Sean Cordrey

November 7, 2014

Stat 3654

Roberts

Weekly Paper 11

Chapter 8 in our R book covers some clustering methods models. Clustering models are useful in that they try to cluster data points that are similar. So we can think of data points being customers or locations and we can find which customers are similar to those of another cluster of customers or locations. Depending on what our goals are, this information can help us formulate solutions to reach our goals.

Clustering revolves around measuring similarity and dissimilarity to make our clusters. This can be thought of as distances. There are a numbers of ways to calculate distances but the most common would be Euclidean distances. Before we can calculate these though between our data points, we need to make sure we get everything in consistent units as well as scale our features. We scale our features to a mean of 0 and std. of 1 so that we don’t have to mess with a complex coordinate system.

The first of two ways we’ll look at to implement clustering is Hierarchical clustering. This is easily applied to our scaled features with the hclust() function. Once we have our clusters with this method, we can visualize them with PCA or a dendrogram which is covered in lecture. The other implementation is K means. Once again we have an easy function kmeans() that we can apply to our scaled features.

This chapter also discusses association rules which formulate rules based on subjects of interest occurring together frequently. A way to do this is look for the subjects of interest that occur more often than in a minimum set fraction of the transactions.