K-Means Clustering

k-Means clustering is one of the most famous and widely used clustering algorithms in practice. It works by using a distance measurement (most commonly the Euclidean distance) to iteratively assign data points in a hyperspace to a set of non-overlapping clusters.

In *K*-means, the anticipated number of clusters, *K*, is chosen at the onset. The clusters are initialized by arbitrarily selecting at random one of the data points as an initial cluster for each *K*. The algorithm now works by iteratively assigning each point in the space to the cluster centroid that it is nearest to using the distance measurement.

After all the points have been assigned to their closest cluster point, the cluster centroid is adjusted to find a new center among the points in the cluster. This process is repeated until the algorithm converges, that is, when the cluster centroids stabilize and points do not readily swap clusters after every reassignment. These steps are illustrated in Figure 25-2.

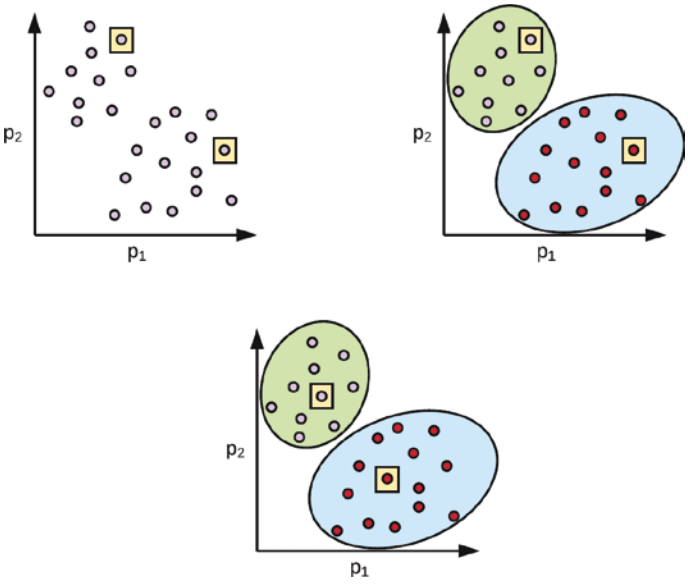

Figure 25-2. *An illustration of k-means clustering with k = 2. Top left: Randomly pick a point for each k. Top right: Iteratively assign each point to its closest cluster centroid. Bottom: Update the cluster centroids for each of the k clusters. Typically, we repeat the iterative assignment of all the points and update the cluster centroid until the algorithm resolves in a stable clustering.*

Considerations for Selecting *K*

There's really no way of telling the number of clusters in a dataset from the onset. The best way of selecting *k* is to try out different values of *K* to see what works best in creating distinct clusters.

Another strategy, which is widely employed in practice, is to compute the average distance of the points in the cluster to the cluster centroid for all clusters. This estimate is plotted on a graph as we progressively increase the value of *K*. We observe that as *K* increases, the distance of points from the centroid of its cluster gradually reduces, and the generated curve resembles the elbow of an arm. From practice, we choose the value of *K* just after the elbow as the best *K* value for that dataset. This method is called the elbow method for selecting *K* as is illustrated in Figure 25-3.

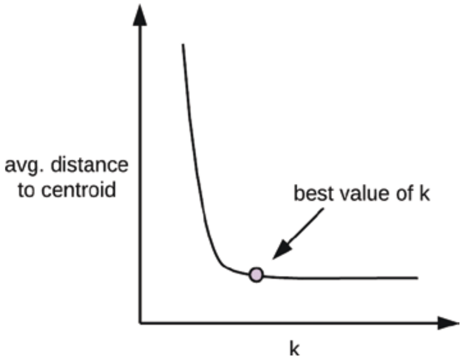

Figure 25-3. *The elbow method for choosing the best value of k*

Considerations for Assigning the Initial *K* Points

The points that determine the initial value of *K* are important in finding a good set of clusters. By selecting the point for *K* at random, two or more points may reside in the same cluster, and this will invariably lead to sub-par results. To mitigate this from occurring, we can employ more sophisticated approaches to selecting the value of *K*. A common strategy is to randomly select the first *K* point and then select the next point as the point that is farthest from the first chosen point. This strategy is repeated until all *K* points have been selected. Another approach is to run hierarchical clustering on a sub-sample of the dataset (this is because hierarchical clustering is a computationally expensive algorithm) and use the number of clusters after cutting off the dendrogram as the value of *K*.