

Appendix A

Proximity Measures

We provide some information on proximity measures in this appendix. *Proximity measures* are important in clustering, multi-dimensional scaling, and nonlinear dimensionality reduction methods, such as ISOMAP. The word *proximity* indicates how close objects or measurements are in space, time, or in some other way (e.g., taste, character, etc.). How we define the nearness of objects is important in our data analysis efforts and results can depend on what measure is used. There are two types of proximity measures: similarity and dissimilarity. We now describe several examples of these and include ways to transform from one to the other.

A.1 Definitions

Similarity measures indicate how alike objects are to each other. A high value means they are similar, while a small value means they are not. We denote the similarity between objects (or observations) i and j as s_{ij} . Similarities are often scaled so the maximum similarity is one (e.g., the similarity of an observation with itself is one, $s_{ii} = 1$). *Dissimilarity measures* are just the opposite. Small values mean the observations are close together and thus are alike. We denote the dissimilarity by δ_{ij} , and we have $\delta_{ii} = 0$. In both cases, we have $s_{ij} \geq 0$ and $\delta_{ij} \geq 0$.

Hartigan [1967] and Cormack [1971] provide a taxonomy of twelve proximity structures; we list just the top four of these below. As we move down in the taxonomy, constraints on the measures are relaxed. Let the observations in our data set be denoted by the set O , then the proximity measure is a real function defined on $O \times O$. The four structures S that we consider here are

- S1 S defined on $O \times O$ is Euclidean distance;
- S2 S defined on $O \times O$ is a metric;
- S3 S defined on $O \times O$ is symmetric real-valued;

S4 S defined on $O \times O$ is real-valued.

The first structure S1 is very strict with the proximity defined as the familiar *Euclidean distance* given by

$$\delta_{ij} = \sqrt{\sum_k (x_{ik} - x_{jk})^2}, \quad (\text{A.1})$$

where x_{ik} is the k -th element of the i -th observation.

If we relax this and require our measure to be a metric only, then we have structure S2. A dissimilarity is a metric if the following holds:

1. $\delta_{ij} = 0$, if and only if $i = j$.
2. $\delta_{ij} = \delta_{ji}$.
3. $\delta_{ij} \leq \delta_{it} + \delta_{tj}$.

The second requirement given above means the dissimilarity is symmetric, and the third one is the triangle inequality. Removing the metric requirement produces S3, and allowing nonsymmetry yields S4. Some of the later structures in the taxonomy correspond to similarity measures.

We often represent the interpoint proximity between all objects i and j in an $n \times n$ matrix, where the ij -th element is the proximity between the i -th and j -th observation. If the proximity measure is symmetric, then we only need to provide the $n(n - 1)/2$ unique values (i.e., the upper or lower part of the *interpoint proximity matrix*). We now give examples of some commonly used proximity measures.

A.1.1 Dissimilarities

We have already defined Euclidean distance (Equation A.1), which is probably used most often. One of the problems with the Euclidean distance is its sensitivity to the scale of the variables. If one of the variables is more dispersed than the rest, then it will dominate the calculations. Thus, we recommend transforming or scaling the data as discussed in [Chapter 1](#).

The Mahalanobis distance (squared) takes the covariance into account and is given by

$$\delta_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)^T \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j),$$

where Σ is the covariance matrix. Often Σ must be estimated using the data, so it can be affected by outliers.

The *Minkowski metric* defines a whole family of dissimilarities, and it is used extensively in nonlinear multidimensional scaling (see [Chapter 3](#)). It is defined by

$$\delta_{ij} = \left\{ \sum_k |x_{ik} - x_{jk}|^\lambda \right\}^{\frac{1}{\lambda}} \quad \lambda \geq 1. \quad (\text{A.2})$$

When the parameter $\lambda = 1$, we have the *city block metric* (sometimes called *Manhattan distance*) given by

$$\delta_{ij} = \sum_k |x_{ik} - x_{jk}|.$$

When $\lambda = 2$, then Equation A.2 becomes the Euclidean distance (Equation A.1).

The MATLAB Statistics Toolbox provides a function called **pdist** that calculates the interpoint distances between the n data points. It returns the values in a vector, but they can be converted into a square matrix using the function **squareform**. The basic syntax for the **pdist** function is

Y = pdist(X, distance)

The **distance** argument can be one of the following choices:

```
'euclidean'      - Euclidean distance
'seuclidean'     - Standardized Euclidean distance, each
                   coordinate in the sum of squares is inverse weighted
                   by the sample variance of that coordinate
'cityblock'      - City Block distance
'mahalanobis'    - Mahalanobis distance
'minkowski'      - Minkowski distance with exponent 2
'cosine'         - One minus the cosine of the included
                   angle between observations (treated as vectors)
'correlation'    - One minus the sample correlation
                   between observations (treated as sequences of
                   values).
'hamming'        - Hamming distance, percentage of
                   coordinates that differ
'jaccard'        - One minus the Jaccard coefficient, the
                   percentage of nonzero coordinates that differ
```

The **cosine**, **correlation**, and **jaccard** measures specified above are originally similarity measures, which have been converted to dissimilarities, two of which are covered below. It should be noted that the **minkowski**

option can be used with an additional argument designating other values of the exponent.

There is a new distance available with Version 5 of the Statistics Toolbox. This is the Chebychev distance, which is the maximum coordinate difference between two vectors. Use the flag '**chebychev**' to get this distance.

A.1.2 Similarity Measures

A common similarity measure that can be used with real-valued vectors is the *cosine similarity measure*. This is the cosine of the angle between the two vectors and is given by

$$s_{ij} = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\sqrt{\mathbf{x}_i^T \mathbf{x}_i} \sqrt{\mathbf{x}_j^T \mathbf{x}_j}} .$$

The *correlation similarity measure* between two real-valued vectors is similar to the one above, and is given by the expression

$$s_{ij} = \frac{(\mathbf{x}_i - \bar{\mathbf{x}})^T (\mathbf{x}_j - \bar{\mathbf{x}})}{\sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})^T (\mathbf{x}_i - \bar{\mathbf{x}})} \sqrt{(\mathbf{x}_j - \bar{\mathbf{x}})^T (\mathbf{x}_j - \bar{\mathbf{x}})}} .$$

A.1.3 Similarity Measures for Binary Data

The proximity measures defined in the previous sections can be used with quantitative data that are continuous or discrete, but not binary. We now describe some measures for binary data.

When we have binary data, we can calculate the following frequencies:

		<i>j</i> -th Observation		
		1	0	
<i>i</i> -th Observation	1	<i>a</i>	<i>b</i>	<i>a</i> + <i>b</i>
	0	<i>c</i>	<i>d</i>	<i>c</i> + <i>d</i>
		<i>a</i> + <i>c</i>	<i>b</i> + <i>d</i>	<i>a</i> + <i>b</i> + <i>c</i> + <i>d</i>

This table shows the number of elements where the *i*-th and *j*-th observations both have a 1 (*a*), both have a 0 (*d*), etc. We use these quantities to define the following similarity measures for binary data.

First is the *Jaccard coefficient*, given by

$$s_{ij} = \frac{a}{a + b + c} .$$

Next we have the *Ochiai measure*:

$$s_{ij} = \frac{a}{\sqrt{(a+b)(a+c)}}.$$

Finally, we have the *simple matching coefficient*, that calculates the proportion of matching elements:

$$s_{ij} = \frac{a+d}{a+b+c+d}.$$

A.1.4 Dissimilarities for Probability Density Functions

In some applications, our observations might be in the form of relative frequencies or probability distributions. For example, this often happens in the case of measuring semantic similarity between two documents. Or, we might be interested in calculating a distance between two probability density functions, where one is estimated and one is the true density function. We discuss three types of measures here: Kullback-Leibler information, L_1 norm, and information radius.

Say we have two probability density functions f and g (or any function in the general case). *Kullback-Leibler (KL) information* measures how well function g approximates function f . The KL information is defined as

$$KL(f, g) = \int f(\mathbf{x}) \log\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) d\mathbf{x},$$

for the continuous case. A summation is used in the discrete case, where f and g are probability mass functions. The KL information measure is sometimes called the *discrimination information*, and it measures the divergence between two functions. The higher the measure, the greater the difference between the functions.

There are two problems with the KL information measure. First, there could be cases where we get a value of infinity, which can happen often in natural language understanding applications [Manning and Schütze, 2000]. The other potential problem is that the measure is not necessarily symmetric.

The second measure we consider is the *information radius* (IRad) that overcomes these problems. It is based on the KL information and is given by

$$IRad(f, g) = KL\left(f, \frac{f+g}{2}\right) + KL\left(g, \frac{f+g}{2}\right).$$

The IRad measure tries to quantify how much information is lost if we describe two random variables that are distributed according to f and g with their average distribution. The information radius ranges from 0 for identical

distributions to $2\log 2$ for distributions that are maximally different. Here we assume that $0\log 0 = 0$. Note that the IRad measure is symmetric and is not subject to infinite values [Manning and Schütze, 2000].

The third measure of this type that we cover is the L_1 norm. It is symmetric and well-defined for arbitrary probability density functions f and g (or any function in the general case). We can think of it as a measure of the expected proportion of events that are going to be different between distributions f and g . This norm is defined as

$$L_1(f, g) = \int |f(\mathbf{x}) - g(\mathbf{x})| d\mathbf{x} .$$

The L_1 norm has a nice property. It is bounded below by zero and bounded above by two:

$$0 \leq L_1(f, g) \leq 2 ,$$

when f and g are valid probability density functions.

A.2 Transformations

In many instances, we start with a similarity measure and then convert to a dissimilarity measure for further processing. There are several transformations that one can use, such as

$$\begin{aligned} \delta_{ij} &= 1 - s_{ij} \\ \delta_{ij} &= c - s_{ij} \quad \text{for some constant } c \\ \delta_{ij} &= \sqrt{2(1 - s_{ij})} . \end{aligned}$$

The last one is valid only when the similarity has been scaled such that $s_{ii} = 1$. For the general case, one can use

$$\delta_{ij} = \sqrt{s_{ii} - 2s_{ij} + s_{jj}} .$$

In some cases, we might want to transform from a dissimilarity to a similarity. One can use the following to accomplish this:

$$s_{ij} = (1 + \delta_{ij})^{-1} .$$

A.3 Further Reading

Most books on clustering and multidimensional scaling have extensive discussions on proximity measures, since the measure used is fundamental to these methods. We describe a few of these here.

The clustering book by Everitt, Landau and Leese [2001] has a chapter on this topic that includes measures for data containing both continuous and categorical variables, weighting variables, standardization, and the choice of proximity measure. For additional information from a classification point of view, we recommend Gordon [1999].

Cox and Cox [2001] has an excellent discussion of proximity measures as they pertain to multidimensional scaling. Their book also includes the complete taxonomy for proximity structures mentioned previously. Finally, Manning and Schütze [2000] provide an excellent discussion of measures of similarity that can be used when measuring the semantic distance between documents, words, etc.

There are also several survey papers on dissimilarity measures. Some useful ones are Gower [1966] and Gower and Legendre [1986]. They review properties of dissimilarity coefficients, and they emphasize their metric and Euclidean nature. For a summary of distance measures that can be used when one wants to measure the distance between two functions or statistical models, we recommend Basseville [1989]. The work described in Basseville takes the signal processing point of view. Jones and Furnas [1987] study proximity measures using a geometric analysis and apply it to several measures used in information retrieval applications. Their goal is to demonstrate the relationship between a measure and its performance.