12

Distance Measures in DNA Microarray Data Analysis.

R. Gentleman, B. Ding, S. Dudoit, and J. Ibrahim

Abstract

Both supervised and unsupervised machine learning techniques require selection of a measure of distance between, or similarity among, the objects to be classified or clustered. Different measures of distance or similarity will lead to different machine learning performance. The appropriateness of a distance measure will typically depend on the types of features being used in the learning process.

In this chapter, we examine the properties of distance measures in the context of the analysis of gene expression data from DNA microarray experiments. The feature vectors represent transcript levels, i.e., mRNA abundance or relative abundance, either across biological samples (if comparing genes) or across genes (if comparing samples).

We consider different aspects of distances that help address the heterogeneity of the data and differences in interpretation depending on the source of the data (cDNA arrays versus short oligonucleotide arrays). Traditional measures, such as Euclidean and Manhattan distances as well as correlation-based distances, are considered. Other dissimilarity functions, which involve comparisons of distributions based on the Kullback-Leibler and mutual information criteria, are also examined.

12.1 Introduction

Genomic experiments generate large and complex multivariate data sets. Machine learning approaches are important tools in microarray data analysis, for the purposes of identifying patterns in expression among genes and/or biological samples, and for predicting clinical or other outcomes using gene expression data. Chapters 13, 16, and 17 consider different aspects

of machine learning in more detail. We briefly review some of the concepts here as motivation for the discussion in this chapter.

Inherent in every machine learning approach is a notion of a distance or similarity between the objects to be clustered or classified. In general, any distance measure can be used with any machine learning algorithm. The choice of distance measure is probably more important than the choice of machine learning algorithm, and some attention should be paid to the selection of an appropriate measure for each problem. In this chapter, we describe distances in quite general terms and consider both their mathematical properties as well as their implementation in different R packages.

The notion of distance is explicit in clustering procedures that operate directly on a matrix of pairwise distances between the objects to be clustered, e.g., partitioning around medoid (PAM) and hierarchical clustering (Kaufman and Rousseeuw, 1990). Certain supervised learning methods, such as nearest neighbor classifiers, also involve explicitly specifying a distance. Although the choice of distance may not be as transparent for other supervised approaches, observations are in fact assigned to classes on the basis of their distances from objects known to be in the classes. For instance, linear discriminant analysis is based on the Mahalanobis distance [Mardia et al. (1979); Ripley (1996a); see Equation (12.3) below] of the observations from the class means. The weighted gene voting scheme of Golub et al. (1999) is a variant of a special case of linear discriminant analysis, also known as naive Bayes classification. In addition, the distance and its behavior are intimately related to the scale on which measurements are made. The choice of a transformation and distance should thus be made jointly and in conjunction with the choice of a classifier or clustering procedure.

In this chapter, we consider the impact of distance selection on the analysis of genomic data. We assume that the data have been preprocessed using appropriate techniques and normalization methods and that the researcher is presented with an array containing G features (genes) for I samples. For microarray data there are potentially two values per feature: an estimate of the abundance of mRNA for that gene and a standard error of estimated abundance.

Our development goes as follows. In the next section, we give a general introduction to distances and discuss specific classes of distances. We provide formal definitions and discuss the relevant resources available in R. Then in Section 12.3, we focus on gene expression data from Affymetrix and two-color cDNA microarray experiments and discuss standardization and issues specific to these two platforms. We provide some examples of the use of different distance measures, in particular we make use of literature co-citation data, in Section 12.4. Visualization methods for distance data are described in Chapter 10.

12.2 Distances

Distances, metrics, dissimilarities, and similarities are related concepts. We provide some general definitions and then consider specific classes of distance measures.

12.2.1 Definitions

Any function d that satisfies the following five properties is termed a *metric*:

(i) non-negativity $d(\mathbf{x}, \mathbf{y}) \geq 0$;

(ii) symmetry $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$;

(iii) identification mark $d(\mathbf{x}, \mathbf{x}) = 0$;

(iv) definiteness $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$;

(v) triangle inequality $d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \geq d(\mathbf{x}, \mathbf{z})$.

A function that satisfied only properties (i)-(iii) is termed a *distance*. For many of the techniques we will consider, distances are sufficient. Hence, we will generally refer to distances (which include metrics) and only mention metrics specifically when properties (iv) and (v) are relevant.

A *similarity function* S is more loosely defined and satisfies the three following properties

(i) non-negativity $S(\mathbf{x}, \mathbf{y}) \geq 0$;

(ii) symmetry $S(\mathbf{x}, \mathbf{y}) = S(\mathbf{y}, \mathbf{x})$;

(iii) $S(\mathbf{x}, \mathbf{y})$ increases in a monotone fashion as objects $\mathbf{x}$ and $\mathbf{y}$ are more and more *similar*.

A *dissimilarity* function satisfies (i) and (ii), but for (iii), $S(\mathbf{x}, \mathbf{y})$ increases as objects $\mathbf{x}$ and $\mathbf{y}$ are more and more dissimilar. It is worth noting that there is, in fact, no need to require symmetry although some adjustments generally need to be made if the measures are not symmetric. The airplane flight time between two cities is an example of an asymmetric distance.

Many options are available in selection of a distance for machine learning tasks. Because there are many different types of data (e.g., ordinal, nominal, continuous) and approaches for analyzing these data, the literature on distances is quite broad. References that consider the application of distances in either clustering or classification include: Duda et al. (2001, Section 4.7); Gordon (1999, Chapter 2); Kaufman and Rousseeuw (1990, Chapter 1); (Mardia et al., 1979, Chapter 13).

As noted above, we are most concerned with a situation where G features have been measured for I observations, or samples. There is substantial interest in applying some form of machine learning to both the samples

(e.g., to identify patients with similar patterns of mRNA expression) and the features (e.g., to identify genes with similar patterns of expression).

We distinguish between two main classes of distance measures. Consider computing the distance between the expression profiles of two genes across I samples. In the first approach, we view the gene expression profiles as two I-vectors in some space and compute distances in a pairwise (within-sample) manner (Section 12.2.2). In contrast, the second approach ignores the natural pairing of observations and instead, views the two gene expression profiles as two different samples generated from underlying probability density functions for mRNA expression measures. In this case, distances between densities or distribution functions are relevant (Section 12.2.3). Of course, one is certainly not limited to an either-or approach. It may, in fact, be quite sensible to devise measures that combine the two. Genes with expression patterns that are similar in both aspects are possibly more interesting than those that are close in only one.

12.2.2 Distances between points

For m-vectors $\mathbf{x} = (x_1, \ldots, x_m)$ and $\mathbf{y} = (y_1, \ldots, y_m)$ consider distances of the form

$$d(\mathbf{x}, \mathbf{y}) = F[d_1(x_1, y_1), \ldots, d_m(x_m, y_m)], \tag{12.1}$$

where the d_k are themselves distances for each of the $k = 1, \ldots, m$ features. We refer to these functions as *pairwise distance functions*, as the pairing of observations within features is preserved. This representation is quite general: there is no need for the d_k to be the same. In particular, features may be of different types (e.g., the data may consist of a mixture of continuous and binary features) and may be weighed differentially (e.g., weighted Euclidean distance).

Common metrics within this class include the Minkowski metric, with $z_k = d_k(x_k, y_k) = |x_k - y_k|$ and $F(z_1, \ldots, z_m) = (\sum_{k=1}^{m} z_k^{\lambda})^{1/\lambda}$. Special cases of the Minkowski metric considered in this chapter are the Manhattan and Euclidean metrics corresponding to $\lambda = 1$ and $\lambda = 2$, respectively.

EUC Euclidean metric

$$d_{euc}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}. \tag{12.2}$$

MAN Manhattan metric

$$d_{man}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{m} |x_i - y_i|.$$

Correlation-based distance measures have been widely used in the microarray literature (Eisen et al., 1998). They include one minus the standard

Pearson correlation coefficient and one minus an uncentered correlation coefficient (or cosine correlation coefficient) considered by Eisen et al. (1998), Spearman's rank correlation, and Kendall's τ (Conover, 1971).

COR Pearson sample correlation distance

$$d_{cor}(\mathbf{x}, \mathbf{y}) = 1 - r(\mathbf{x}, \mathbf{y}) = 1 - \frac{\sum_{i=1}^{m}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \overline{x})^2 \sum_{i=1}^{m}(y_i - \overline{y})^2}}.$$

EISEN Cosine correlation distance

$$d_{eisen}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x}'\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = 1 - \frac{|\sum_{i=1}^{m} x_i y_i|}{\sqrt{\sum_{i=1}^{m} x_i^2 \sum_{i=1}^{m} y_i^2}}$$

which is a special case of Pearson's correlation with $\overline{x}$ and $\overline{y}$ both replaced by zero.

SPEAR Spearman sample correlation distance

$$d_{spear}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\sum_{i=1}^{m}(x'_i - \overline{x'})(y'_i - \overline{y'})}{\sqrt{\sum_{i=1}^{m}(x'_i - \overline{x'})^2 \sum_{i=1}^{m}(y'_i - \overline{y'})^2}}.$$

where $x'_i = rank(x_i)$ and $y'_i = rank(y_i)$.

TAU Kendall's τ sample correlation

$$d_{tau}(\mathbf{x}, \mathbf{y}) = 1 - \tau(\mathbf{x}, \mathbf{y})| = 1 - \frac{|\sum_{i=1}^{m}\sum_{j=1}^{m} C_{x_{ij}} C_{y_{ij}}}{m(m-1)}$$

where $C_{x_{ij}} = sign(x_i - x_j)$ and $C_{y_{ij}} = sign(y_i - y_j)$.

Note that we have transformed the correlations by subtracting them from one. This is done so that two vectors that are strongly positively correlated are regarded as close together. Using this transformation, data that exhibit a strong negative correlation will be far apart. In some cases, you might want to treat negative and positive correlations similarly, and that can be achieved by using the absolute value of the correlation. Correlation-based measures are in general invariant to location and scale transformations and tend to group together genes whose expression patterns are linearly related. While correlation-based distances have many nice properties, they tend to be adversely affected by outliers and then the non-parametric versions (SPEAR or TAU) are preferred.

When the data are standardized using the mean and variance, so that both $\mathbf{x}$ and $\mathbf{y}$ are m-vectors with zero mean and unit length, there is a functional relationship between the Pearson correlation coefficient $r(\mathbf{x}, \mathbf{y})$ and the Euclidean distance. The relationship is

$$d_{euc}(\mathbf{x}, \mathbf{y}) = \sqrt{2m[1 - r(\mathbf{x}, \mathbf{y})]}.$$

We note that expression values are generally measured with error. The standard deviation of measurement errors can be estimated and is sometimes available along with intensity measures in the form of "standard

errors." This variability information can be exploited in *errors-in-variables* models. This is the approach taken by Tadesse et al. (2005) for modeling survival data. Estimated standard errors can also be used when considering Kullback-Leibler distances, as is shown below.

Finally, we mention the *Mahalanobis distance*. Consider a situation where a pair of vectors, $\mathbf{x}$ and $\mathbf{y}$, are generated from some multivariate distribution with mean vector μ and variance-covariance matrix $\mathbf{\Sigma}$. Then the Mahalanobis distance between them is defined as

$$(\mathbf{x} - \mathbf{y})' \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{y}). \tag{12.3}$$

When $\mathbf{\Sigma}$ is unknown, it is generally replaced with the sample variance-covariance matrix. In general terms, the Mahalanobis distance reflects the notion that the data are more variable in some directions than in others.

Distances and transformations. Distances and data transformations are closely related. If g is an invertible, possibly non-linear, transformation $g : x \to x'$, then this can be used to induce a new metric d' via

$$d(\mathbf{x}, \mathbf{y}) = d[g^{-1}(\mathbf{x}'), g^{-1}(\mathbf{y}')] = d'(\mathbf{x}', \mathbf{y}').$$

The metric d operates on the original variables $\mathbf{x}$, whereas d' works on the transformed variables $\mathbf{x}'$, and the two are equivalent, even though they can have quite different functional forms. Conversely, the same distance function, say d_{euc} from Equation (12.2), can lead to quite different distances, between the same data points, when applied on different scales. Hence the choice of the scale is important. For microarray data, at least three different scales are generally considered: that of the original scanned fluorescence intensities, the logarithmically transformed scale, or the generalized logarithmic (variance-stabilized) scale proposed by Huber et al. (2002) and Durbin et al. (2002). A more general discussion of transformations in regression can be found, for example, in Ryan (1997).

Practicalities. Many pairwise distances can be computed in R using the `dist` function, including `euclidean`, `manhattan`. The function returns an object of class *dist*, which represents the distances between the rows of the input argument (which can be either a matrix or a dataframe). This function assumes that distances are symmetric and saves storage space by using a lower-triangular representation for the returned value.

The function `daisy` in the `cluster` package also provides distance computations. This function returns an object of class *dissimilarity* which contains the distances between the rows of the input matrix or dataframe. This class *inherits* from the *dist* class so that it will automatically use methods appropriate to that class. When some of the input variables are categorical, such as sex, then it makes no sense to compute distances between the numerical encodings and `daisy` has functionality to compute appropriate between-observation distances.

The package bioDist has implementations of the various correlation distances, such as `spearman.dist` and `tau.dist`. These functions return objects of class *dist*.

The functions in cluster take either a data matrix or a dissimilarity matrix as input. Other machine learning algorithms are less flexible and may require that the user manipulate the data in order to alter the distance measure that is used. An approach toward standardization is considered in Chapter 16.

12.2.3 Distances between distributions

The distances enumerated in the preceding section treat the expression measurements as points in some metric space, where each observation (gene or sample, depending on the problem) contributes one point and the coordinates are given by the corresponding expression measures. Distances are computed in a pairwise manner within features (samples when genes are being compared, and vice versa). A different approach is to consider the data for each feature as an independent sample from a population. In this case, we are interested in questions such as whether the shape of the distribution of features is similar between two genes. For example whether they are bimodal or, perhaps have long right-tails. Other authors have also considered using distances between distributions as a means of analyzing genomic data. For example, Butte and Kohane (2000) suggest binning the data and then using a mutual information distance. Quite a different approach to the comparison of distributions is taken in Gentleman and Carey (2003, Section 2.4.3); see also Section 16.4.6 below.

Alternatively, for each gene, across samples, we can consider the data as random I-vectors from some distribution. The simplest case is to assume that the expression measures for a particular gene follow an I-dimensional multivariate normal distribution with diagonal variance-covariance matrix. Using this approach, each gene provides a multivariate observation. Each of the I measurements for a given gene come from different samples, which are assumed to be independent, and hence the estimated variance-covariance matrix is diagonal. This approach can be used when both expression levels and their associated standard errors are available. The observed expression values are used to estimate the mean vector and the observed standard errors are used to estimate the variance-covariance matrix.

Many different distance measures can be used to assess the similarities between two densities. We consider two measures that are not actually distances: the Kullback-Leibler information and Hamming's mutual information.

Kullback-Leibler Information. The *Kullback-Leibler Information* (KLI) measure between densities f_1 and f_2 is defined as

$$\begin{aligned} KLI(f_1, f_2) &= E_1 \{\log[f_1(X)/f_2(X)]\} \\ &= \int \log[f_1(x)/f_2(x)] f_1(x) dx, \end{aligned} \tag{12.4}$$

where X is a random variable with density f_1, and E_1 denotes expectation with respect to f_1. This ratio can be infinite and hence so can the KLI. The KLI is not a distance because it is not symmetric. KLI does not satisfy the triangle inequality either.

The KLI can be symmetrized in a number of ways, including the approach described in Cook and Weisberg (1982, p. 163). They define the *Kullback-Leibler Distance* (KLD) to be,

$$2d_{KLD}(f_1, f_2) = KLI(f_1, f_2) + KLI(f_2, f_1).$$

The measure is symmetric and positive if f_1 and f_2 are different, however, it still does not satisfy the triangle inequality.

In the special case where $f_1 = N_m(\mu_1, \Sigma_1)$ and $f_2 = N_m(\mu_2, \Sigma_2)$, and assuming that Σ_1 and Σ_2 are positive definite, the expression for $d_{KLD}(f_1, f_2)$ simplifies and we get:

$$\begin{aligned} 2d_{KLD}(f_1, f_2) &= (\mu_1 - \mu_2)^T \Sigma_2^{-1} (\mu_1 - \mu_2) \\ &\quad + \log(|\Sigma_1|/|\Sigma_2|) + \mathrm{tr}(\Sigma_1 \Sigma_2^{-1}) - m. \end{aligned} \tag{12.5}$$

However, this simplification involves making a strong assumption and requires knowledge of both variance-covariance matrices. Note that if Σ_1 and Σ_2 are identical, this is a form of Mahalanobis distance. However, we should emphasize that the treatment here is slightly different.

To compute between gene distances from microarray data, the expression measures for a given gene, across samples, can be treated as a single observation from an I-dimensional multivariate normal distribution. For each gene, we estimate the mean in each coordinate (sample) by the observed expression measure for that sample, and we estimate the variances using, for example, the Li and Wong estimated standard errors for Affymetrix data (Li and Wong, 2001a). When viewed from this perspective, KLD (Equation 12.5) is more similar to the distances in Section 12.2.2 than it is to either KLI or the mutual information distance described below. This is a model that accounts for measurement error, though not as explicitly as an errors-in-variables approach.

Mutual Information. Closely related to the KLI is the *mutual information* (MI). The MI measures the extent to which two random variables X and Y are dependent. Let $f(\cdot, \cdot)$ denote the joint density function and $f_1(\cdot)$ and $f_2(\cdot)$ the two marginal densities for X and Y, respectively. Then the MI is defined as

$$MI(f_1, f_2) = E_f \left\{ \log \left[\frac{f(X, Y)}{f_1(X) f_2(Y)} \right] \right\}, \tag{12.6}$$

and is zero in the case of independence. We note that like KLI, MI is not a distance although we will sometimes refer to it as if it were. This can easily be determined by noticing the relationship between the MI distance and the KLI. The MI is basically the KLI between $f(x, y)$ and $g(x, y) = f_1(x)f_2(y)$, where $g(x, y)$ is the joint distribution obtained by assuming that the two marginals are independent,

$$\begin{aligned} KLI(f, g) &= \int_x \int_y \log[f(x, y)/g(x, y)]f(x, y)dxdy \\ &= E_f \left\{ \log \left[\frac{f(X, Y)}{f_1(X)f_2(Y)} \right] \right\} \\ &= MI(f_1, f_2). \end{aligned}$$

MI and KLD focus on very different aspects of distributions and that is reflected in their performance. MI is large when the joint distribution is quite different from the product of the marginals. Thus, it attempts to measure the distance from *independence*. KLD, on the other hand, measures how much the shape of one distribution resembles that of the other.

Joe (Joe, 1989) considers MI and its role as a multivariate measure of association. He shows that if the transformation,

$$\delta^* = [1 - \exp(-2MI)]^{1/2} \tag{12.7}$$

is used, then δ^* takes values in the interval $[0, 1]$ and can be interpreted as a a generalization of the correlation. He further notes that δ^* is related to Kendall's τ. We will make the further transformation to $1 - \delta^*$ so that our measure has the same interpretation as the other correlation-based distance measures discussed in this chapter.

Practicalities. The distances being considered are functionals of the underlying probability density functions. Given the observed data, there are many different methods for providing the appropriate estimates. We consider three of the more commonly used methods in this chapter. The simplest method is to assume some parametric distribution for the data and to estimate the parameters for that distribution; these can then be used, together with the functional form of the density, to estimate the mutual information. A second approach is to roughly group the data and to then treat it as discrete. A third approach is to use density estimation followed by either numerical integration or explicit calculation. The second and third approaches involve some form of smoothing, and this should be dealt with explicitly. Much work remains to be done before any method can be recommended for general use.

To apply the binning approach, the samples are separately divided into k common bins and then each sample is treated as if it were data from a discrete distribution. This approach can be problematic, as the estimated KLI will be infinite whenever a bin has an observation from f_1 but not one

from f_2. In our experience, this occurs quite often. We note that there are other problems with the binning approach; a straightforward calculation shows that the binned version of MI distance tends to the logarithm of the number of sample points as the number of bins goes to infinity, since in the limit every point will end up in a bin of its own.

An alternative procedure is to employ a density estimation procedure followed by numerical integration. One could standardize the data (shift so that a measure of central location is approximately zero and scale so that a measure of dispersion is approximately unity), estimate the densities and then apply numerical integration (using the range -3 to 3) to estimate KLI in Equation (12.4). This approach could be extended to MI as well. There are many good density estimation routines available in R, and one-dimensional integration is straightforward. In our examples for MI, we used the binning approach because density estimation followed by numerical integration proved too computationally expensive.

We have created the bioDist package, which contains code for some of the methods described here. It is used in the examples given later in this chapter. bioDist contains an implementation of the KL distances that rely on binning; `KLdist.matrix` and one that uses density estimation followed by numerical integration, `KLD.matrix` . For mutual information there are two functions, `mutualInfo` that computes the distance from independence and `MIdist` that computes the transformation in Equation (12.7). We note that the computations are not terribly fast computing these distances on very large data sets is time consuming.

12.2.4 Experiment-specific distances between genes

The between-gene distances considered thus far do not take into account the *structure* or *design* of the microarray experiment. Such distance measures may be appropriate for situations where there is no particular structure of interest among the arrays, e.g., when the target samples hybridized to the arrays are viewed as a random sample from a particular population. However, microarray experiments can be highly structured, as in time-course and multifactorial experiments. It is desirable to derive between-gene distances that reflect the design of the experiment under consideration. Such distances may serve to produce a more vivid visualization of the data and to permit focus on more meaningful patterns in gene expression. In this section, we consider some modifications that are more suitable for data arising from designed experiments or other situations where the samples have specific relationships to one another.

Instead of computing distances directly on the genes-by-arrays data matrices, one may use covariate information (e.g., treatment, cell type, dose, time) to derive suitable transformations of this matrix. Linear models and extensions thereof (e.g., generalized linear models) can be used to estimate experiment specific effects for each gene and hence produce new gene pro-

files. For factorial experiments studying the simultaneous gene expression response to two treatments, say, the new profiles could be based on main effects and interactions. In time-course experiments, it makes sense to consider distances that are not time-exchangeable and use the time index in an essential way. This could be done by penalizing for non-smoothness as in Sobolev metrics, where the squared Sobolev distance between two functions is based on the sum of squared distances, in some standard metric (e.g., $\mathcal{L}_2$), between the two functions, their first derivatives, second derivatives, etc., up to some order p. For time-course data with a large enough number of equally spaced time points, one of the standard wavelet decompositions could be used to decompose expression profiles into potentially interpretable quantities corresponding to local frequency components.

The use of covariate information as described above produces new profiles for each gene. Distances can then be computed for the new profiles, and genes can be clustered based on these distances. A preliminary application of such an approach can be found in Lin et al. (2004), for a study of spatial differential expression in the mouse olfactory bulb experiment. Distances on the new profiles can also be used to match profiles to a library of profiles of interest for a particular experiment, by ranking projections of the new gene profiles along specified directions in an appropriate geometric representation of the problem. For instance, in factorial experiments across time, interesting reference profiles for main effects and interactions might include: cyclical, early, or late effects, or the effects over time for a known gene.

12.3 Microarray data

For our purpose, gene expression data on G genes for I mRNA samples may be summarized by a $G \times I$ matrix $X = (x_{gi})$, where x_{gi} denotes the expression measure of gene g in mRNA sample i. The expression levels might be either absolute (e.g., Affymetrix oligonucleotide arrays) or relative to the expression levels of a suitably defined common reference sample (e.g., cDNA microarrays.)

12.3.1 Distances and standardization

The behavior of the distance is closely related to the scale on which the observations have been made. Standardization of features is thus an important issue when considering distances between objects and is one method of making the features comparable. However, standardization also has the effect of removing some of the potentially interesting features in the data. Thus, in some cases it will be sensible to explore other approaches to obtaining comparability across features.

In the context of microarray data, one may standardize genes and/or samples. When standardizing genes, expression measures are transformed as follows

$$x_{gi} = \frac{x_{gi} - \text{center}(x_{g.})}{\text{scale}(x_{g.})}$$

where $\text{center}(x_{g.})$ is some measure of the center of the distribution of the set of values x_{gi}, $i = 1, \ldots, I$, such as mean or median, and $\text{scale}(x_{g.})$ is a measure of scale such as the standard deviation, interquartile range, or MAD (median absolute deviation about the median). Alternatively, one may want to standardize arrays (samples) if there is interest in clustering or classifying them (rather than clustering or classifying the genes). Now we use

$$x_{gi} = \frac{x_{gi} - \text{center}(x_{.i.})}{\text{scale}(x_{.i})},$$

where the centering and scaling operations are carried out across all genes measured on sample (or array) i.

We now consider the implications of the preceding discussion on standardization in the context of both relative mRNA expression measurements (cDNA) and absolute (Affymetrix) mRNA expression measurements. Consider the standard situation where x_{gi} represents the expression measure on a log scale for gene g on patient (i.e., array or sample) i. Let $y_{gi} = x_{gi} - x_{gA}$, where patient A is our reference. Then, the relative expression measures y_{gi} correspond to the standard data available from a cDNA experiment with a common reference. The use of relative expression measures represents a location transformation for each gene (gene centering). Now, suppose that we want to measure the distance between patient samples i and j. Then, for the classes of distances considered in Equation (12.1) of Section 12.2.2,

$$d(\mathbf{y}_{.i}, \mathbf{y}_{.j}) = \sum_{g=1}^{G} d_g(y_{gi}, y_{gj}) = \sum_{g=1}^{G} d_g(x_{gi} - x_{gA}, x_{gj} - x_{gA}).$$

When the $d_g(x, y)$ are functions of $x - y$ alone, then $d(\mathbf{y}_{.i}, \mathbf{y}_{.j}) = d(\mathbf{x}_{.i}, \mathbf{x}_{.j})$, and it does not matter if we look at relative (the $\mathbf{y}$'s) or absolute (the $\mathbf{x}$'s) expression measures.

Suppose that we are interested instead in comparing genes and not samples. Then the distance between genes g and h is

$$d(\mathbf{y}_{g.}, \mathbf{y}_{h.}) = \sum_{i=1}^{I} d_i(y_{gi}, y_{hi}) = \sum_{i=1}^{I} d_i(x_{gi} - x_{gA}, x_{hi} - x_{hA}).$$

If $d(\mathbf{x}, \mathbf{y})$ has the property that $d(\mathbf{x} - \mathbf{c}, \mathbf{y}) = d(\mathbf{x}, \mathbf{y})$ for any c, then the distance measure is the same for absolute and relative expression measures.

Thus, for Minkowski distances, the distance between samples is the same for relative and absolute expression measures. This does not hold for the

distance between genes. On the other hand, distances based on the Pearson correlation yield the same distances between genes for both relative and absolute measures. This does not hold for the distance between samples. Arguments can be made in favor of either approach: invariance of (i) gene distances or (ii) sample distances, for absolute and relative expression measures. The data analyst will have to weigh these and other biological considerations when selecting a distance measure.

12.4 Examples

For our examples in this chapter we make use of the data reported in Chiaretti et al. (2004) and described in Appendix A. We consider only the subset of patients that have a reciprocal translocation between the long arms of Chromosomes 9 and 22 that has been causally related to chronic and acute leukemia (Cilloni et al., 2002). They are labeled `BCR/ABL`.

We select genes for our distance measurements by first carrying out a non-specific filtering (as described in Chapter 14) where we imposed three requirements: the gene must have an expression level greater than log(100) in at least 25% of the samples, it must have an IQR that is larger than 0.5, and it must have median expression level greater than log(300). Genes that passed all three filters will be referred to as *expressed*. We then adjusted each gene across samples by subtracting the median and dividing by the MAD (median absolute deviation from the median). This step makes computations between genes and across different distance measures more comparable. By standardizing the genes, we have made the four distances, EISEN, COR, EUC, and MAN, more similar than they would be if we worked with untransformed data.

The code below shows how we constructed our filters, using genefilter and then the resulting manipulations, to restrict the data to those selected genes. Finally, we standardize the genes, across samples, as described above.

```
> library("genefilter")
> data(ALL)
> Bsub <- (ALL$mol == "BCR/ABL")
> Bs <- ALL[, Bsub]

> f1 <- pOverA(0.25, log2(100))
> f2 <- function(x) (IQR(x) > 0.5)
> f3 <- function(x) (median(2^x) > 300)
> ff <- filterfun(f1, f2, f3)
> selected <- genefilter(Bs, ff)
> sum(selected)

[1] 637

> BSub <- Bs[selected, ]
> eS <- exprs(BSub)
```

```
> mads <- apply(eS, 1, mad)
> meds <- apply(eS, 1, median)
> e1 <- sweep(eS, 1, meds)
> e2 <- sweep(e1, 1, mads, FUN = "/")
> BSubStd <- BSub
> exprs(BSubStd) <- e2
```

We now show how some of the distance measures we have discussed can be applied to the ALL data. In order to have a small set of genes to work with, we select genes that are in the GO BP category `GO:0006917`, which corresponds to the induction of apoptosis.

```
> library("GO")
> library("annotate")
> GOTERM$"GO:0006917"
```

```
GOID = GO:0006917
Term = induction of apoptosis
Synonym = apoptosis signaling
Synonym = positive regulation of apoptosis
Definition = A process that directly activates any
     of the steps required for cell death by
     apoptosis.
Ontology = BP
```

```
> library("hgu95av2")
> apop <- hgu95av2GO2ALLPROBES$"GO:0006917"
> inboth <- apop %in% row.names(e2)
> whsel <- apop[inboth]
> exprApop <- e2[whsel, ]
> unlist(mget(whsel, hgu95av2LOCUSID))
```

```
  36199_at   39020_at  2031_s_at   39723_at    1635_at
      1611      10572       1026       8454         25
 1636_g_at   39730_at   34740_at 41763_g_at   38050_at
        25         25       2309       7073       9774
```

Next we load the bioDist package and compute some pairwise distances between probesets.

```
> library("bioDist")
> man <- dist(exprApop, "manhattan")
> MI <- MIdist(exprApop)
> KLsmooth <- KLD.matrix(exprApop)
> KLbin <- KLdist.matrix(exprApop)
```

False color representations of the distance matrices are shown in Figure 12.1. We have used the transformation of mutual information distance described in Equation (12.7). The KL distances are small the more similar the shape of the two densities and are larger if the shapes are quite different.

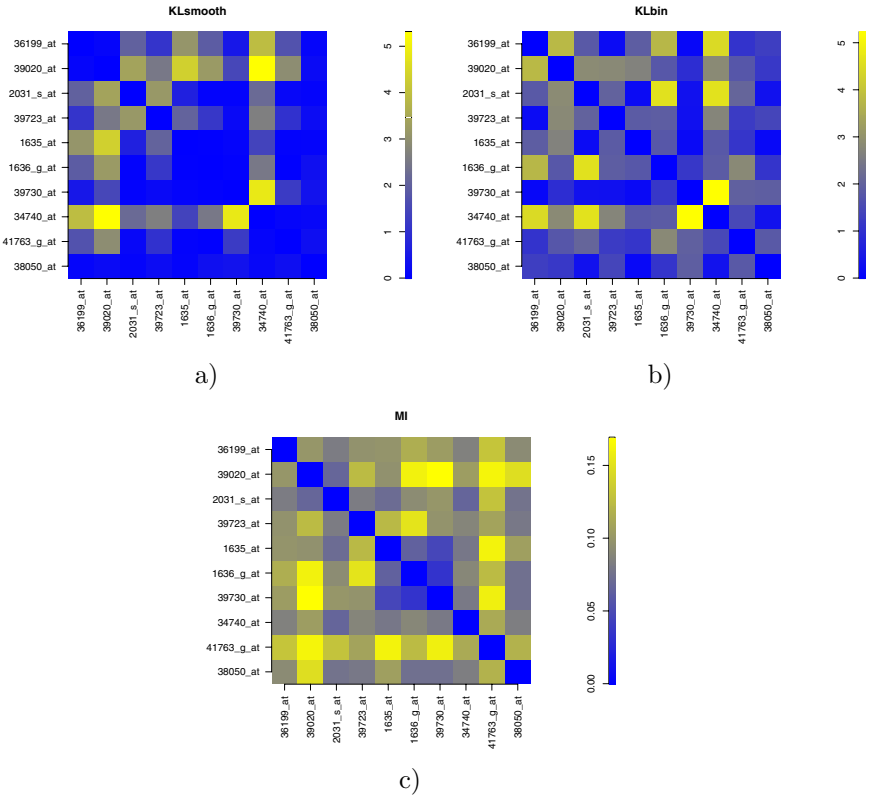

Figure 12.1. False color representation of the distance matrices. a) `KLsmooth`, b) `KLbin`, c) `MI`.

To further compare the distances, we produced pairwise scatterplots of the different distances in Figure 12.2. From that we can see the general positive correlation of the KL based distances and note that, as expected, there is little relationship between the MI distance and the KL distances – they are measuring different things.

12.4.1 A co-citation example

We now consider an example that relates distances and co-citation in the medical literature as measures of biological similarity. This approach can be contrasted with the one taken in Chapter 22. Two or more genes that share a common reference (i.e. were written about in the same paper) are more likely to be meaningfully biologically related than genes that are never jointly mentioned in any paper. Joint mention of genes A and B does not imply that these genes are strongly or even remotely biologically related.

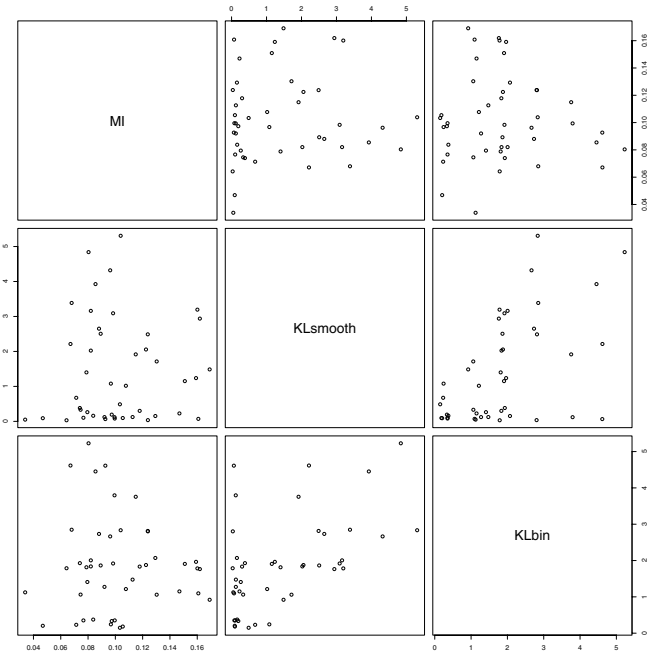

Figure 12.2. Pairwise scatterplots of different distances computed between the same set of points.

However, the data resource is large, there have been a number of studies that make use of co-citation data, and it is an active research area (Jenssen et al., 2001; Masys et al., 2001). Our approach is simple but could easily be extended to make use of other data sources as they become available.

Data on co-citation were obtained from PubMed. See Chapter 7 for more details on this reference database. We mapped Affymetrix identifiers to their corresponding LocusLink values and from there to PubMed identifiers (PMIDs).

We distinguish two relationships between genes that can be identified from these data. Any two genes that directly share a citation are called *adjacent*. Two genes will be called *accessible* if they can be connected, possibly by other genes, through a co-citation path. To be specific, suppose that genes X and Y are co-cited and that X and Z are also co-cited. Then we would say that X and Y are adjacent (as are X and Z) and that Y and Z are accessible. In the literature co-citation context, accessibility is too weak a relationship to explore further.

Next we selected a target gene and then determined the 100 genes closest to that target using the different distance measures under consideration, namely COR, SPEAR, TAU, EUC, MAN, KLD, and MI. We first look at

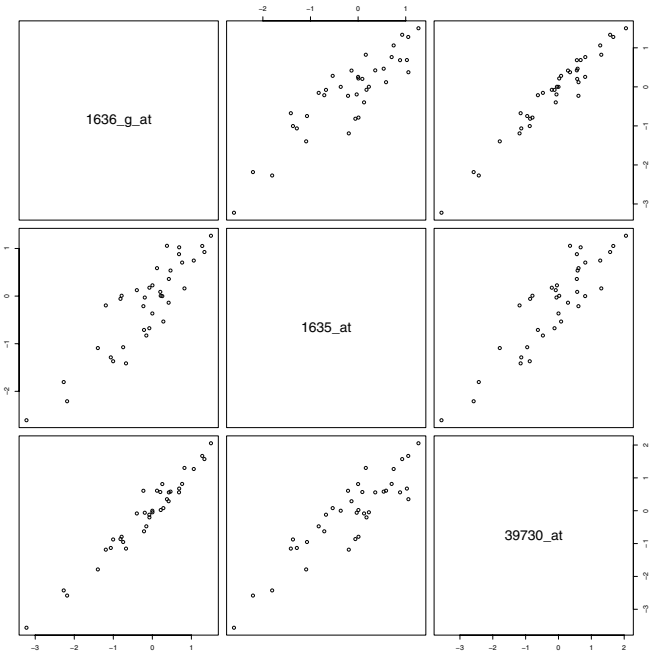

Figure 12.3. Scatterplot of multiple expressed probesets for ABL1

the agreement among distances in terms of percentages of common genes chosen.

As an example, we selected ABL1 as the target gene, which has 8 probesets on the HG-U95Av2 GeneChip. The filtering steps left us with 3 probesets that satisfied our selection criterion (see Figure 12.3). We emphasize the fact that these three probesets should be measuring the same thing and certainly the correlations, Figure 12.3, do appear quite strong. So we anticipate that starting with any one of them, we would find that the other two were close to the one we started with. To test this, we carried out the following experiment.

For each of the 3 expressed probesets a list of the 100 closest probesets, using each of the 7 distance measures, was computed. The between-list agreement, in terms of whether a probeset was selected or not, is shown next.

```
$"1636_g_at"
       cor spear  tau  euc  man  kld
spear 0.73
tau   0.72  0.94
euc   0.74  0.66 0.67
man   0.65  0.65 0.66 0.81
```

```
kld   0.19  0.15 0.13 0.10 0.09
mi    0.23  0.26 0.28 0.26 0.32 0.13

$"1635_at"
       cor spear  tau  euc  man  kld
spear 0.85
tau   0.85  0.93
euc   0.75  0.71 0.69
man   0.69  0.70 0.69 0.84
kld   0.15  0.17 0.18 0.15 0.15
mi    0.35  0.36 0.36 0.39 0.41 0.21

$"39730_at"
       cor spear  tau  euc  man  kld
spear 0.77
tau   0.75  0.95
euc   0.78  0.71 0.71
man   0.69  0.74 0.74 0.81
kld   0.22  0.19 0.20 0.23 0.23
mi    0.29  0.31 0.30 0.31 0.34 0.22
```

As we can see, the agreement among the correlational distances (i.e. COR, SPEAR and TAU) and Minkowski metric distances (i.e., EUC and MAN) are good, especially those between the nonparametric correlational distances (i.e., SPEAR and TAU). There is not much commonality between the distributional distances (i.e., KLD and MI) and other distances and the agreement within the distributional distances is also low. For each of the multiple probesets, we further look at the rank of the other two probesets chosen by the various distance measures.

```
1636_g_at
     cor spear tau euc man kld mi
[1,]   2     2   2   2   2  NA 10
[2,]   1     1   1   1   1  NA  1
1635_at
     cor spear tau euc man kld mi
[1,]   2     1   2   1   1  NA 23
[2,]   1     2   1   2   2  55  1
39730_at
     cor spear tau euc man kld mi
[1,]   1     1   1   1   1  NA  1
[2,]   2     2   2   2   2  NA  3
```

We can see that whenever one of the multiple probesets for ABL1 was chosen as target probeset, the other two probesets for the same gene were always the top two probesets using correlational or metric distances. MI also captured the other multiple probesets in the top list although the ranks tended to be larger. Again we see that KLD fared worst. Note that here we used discrete versions of KLD (symmetrized) and MI. The missing

values reported for `kld` arise because the other probesets for ABL1 were not always chosen.

12.4.2 Adjacency

We now compare co-citation and distance. Recall that two genes are called adjacent if they are cited together in any article, and in a sense this is a measure of similarity. So we might then ask whether any of the distance measures under consideration reflect this same measure of similarity. To answer that question we examine how many of the co-cited genes are within the top 100 list for each of the different distances. We first subset the original data set to include probesets that have been cited, this reduced the number of probesets to 632. There were 38 genes that were co-cited with ABL1 in the humanLLMappings package, out of which 2 were in the filtered *ALL* data set. We then used the 7 distance measures to generate the top 100 probesets and computed the number of genes among those 100 closest that had co-citations with the target gene. The results are shown below

```
      1636_g_at 1635_at 39730_at
cor           0       1        1
spear         1       1        1
tau           1       0        1
euc           1       2        1
man           1       1        1
kld           0       1        0
mi            0       1        0
```

Notice that both EUC and MAN did quite well and that KLD seemed to fare the worse. The Hypergeometric distribution can be used to assess the significance of the results given above. Consider an urn with 629 balls in it. Of these 2 are colored white, the remainder are black. Under the null hypothesis that there is no relationship between co-citation and being close, as computed by one of the distances, then each selection of the 100 nearest genes is like drawing 100 balls from the urn and counting how many white ones were selected. The computation of p-values is easily carried out.

```
P(X >= 1) P(X >= 2) P(X >= 3)
   0.4068    0.0678    0.0040
```

Note that the assumption of independence among the enumerated events required for applicability of the Hypergeometric model is not tenable for these data. Thus this Hypergeometric computation should be regarded as a rough approximation to the truth.

12.5 Discussion

Distances are an integral part of all machine learning algorithms and hence play a central role in the analysis of most experimental data. The distance that is used for any particular task can have a profound effect on the output of the machine learning method, and it is therefore essential that users ensure that the same distance method is used when comparing machine learning algorithms.

It is also important that the investigator be able to select, and use, a distance that is appropriate for the task at hand. There is no single distance that is always relevant, and similarity can be measured in many ways. We find R to be a good platform for these sorts of analyses, as it has a wealth of built-in distance functions, and supports the addition of new distance functions straightforwardly.