Introduction to Clustering

Philip Stallworth

October 15, 2015

Types of Data and How to Handle Them

Throughout this section, suppose that there are n objects to be clustered. Clustering algorithms are based on two matrix structures. The first is an $n \times p$ matrix where each row represents an object and each column represents a measurement or attribute. This is called an *objects-by-variables* matrix and is *two mode* because the rows and columns represent different things. The other type of object is an $n \times n$ proximity table. This is a *one mode* matrix where each entry, $x_{i,j}$, represents the proximity between objects i and j. We will consider proximities in terms of distances, dissimilarities, and similarities.

Interval-Scaled Variables

Suppose that the n objects are charazcterized by p continuous measurements. Measurements will often be real numbers. The measurements can be organized into an $n \times p$ objects-by-variables matrix, where each row corresponds to an object and each column to a variable. If we have 2 variables, the data lend themselves to visualization. However, clustering by eyeballing poses serious issues. In particular, different measurement units might yield very different clusters(see Kaufman and Rousseeuw pages 6, 7, 8).

One can standardize the data to alleviate this issue. First, calculate the mean of the variable f, given by

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf})$$

for each $f = 1, \ldots, p$. Then compute a measurement of the spread. Typically, people use standard deviation. But Kauffman and Rousseeuw use *mean absolute deviation*,

$$s_f = \frac{1}{n}\{|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|\}$$

because it is less sensitive to outliers. By less senstive, I mean outliers have less of an influence of mean absolute deviation than standard deviation. If we assume s_f is non-zero, the standardized measurements are defined as

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

. These measurements are unitless. After standardization, re-create the observations-by-factors matrix with z-scores as entries.

Standardizing isn't always ideal. Sometimes variables have an absolute meaning and essential information is lost through standardization. The text also has an interesting discussion about how measurements and weights are essentially the same. Standardization tries to give all variables the exact same weight.

We then compute dissimilarities between objects after creating the objects-by-variables matrix. In order to do so, we need to create a distance metric between any two objects i and j. Typically choises are:

1. Euclidean Distance
2. Manhattan Distance

Manhattan distance is helpful when, for instance, it doesn't matter whether we have a distance of 3 in one variable and 1 in the other or a distance of 2 in each variable. The Euclidean distance would find different distances for each. Any metric(real analysis definition) works as a distance measure. A good generalization of the Euclidean and Manhattan metrics is the *Minkowski distance*:

$$d(i,j) = (|x_{i1} - x_{ji}|^q + \ldots + |x_{ip} - x_{jp}|^j)^{1/q}$$

. The Minkowskki distance is also called the L_q metric. There are also weighted Euclidean distances. Suppose we have a variable that provides no relevant information. In an unweighted model, this variable will impact clustering even though it has no relevance. Consequently, we should give it 0 weight.

Dissimilarities

Now consider the $n \times n$ matrix with dissimilarities as entries. These entries might be distances between objects i and j. However, the don't necessarily have to be. A dissimilarity is any function which takes two objects, i and j, and returns a non-negative real number. The returned value should be close to 0 when the objects are similar and large when the objects are different. The difference between a dissimilarity and a distance is the triangle inequality. Dissimilarities don't require it. Dissimilarities can be computed from any numeric variable: binary, nominal, ordinal, interval, or a combination of these. They can just be subjective measures of 'difference'. They could also be levels of correlation between two sets of variables.

Similarities

Everything from dissimilarities apply. Except now distant objects have similarities near 0 and close objects have similarities neaer 1. Also, $s(i,i) = 1$.