Partitioning Methods of Clustering

Stephen Vardeman
Analytics Iowa LLC
ISU Statistics and IMSE

Centroid-based methods

By far the most commonly used clustering methods are based on partitioning related to "centroids," particularly the so called "K-Means" clustering algorithm for the rows of $\mathbf{X}$ in cases where the columns contain values of continuous variables x_j (for which arithmetic averaging makes sense). (In this context, a natural choice of $d(\mathbf{x}, \mathbf{z})$ is $\|\mathbf{x} - \mathbf{z}\|^2$. A fancier option might be built on squared Mahalanobis distance, $(\mathbf{x} - \mathbf{z})' \mathbf{Q} (\mathbf{x} - \mathbf{z})$ for some non-negative definite $\mathbf{Q}$.)

A first iteration of the K-means algorithm

The algorithm begins with some set of K distinct "centers" $\mathbf{c}_1^0, \mathbf{c}_2^0, \ldots, \mathbf{c}_K^0$. They might, for example, be a random selection of the rows of $\mathbf{X}$ (subject to the constraint that they are distinct). One then assigns each $\mathbf{x}_i$ to that center $\mathbf{c}_{k^0(i)}^0$ minimizing

$$d\left(\mathbf{x}_i, \mathbf{c}_l^0\right)$$

over choice of l (creating K clusters around the centers). One then replaces all of the $\mathbf{c}_k^0$ with the corresponding cluster means

$$\mathbf{c}_k^1 = \frac{1}{\#\text{ of } i \text{ with } k^0(i) = k} \sum I\left[k^0(i) = k\right] \mathbf{x}_i$$

m-th iteration of the K-means algorithm

At stage m with all $\mathbf{c}_k^{m-1}$ available, one then assigns each $\mathbf{x}_i$ to that center $\mathbf{c}_{k^{m-1}(i)}^{m-1}$ minimizing

$$d\left(\mathbf{x}_i, \mathbf{c}_l^{m-1}\right)$$

over choice of l (creating K clusters around the centers) and replaces all of the $\mathbf{c}_k^{m-1}$ with the corresponding cluster means

$$\mathbf{c}_k^m = \frac{1}{\# \text{ of } i \text{ with } k^{m-1}(i) = k} \sum I\left[k^{m-1}(i) = k\right] \mathbf{x}_i$$

This iteration goes on to convergence.

Multiple starts and comparison across K

One compares multiple random starts for a given K (and then minimum values found for each K) in terms of

$$\text{Total Within-Cluster Dissimilarity}\,(K) = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \text{ in cluster } k} d\left(\mathbf{x}_i, \mathbf{c}_k\right)$$

for $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K$ the final means produced by the iterations. (For a squared Euclidean distance d, this is a total squared distance of $\mathbf{x}_i$s to their corresponding cluster means.)

One may then consider the monotone sequence of Total Within-Cluster Dissimilarities and try to identify a value K beyond which there seem to be diminishing returns for increased K.

A first iteration of a *K*-mediods algorithm

A more general version of this algorithm (a "K-medoid" algorithm) doesn't require that the entries of the $\mathbf{x}_i$ be values of continuous variables, but (since it is then unclear that one can even evaluate, let alone find a general minimizer of, $d(\mathbf{x}_i, \cdot)$) restricts the "centers" to be original items.

This algorithm begins with some set of K distinct "medoids" $\mathbf{c}_1^0, \mathbf{c}_2^0, \ldots, \mathbf{c}_K^0$ that are a random selection from the r items $\mathbf{x}_i$ (subject to the constraint that they are distinct). One then assigns each $\mathbf{x}_i$ to that medoid $\mathbf{c}_{k^0(i)}^0$ minimizing

$$d\left(\mathbf{x}_i, \mathbf{c}_l^0\right)$$

over choice of l (creating K clusters associated with the medoids) and replaces all of the $\mathbf{c}_k^0$ with $\mathbf{c}_k^1$ the minimizers over the $\mathbf{x}_{i'}$ belonging to cluster k of the sums

$$\sum_{i \text{ with } k^0(i)=k} d\left(\mathbf{x}_i, \mathbf{x}_{i'}\right)$$

m-th iteration of the K-medoids algorithm

At stage m with all $\mathbf{c}_k^{m-1}$ available, one then assigns each $\mathbf{x}_i$ to that medoid $\mathbf{c}_{k^{m-1}(i)}^{m-1}$ minimizing

$$d\left(\mathbf{x}_i, \mathbf{c}_l^{m-1}\right)$$

over choice of l (creating K clusters) and replaces the $\mathbf{c}_k^{m-1}$ with $\mathbf{c}_k^m$ the minimizers over the $\mathbf{x}_{i'}$ belonging to cluster k of the sums

$$\sum_{i \text{ with } k^{m-1}(i)=k} d\left(\mathbf{x}_i, \mathbf{x}_{i'}\right)$$

This iteration goes on to convergence. One compares multiple random starts for a given K (and then minimum values found for K) in terms of

$$\sum_{k=1}^{K} \sum_{\mathbf{x}_i \text{ in cluster } k} d\left(\mathbf{x}_i, \mathbf{c}_k\right)$$

for $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K$ the final medoids produced by the iterations.