Clustering

One of the most prominent domains among unsupervised learning models are those that use clustering techniques. These models attempt to build a simplified picture of a dataset by grouping together similar observations and distinguishing these from observations that fall into other groups. A common use case for this approach is in Market Segmentation, where one uses demographic and/or past purchase data to group consumers together. This process can be helpful for identifying the characteristics of a company's consumers, for the purpose of designing targeted advertising or promotions likely to appeal to those customers.

There are a large number of clustering algorithms used by data scientists working in various niches. Here we provide brief descriptions of a few of the most broadly applicable ones, starting with two classical combinatorial algorithms: *k*-Means and Hierarchical clustering. All methods of clustering apply some measure of similarity between observations to group them together, and these two take a geometric perspective by treating observations as points in space and measuring the distances between them. This approach has some drawbacks, which some more contemporary algorithms attempt to address via ranking systems and/or transformations of the data. We introduce two of these: Spectral Clustering and Affinity Propagation. An additional way of grouping data is to do so probabilistically, that is to assume that the data are drawn from a finite number of population distributions and to try to characterize these underlying distributions. Techniques of this form are referred to as Mixture Models, and are our final clustering topic.