Week 3 Topics+

1. Clustering

# What is Clustering?

Clustering involves the grouping of similar objects into a set known as cluster. Objects in one cluster are likely to be different when compared to objects grouped under another cluster. Clustering is one of the main tasks in exploratory data mining and is also a technique used in statistical data analysis. While clustering is not one specific algorithm, it is a general task that can be solved by means of several algorithms. Some of the popular clustering methods that are used include hierarchical, partitioning, density-based and model-based.

Clustering is also known as clustering analysis

Clustering is the act of creating various clusters that have all objects under the data set. Further, clustering can be distinguished into hard and soft clustering. Under hard clustering, an object either belongs to a cluster or it does not. However, with soft clustering (fuzzy clustering) an object can belong to many clusters. The ultimate aim of clustering is to intrinsically group unlabeled data. It finds applications in market research, pattern recognition, data mining and analysis, data compression, image recognition and more.

The concept of a cluster cannot be easily defined, and this is largely why several algorithms are available for clustering. These algorithms differ in their properties, and therefore, researchers are known to apply different cluster models based on the data set in question and also what it is intended to be used for. For example, hierarchical clustering is based on distance connectivity, while distribution models are based on statistical distributions.

Clustering is considered an unsupervised learning method since we don’t have the ground truth to compare the output of the clustering algorithm to the true labels to evaluate its performance. We only want to try to investigate the structure of the data by grouping the data points into distinct subgroups.

In this course, we will cover only K-means which is considered as one of the most used clustering algorithms due to its simplicity.

# K-means Algorithm

K-means algorithm is an iterative algorithm that tries to partition the dataset into K-predefined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

# k-means algorithm:

Algorithm goal is to *find the k cluster centers and assign the observations (records) to the nearest cluster center, such that the squared distances from the cluster are minimized*. That is this algorithm aims at minimizing an objective function, in this case a squared error function.

The objective function definition:

Given a set of observations (x1, x2, …, x*n*), where each observation is a *d*-dimensional real vector (i.e. a record with many attributes), *k*-means clustering aims to partition the *n* observations into *k* (≤ *n*) sets **S** = {*S*1, *S*2, …, *Sk*} so as to minimize the Within-Cluster Sum of Squares (WCSS) (sum of distance functions of each point in the cluster to the K center). In other words, its objective is to find:

Minimized!

Where , which mean the length of a vector, is a chosen distance measure between a point x and cluster center of the cluster *i*, and is an indicator of the distance of the *n* data points (observations) from their respective cluster centers.

Steps

1. Specify number of clusters K.
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e. assignment of data points to clusters isn’t changing. (this is a loop and stops when there is no change in part c of the loop)
   1. Compute the sum of the squared distance between data points and all centroids.
   2. Assign each data point to the closest cluster (centroid).
   3. Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

# Applications

k-means algorithm is very popular and used in a variety of applications such as market segmentation, document clustering, image segmentation and image compression, etc. The goal usually when we undergo a cluster analysis is either:

1. Get a meaningful intuition of the structure of the data we’re dealing with.
2. Cluster-then-predict where different models will be built for different subgroups if we believe there is a wide variation in the behaviors of different subgroups. An example of that is clustering patients into different subgroups and build a model for each subgroup to predict the probability of the risk of having heart attack.
3. Clustering with K-means using R

The following code is from example listed in the page 430 of your textbook.

##\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

#load and normalize data

utilities.df<-read.csv("Utilities.csv")

utilities.df[,1]

row.names(utilities.df)<-utilities.df[,1]

utilities.df[,1]

utilities.df<-utilities.df[,-1]

##\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

#Normalized distance

utilities.df.norm<-sapply(utilities.df, scale)

row.names(utilities.df.norm)<-row.names(utilities.df)

##\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

#Appy the K-means

utilities.km<-kmeans(utilities.df.norm, 6)

##\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

utilities.km$cluster

utilities.km$centers