**MACHINE LEARNING FOR DATA SCIENCE**

**Q1.** **What is the difference between supervised and unsupervised learning? Give one example (not a technique or algorithm) to demonstrate.**

**Solution:**

Supervised Learning:

Supervised Learning is training the machine learning task for every input with corresponding target (label class) and it will be able to provide target (label) for any new input, after sufficient training. In simpler terms, we can say that if there is training data (input information) and if we already know what will be the outcome variable, then the new input will get the prediction based on the training model, this is called as supervised learning. Regression and Classification are widely used for supervised learning.

Example: Showing kids different types of animals, and telling them that a type in it is dog. After giving few more tips on that type, if kids identify a new species of dog (which is never showed to them so far) correctly as Dog itself not as any other species, then the kids undergone supervised learning.

Unsupervised Learning:

Unsupervised learning is training the machine learning task only with the set of inputs, no target variable (no Y label) is known and it will be able to find the structure or relationships between different inputs. In simpler terms, no training data available, directly testing data is given with no knowledge of what the outcome variable, So, it is left with the learning algorithm to figure out patterns in the data provided and make the decision. Clustering and Association are used for unsupervised learning.

Example: Going Bag-packing to a new country, without knowing about their food, culture, language etc., But from day 1, we will start learning to eat new cuisines, finding places to stay. There is no one to guide us, we will find way on our own. Based on some criteria, we will start churning out the information into groups that make sense. This can be an example of unsupervised learning.

**Q2. Explain how/why k-means is (almost) guaranteed to converge**.

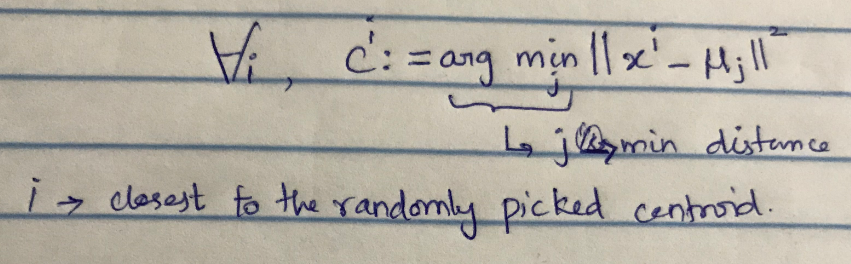
**Solution:**

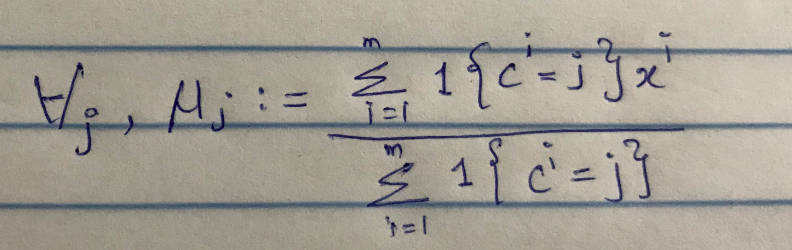
k-means clustering algorithm is an unsupervised learning algorithm, where a given input datasets should be clustered with some relationships among them, with no labels provided.

k-means clustering algorithm:

Step 1: Initialize cluster centroids randomly (for an n dimensional space). µ1 andµ2 for 2-dimensional space.

Step 2: Repeat until convergence: {

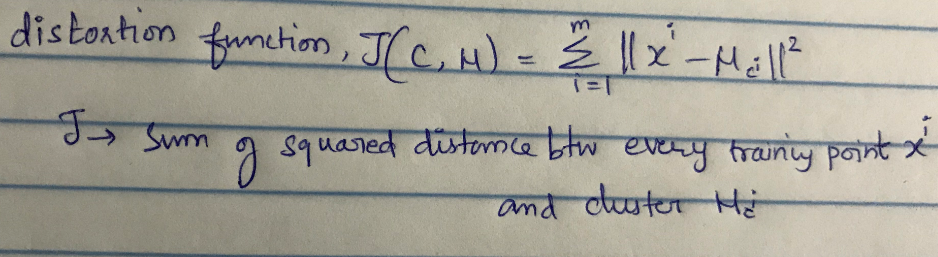
For every i, 

For every j, 

This iteration of assigning each training example x(i) to the closest cluster centroid µj, and moving each cluster centroid µj to the mean of the points assigned to it keeps repeating.

**k-means is (almost) guaranteed to converge**:

k-means is exactly coordinate descent on J (distortion function) and it is recommended to minimize J.



The step 2 loop of k-means repeatedly minimizes J with respect to c while holding µ fixed, and then minimizes J with respect to µ while holding c fixed. Thus, J must monotonically decrease, and the value of J must converge which means that no points in the cluster gets reassigned after its reaches the convergence. Hence, it is guaranteed that k-means is(almost) guaranteed to converge.

**Q3. Use this dataset and two different clustering approaches (agglomerative and divisive) to group the airlines with similar safety records. Do these two approaches lead to same/similar results? Provide your brief comments.**

**Solution:** Agglomerative hierarchical Clustering

Agglomerative clustering is called as the bottom-up approach. In a bottom-up approach, all the data points are treated as individual clusters to start with and then gradually merged into bigger clusters.

Steps to be followed:

1. Loading the airline\_safety dataset into the R studio.

2. Remove the categorical variable “airline” as these techniques can be only applied to the numerical variables.



3. Load the “cluster” package and use the agnes () function (agglomerative nesting algorithm) to perform the agglomerative clustering.

4. There are three method which can be used for the linkage of the clusters.

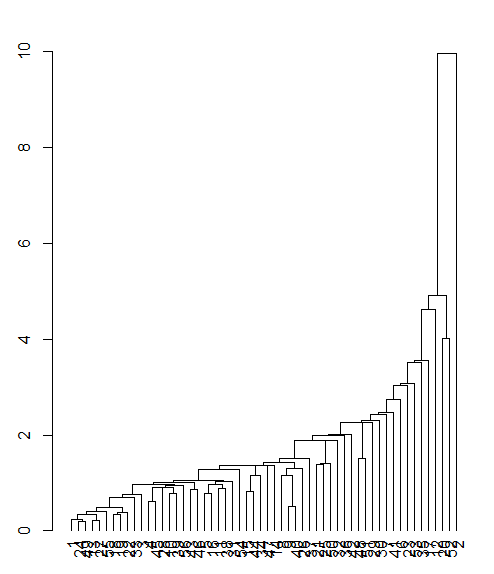
a. single method – single linkage clustering

b. complete method – complete linkage clustering

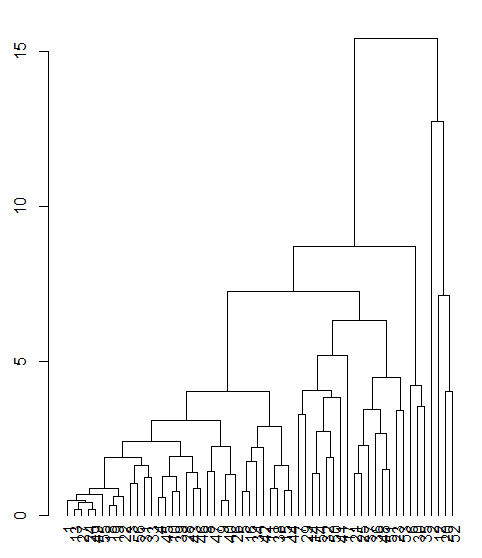
c. average method – average linkage clustering

5. And then plot the dendrograms to visualize the clusters.

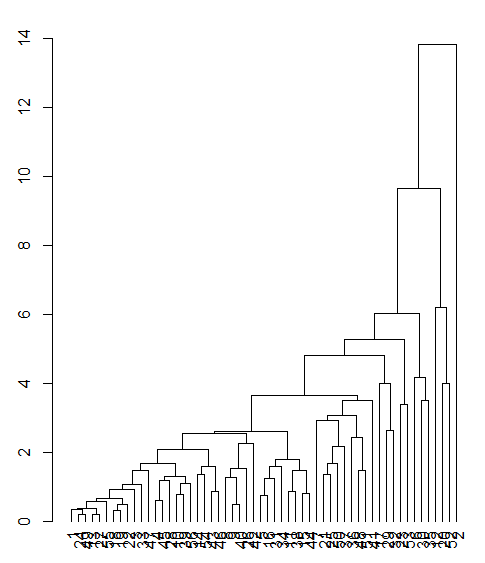
6. Dendrogram for single linkage cluster



7. Dendrogram for complete linkage cluster



8. Dendrogram for average linkage cluster



Divisive Clustering:

This is also hierarchical clustering but as opposed by agglomerative, this follows a top-down approach. Make all data points as single cluster and it is split using the clustering algorithm repeatedly until each point gets its own singleton cluster.

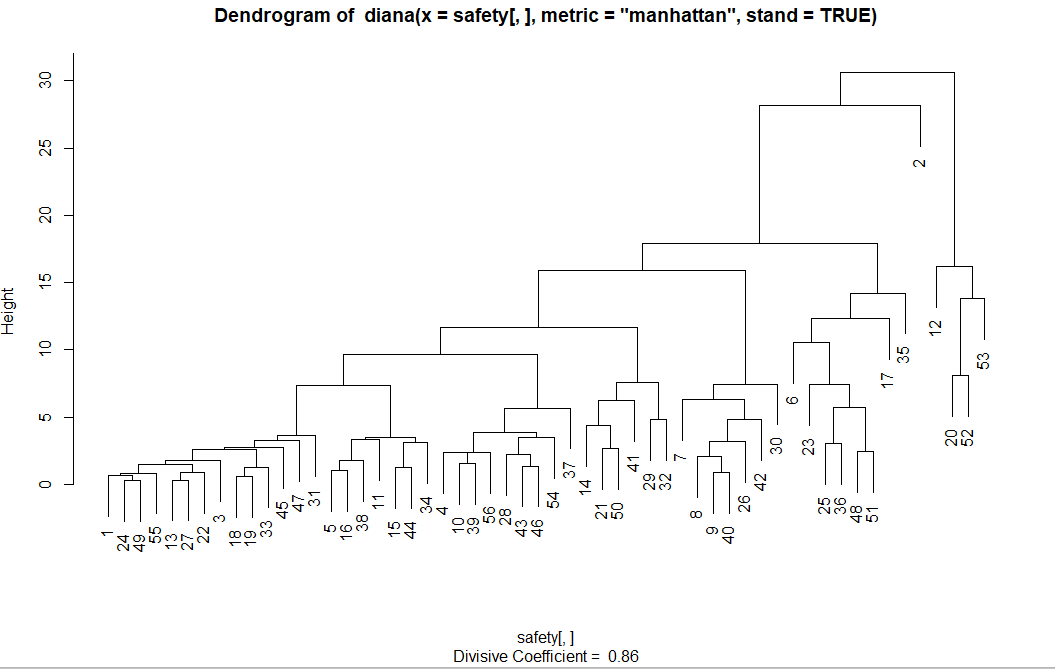
Steps to be followed:

1. Load the dataset.

2. Remove the categorical variable airline

3. Use diana () function (divisive analysis clustering) to perform divisive clustering and then visualize the cluster in the dendrogram.

4. Dendrogram for Divisive Clustering for our dataset looks like this –



From the dendrogram of both the approaches, the result looks slightly similar with divisive clustering and agglomerative complete linkage. But if we compare the divisive clustering with the agglomerative average and single method, they look different.

May be, if we normalize the data in both approaches and set a threshold point to draw the tree cut, then it depends on how many clusters we get as a result in both, might tell us if their outputs are similar or not. Like if the number of clusters are same in both the approaches, then we can conclude that they gave similar results.