CHAPTER 25

Clustering

Clustering is an unsupervised machine learning technique for grouping homogeneous data points into partitions called clusters. In the example dataset illustrated in Figure 25-1, suppose we have a set of *n* points and 2 features. A clustering algorithm can be applied to determine the number of distinct subclasses or groups among the data samples.

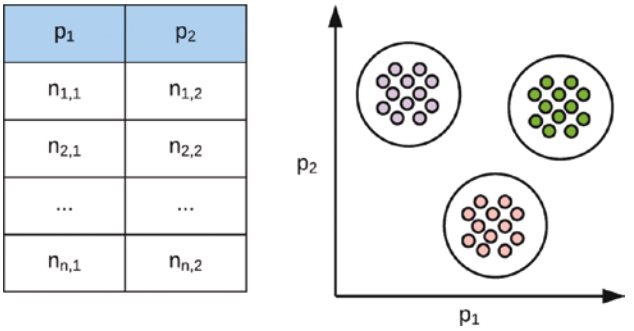

Figure 25-1. *An illustration of clustering in a 2-D space*

Clustering a 2-D dataset as seen in Figure 25-1 is relatively trivial. The real challenge arises when we have to perform clustering in higher-dimensional spaces. The question now is how do we ascertain or find out if a set of points are similar or if a set of points should be in the same group? In this section, we would cover two essential types of clustering algorithms known as k-means clustering and hierarchical clustering.

K-means clustering is used when the number of anticipated distinct classes or sub-groups is known in advance. In hierarchical clustering, the exact number of clusters is not known, and the algorithm is tasked to find the optimal number of heterogeneous sub-groups in the dataset.

E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*,
https://doi.org/10.1007/978-1-4842-4470-8_25

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.

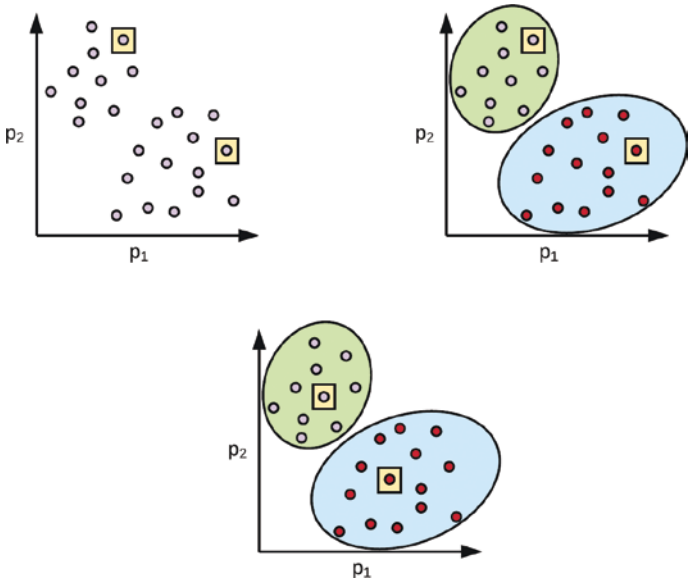

Figure 25-2. *An illustration of k-means clustering with k = 2. Top left: Randomly pick a point for each k. Top right: Iteratively assign each point to its closest cluster centroid. Bottom: Update the cluster centroids for each of the k clusters. Typically, we repeat the iterative assignment of all the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