unity, independent of the other value. Thus, 0 and 1 are equally dissimilar to a pair of elements 0 and 10^6. Sometimes values of 0 are replaced by small positive numbers (smaller than the recorded values of that variable).

Nonlinear distance

$$d_N = \begin{cases} 0 & \text{if } d_e(\boldsymbol{x}, \boldsymbol{y}) < D \\ H & \text{if } d_e(\boldsymbol{x}, \boldsymbol{y}) \geq D \end{cases}$$

where D is a threshold and H is a constant. Kittler (1975a) shows that an appropriate choice for H and D for feature selection is that they should satisfy

$$H = \frac{\Gamma(p/2)}{D^p 2\sqrt{\pi^p}}$$

and that D satisfies the unbiasedness and consistency conditions of the Parzen estimator, namely $D^p n \to \infty$ and $D \to 0$ as $n \to \infty$, where n is the number of samples in the data set.

Angular separation

$$\frac{\sum_{i=1}^{p} x_i y_i}{\left[\sum_{i=1}^{p} x_i^2 \sum_{i=1}^{p} y_i^2\right]}$$

The angular separation is a similarity rather than a dissimilarity measure that measures the angle between the unit vectors in the direction of the two pattern vectors of interest. This is appropriate when data are collected for which only the relative magnitudes are important.

The choice of a particular proximity measure depends on the application and may depend on several factors, including distribution of data and computational considerations. It is not possible to make recommendations, and studies in this area have been largely empirical, but the method you choose should be the one that you believe will capture the essential differences between objects.

A.1.2 Nominal and ordinal variables

Nominal and ordinal variables are usually represented as a set of binary variables. For example, a nominal variable with s states is represented as s binary variables. If it is in the mth state, then each of the s binary variables has value 0 except the mth, which has the value unity. The dissimilarity between two objects can be obtained by summing the contributions from the individual variables.

For ordinal variables, the contribution to the dissimilarity between two objects from a single variable does not simply depend on whether or not the values are identical. If the contribution for one variable in state m and one in state l ($m < l$) is δ_{ml}, then we require

$$\begin{aligned} \delta_{ml} &\geq \delta_{ms} \quad \text{for } s < l \\ \delta_{ml} &\geq \delta_{sl} \quad \text{for } s > m \end{aligned}$$

that is, δ_{ml} is monotonic down each row and across each column of the half-matrix of distances between states ($\delta_{14} > \delta_{13} > \delta_{12}$ etc.; $\delta_{14} > \delta_{24} > \delta_{34}$). The values chosen for δ_{ml} depend very much on the problem. For example, we may have a variable describing fruits of a plant that can take the values short, long or very long. We would want the dissimilarity between a plant with very long fruit and one with short fruit to be greater than that between one with long fruit and one with short fruit (all other attributes having equal values). A numeric coding of 1, 2, 3 would achieve this, but so would 1, 10, 100.

A.1.3 Binary variables

Various dissimilarity measures have been proposed for binary variables. For vectors of binary variables $\boldsymbol{x}$ and $\boldsymbol{y}$ these may be expressed in terms of quantities a, b, c, and d where

a is equal to the number of occurrences of $x_i = 1$ and $y_i = 1$
b is equal to the number of occurrences of $x_i = 0$ and $y_i = 1$
c is equal to the number of occurrences of $x_i = 1$ and $y_i = 0$
d is equal to the number of occurrences of $x_i = 0$ and $y_i = 0$

This is summarised in Table A.2. Note that $a+b+c+d = p$, the total number of variables (attributes). It is customary to define a similarity measure rather than a dissimilarity measure. Table A.3 summarises some of the more commonly-used similarity measures for binary data.

Table A.2 Co-occurrence table for binary variables

		x_i	
		1	0
y_i	1	a	b
	0	c	d

Table A.3 Similarity measures for binary data

Similarity measure	Mathematical form
Simple matching coefficient	$d_{sm} = \frac{a+d}{a+b+c+d}$
Russell and Rao	$d_{rr} = \frac{a}{a+b+c+d}$
Jaccard	$d_j = \frac{a}{a+b+c}$
Czekanowski	$d_{Cz} = \frac{2a}{2a+b+c}$

Simple matching coefficient The simple matching coefficient is the proportion of variables for which two variables have the same value. The dissatisfaction with this measure has been with the term d representing *conjoint absences*. The fact that two sites in an ecological survey both lack something should not make them more similar. The dissimilarity measure defined by $d_{xy} = 1 - s_{xy} = (b+c)/p$ is proportional to the square of the Euclidean distance, $b+c$, which is the Hamming distance in communication theory.

Russell and Rao This does not involve the term d in the numerator and is appropriate in certain circumstances. The quantity $1 - s_{xy}$ is not a dissimilarity coefficient since the dissimilarity between an object and itself is not necessarily zero.

Jaccard This does not involve the quantity d at all and is used extensively by ecologists. The term $d_{xy} = 1 - s_{xy}$ is a metric dissimilarity coefficient.

Czekanowski This is similar to the Jaccard measure except that coincidences carry double weight.

Many other coefficients have been proposed that handle the conjoint absences in various ways (Clifford and Stephenson, 1975; Diday and Simon, 1976).

A.1.4 Summary

We have listed some of the measures of proximity which can be found in the pattern processing and classification literature. A general similarity coefficient between two objects $\boldsymbol{x}$ and $\boldsymbol{y}$ encompassing variables of mixed type has been proposed by Gower

(1971). Of course, there is no such thing as a best measure. Some will be more appropriate for certain tasks than others. Therefore, we cannot make recommendations. However, the user should consider the following points when making a choice: (1) simplicity and ease of understanding; (2) ease of implementation; (3) speed requirements; (4) knowledge of data.

A.2 Distances between distributions

All of the distance measures described so far have been defined between two patterns or objects. We now turn to measures of distances between groups of objects or distributions. These measures are used to determine the discriminatory power of a feature set, discussed in Chapter 9. Many measures have been proposed in the pattern recognition literature, and we introduce two basic types here. The first uses prototype vectors for each class together with the distance metrics of the previous section. The second uses knowledge of the class-conditional probability density functions. Many methods of this type are of academic interest only. Their practical application is rather limited since they involve numerical integration and estimation of the probability density functions from samples. They do simplify if the density functions belong to a family of parametric functions such as the exponential family, which includes the normal distribution. The use of both of these approaches for feature selection is described in Chapter 9.

A.2.1 Methods based on prototype vectors

There are many measures of inter-group dissimilarities based on prototype vectors. In the context of clustering, these give rise to different hierarchical schemes, which are discussed in Chapter 10. Here we introduce the average separation, defined to be the average distance between all pairs of points, with one point in each pair coming from each distribution. That is, for n_1 points in ω_1 ($\boldsymbol{x}_i, i = 1, \ldots, n_1$) and n_2 points in ω_2 ($\boldsymbol{y}_i, i = 1, \ldots, n_2$),

$$J_{as}(\omega_1, \omega_2) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} d(\boldsymbol{x}_i, \boldsymbol{y}_j)$$

where d is a distance between $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$.

A.2.2 Methods based on probabilistic distance

These measures use the complete information about the structure of the classes provided by the conditional densities. The distance measure, J, satisfies the following conditions:

1. $J = 0$ if the probability density functions are identical, $p(x|\omega_1) = p(x|\omega_2)$;

2. $J \geq 0$;

3. J attains its maximum when the classes are disjoint, i.e. when $p(x|\omega_1) = 0$ and $p(x|\omega_2) \neq 0$.

Many measures satisfying these conditions have been proposed (Chen, 1976; Devijver and Kittler, 1982). As an introduction, consider two overlapping distributions with conditional densities $p(\boldsymbol{x}|\omega_1)$ and $p(\boldsymbol{x}|\omega_2)$. The classification error, e (see Chapter 8), is given by

$$e = \frac{1}{2}\left\{1 - \int |p(\omega_1|\boldsymbol{x}) - p(\omega_2|\boldsymbol{x})| p(\boldsymbol{x})\, d\boldsymbol{x}\right\}$$

The integral in the equation above,

$$J_K = \int |p(\omega_1|\boldsymbol{x}) - p(\omega_2|\boldsymbol{x})| p(\boldsymbol{x})\, d\boldsymbol{x}$$

is called the *Kolmogorov variational distance* and has the important property that it is directly related to the classification error. Other measures cannot be expressed in terms of the classification error, but can be used to provide bounds on the error. Three of these are given in Table A.4. A more complete list can be found in the books by Chen (1976) and Devijver and Kittler (1982).

One of the main disadvantages of the probabilistic dependence criteria is that they require an estimate of a probability density function and its numerical integration. This restricts their usefulness in many practical situations. However, under certain assumptions regarding the form of the distributions, the expressions can be evaluated analytically.

First of all, we shall consider a specific parametric form for the distributions, namely normally distributed with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. Under these assumptions, the distance measures can be written down as follows.

Table A.4 Probabilistic distance measures

Dissimilarity measure	Mathematical form
Average separation	$\frac{1}{n_a n_b}\sum_{i=1}^{n_a}\sum_{j=1}^{n_b} d(\boldsymbol{x}_i, \boldsymbol{y}_j), \boldsymbol{x}_i \in \omega_A; \boldsymbol{y}_j \in \omega_B;$ d any distance metric
Chernoff	$J_c = -\log \int p^s(\boldsymbol{x}|\omega_1) p^{1-s}(\boldsymbol{x}|\omega_2)\, d\boldsymbol{x}$
Bhattacharyya	$J_B = -\log \int (p(\boldsymbol{x}|\omega_1) p(\boldsymbol{x}|\omega_2))^{\frac{1}{2}}\, d\boldsymbol{x}$
Divergence	$J_D = \int [p(\boldsymbol{x}|\omega_1) - p(\boldsymbol{x}|\omega_2)] \log\left(\frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)}\right) d\boldsymbol{x}$
Patrick–Fischer	$J_P = \left\{\int [p(\boldsymbol{x}|\omega_1)p_1 - p(\boldsymbol{x}|\omega_2)p_2]^2\, d\boldsymbol{x}\right\}^{\frac{1}{2}}$

Chernoff

$$J_c = \frac{1}{2}s(1-s)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T[\boldsymbol{\Sigma}_s]^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \frac{1}{2}\log\left(\frac{|\boldsymbol{\Sigma}_s|}{|\boldsymbol{\Sigma}_1|^{1-s}|\boldsymbol{\Sigma}_2|^s}\right)$$

where $\boldsymbol{\Sigma}_s = (1-s)\boldsymbol{\Sigma}_1 + s\boldsymbol{\Sigma}_2$ and $s \in [0, 1]$. For $s = 0.5$, we have the Bhattacharyya distance.

Bhattacharyya

$$J_B = \frac{1}{4}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T[\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2]^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \frac{1}{2}\log\left(\frac{|\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2|}{2(|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|)^{\frac{1}{2}}}\right)$$

Divergence

$$J_D = \tfrac{1}{2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T(\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \mathrm{Tr}\{\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2 - 2\boldsymbol{I}\}$$

Patrick–Fischer

$$\begin{aligned} J_P =& (2\pi)^{-p/2}\Bigg[|2\boldsymbol{\Sigma}_1|^{-\frac{1}{2}} + |2\boldsymbol{\Sigma}_2|^{-\frac{1}{2}} \\ & -2|\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\right\}\Bigg] \end{aligned}$$

Finally, if the covariance matrices are equal, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, the Bhattacharyya and divergence distances simplify to

$$J_M = J_D = 8J_B = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$

which is the Mahalanobis distance.

Of course, the means and covariance matrices are not known in practice and must be estimated from the available training data.

The above parametric forms are useful both in feature selection and extraction. In feature selection, the set of features at the kth stage of an algorithm is constructed from the set of features at the $(k-1)$th stage by the addition or subtraction of a small number of features. The value of the feature selection criterion at stage $k+1$ may be computed from that at stage k rather than evaluating the above expressions directly. This saves on computation. Recursive calculation of separability measures is discussed in Chapter 9.

Probabilistic distance measures can also be extended to the multigroup case by evaluating all pairwise distances between classes,

$$J = \sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j J_{ij}$$

where J_{ij} is the chosen distance measure evaluated for class ω_i and class ω_j.

A.2.3 Probabilistic dependence

The probabilistic distance measures are based on discrimination between a pair of classes, using the class-conditional densities to describe each class. Probabilistic dependence measures are multiclass feature selection criteria that measure the distance between the class-conditional density and the mixture probability density function (see Figure A.2). If $p(\boldsymbol{x}|\omega_i)$ and $p(\boldsymbol{x})$ are identical then we gain no information about class by observing $\boldsymbol{x}$, and the 'distance' between the two distributions is zero. Thus, $\boldsymbol{x}$ and ω_i are independent. If the distance between $p(\boldsymbol{x}|\omega_i)$ and $p(\boldsymbol{x})$ is large, then the observation $\boldsymbol{x}$ is dependent on ω_i. The greater the distance, the greater the dependence of $\boldsymbol{x}$ on the class ω_i. Table A.5

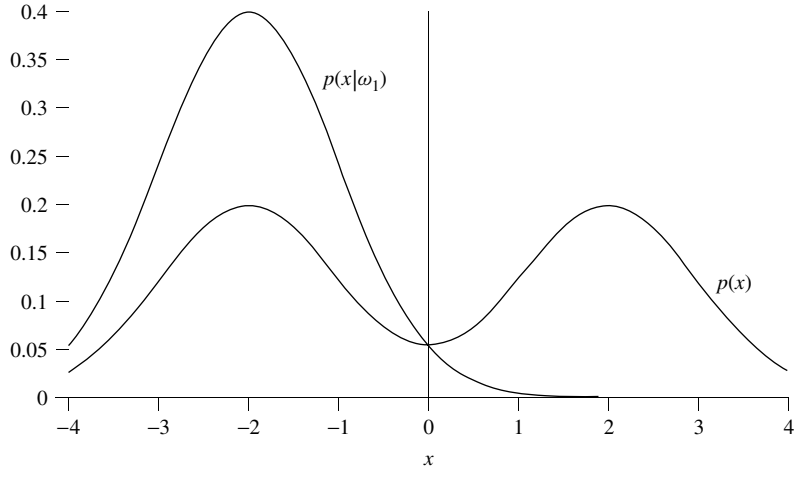

Figure A.2 Probabilistic dependence

Table A.5 Probabilistic dependence measures

Dissimilarity measure	Mathematical form
Chernoff	$J_c = \sum_{i=1}^{C} p_i \left\{ -\log \int p^s(\boldsymbol{x}\|\omega_i) p^{1-s}(\boldsymbol{x})\, d\boldsymbol{x} \right\}$
Bhattacharyya	$J_B = \sum_{i=1}^{C} p_i \left\{ -\log \int (p(\boldsymbol{x}\|\omega_i) p(\boldsymbol{x}))^{\frac{1}{2}}\, d\boldsymbol{x} \right\}$
Joshi	$J_D = \sum_{i=1}^{C} p_i \int [p(\boldsymbol{x}\|\omega_i) - p(\boldsymbol{x})] \log \left(\frac{p(\boldsymbol{x}\|\omega_i)}{p(\boldsymbol{x})} \right) d\boldsymbol{x}$
Patrick–Fischer	$J_P = \sum_{i=1}^{C} p_i \left\{ \int [p(\boldsymbol{x}\|\omega_i) - p(\boldsymbol{x})]^2\, d\boldsymbol{x} \right\}^{\frac{1}{2}}$

gives the probabilistic dependence measures corresponding to the probabilistic distance measures in Table A.4. In practice, application of probabilistic dependence measures is limited because, even for normally distributed classes, the expressions given in Table A.5 cannot be evaluated analytically since the mixture distribution $p(\boldsymbol{x})$ is not normal.

A.3 Discussion

This appendix has reviewed some of the distance and dissimilarity measures used in Chapter 9 on feature selection and extraction and Chapter 10 on clustering. Of course the list is not exhaustive and those that are presented may not be the best for your problem. We cannot make rigid recommendations as to which ones you should use since the choice is highly problem-specific. However, it may be advantageous from a computational point of view to use one that simplifies for normal distributions even if your data are not normally distributed.

The book by Gordon (1999) provides a good introduction to classification methods. The chapter on dissimilarity measures also highlights difficulties encountered in practice with real data sets. There are many other books and papers on clustering which list other measures of dissimilarity: for example, Diday and Simon (1976), Cormack (1971) and Clifford and Stephenson (1975), which is written primarily for biologists but the issues treated occur in many areas of scientific endeavour. The papers by Kittler (1975b, 1986) provide very good introductions to feature selection and list some of the more commonly used distance measures. Others may be found in Chen (1976). A good account of probabilistic distance and dependence measures can be found in the book by Devijver and Kittler (1982).