Clustering Generalities

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Generalities

Typically, the object in "clustering" is to find natural groups of rows or columns of

$$\underset{N\times p}{\mathbf{X}} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_N' \end{pmatrix}$$

(In some contexts one may want to somehow find homogenous "blocks" in a properly rearranged $\mathbf{X}$.) Sometimes all columns of $\mathbf{X}$ represent values of continuous variables (so that ordinary arithmetic applied to all its elements is meaningful). But sometimes some columns correspond to ordinal or even categorical variables.

Here, we will smply let $\mathbf{x}_i$ $i = 1, 2, \ldots, r$ stand for "items" to be clustered (that might be rows or columns of $\mathbf{X}$ with entries that need not necessarily be continuous variables).

Dissimilarities between objects

In developing and describing clustering methods, one begins with a dissimilarity measure $d(\mathbf{x}, \mathbf{z})$ that (at least for the items to be clustered and perhaps for other possible items) quantifies how "unalike" items are. This measure is usually chosen to satisfy

1. $d(\mathbf{x}, \mathbf{z}) \geq 0 \ \forall \mathbf{x}, \mathbf{z}$
2. $d(\mathbf{x}, \mathbf{x}) = 0 \ \forall \mathbf{x}$, and
3. $d(\mathbf{x}, \mathbf{z}) = d(\mathbf{z}, \mathbf{x}) \ \forall \mathbf{x}, \mathbf{z}$.

It may be chosen to further satisfy

4. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{w}) + d(\mathbf{z}, \mathbf{w}) \ \forall \mathbf{x}, \mathbf{z}$, and $\mathbf{w}$, or

4′. $d(\mathbf{x}, \mathbf{z}) \leq \max[d(\mathbf{x}, \mathbf{w}), d(\mathbf{z}, \mathbf{w})] \ \forall \mathbf{x}, \mathbf{z}$, and $\mathbf{w}$.

Where 1-4 hold, d is a "metric." Where 1-3 hold and the stronger condition 4′ holds, d is an "ultrametric."

Two natural dissimilarity measures

In a case where one is clustering rows of $\mathbf{X}$ and each column of $\mathbf{X}$ contains values of a continuous variable, a squared Euclidean distance is a natural choice for a dissimilarity measure

$$d\left(\mathbf{x}_i, \mathbf{x}_{i'}\right) = \left\|\mathbf{x}_i - \mathbf{x}_{i'}\right\|^2 = \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2$$

In a case where one is clustering columns of $\mathbf{X}$ and each column of $\mathbf{X}$ contains values of a continuous variable, with $r_{jj'}$ the sample correlation between values in columns j and j', a plausible dissimilarity measure is

$$d\left(\mathbf{x}_j, \mathbf{x}_{j'}\right) = 1 - \left|r_{jj'}\right|$$

Proximity matrices

When dissimilarities between pairs of r items are organized into a (non-negative symmetric) $r \times r$ matrix

$$\mathbf{D} = (d_{ij}) = (d(\mathbf{x}_i, \mathbf{x}_j))$$

with 0's down its diagonal, the terminology "**proximity matrix**" is often used. For some clustering algorithms and for some purposes, the proximity matrix encodes all one needs to know about the items to do clustering. One seeks a partition of the index set $\{1, 2, \ldots, r\}$ into subsets such that the d_{ij} for indices within a subset are small (and the d_{ij} for indices i and j from different subsets are large).