**Clustering Evaluation strategies**

Clustering is an unsupervised machine learning algorithm. It helps in clustering data points to groups. Validating the clustering algorithm is bit tricky compared to supervised machine learning algorithm as clustering process does not contain ground truth labels. If one want to do clustering with ground truth labels being present, validation methods and metrics of supervised machine learning algorithms can be used?.

**How Clustering can be evaluated?**

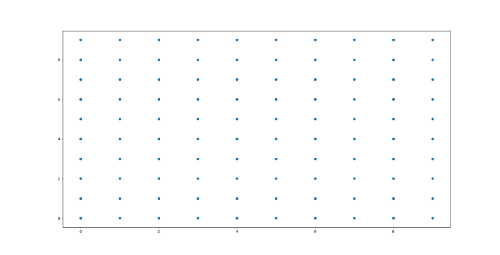
Three important factors by which clustering can be evaluated are

*(a) Clustering tendency (b) Number of clusters,****k****(c) Clustering quality*

**Clustering tendency**

Before evaluating the clustering performance, making sure that data set we are working has clustering tendency and does not contain uniformly distributed points is very important. If the data does not contain clustering tendency, then clusters identified by any state of the art clustering algorithms may be irrelevant. Non-uniform distribution of points in data set becomes important in clustering.

To solve this, Hopkins test, a statistical test for spatial randomness of a variable, can be used to measure the probability of data points generated by uniform data distribution.



Plot for data from Uniform distribution

***Null Hypothesis (Ho) :****Data points are generated by uniform distribution (implying no meaningful clusters)****Alternate Hypothesis (Ha):****Data points are generated by random data points (presence of clusters)*  
If H>0.5, null hypothesis can be rejected and it is very much likely that data contains clusters. If H is more close to 0, then data set doesn’t have clustering tendency.

**Number of Optimal Clusters, *k***

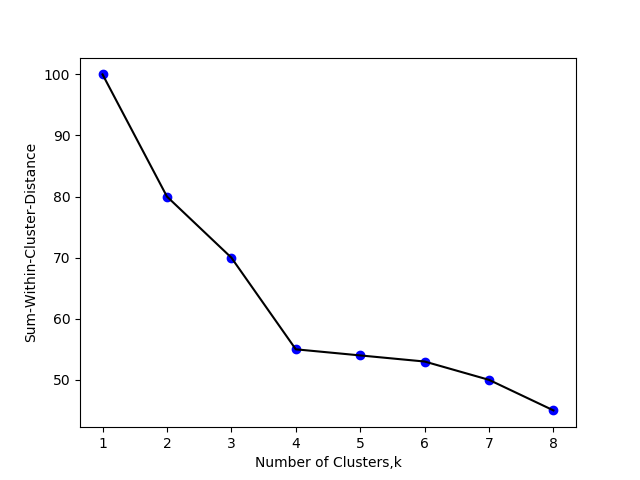
Some of the clustering algorithms like K-means, require number of clusters, *k*, as clustering parameter. Getting the optimal number of clusters is very significant in the analysis. If *k* is too high, each point will broadly start representing a cluster and if *k* is too low, then data points are incorrectly clustered. Finding the optimal number of clusters leads to granularity in clustering.  
  
There is no definitive answer for finding right number of cluster as it depends upon (a) Distribution shape (b) scale in the data set (c) clustering resolution required by user. Although finding number of clusters is a very subjective problem. There are two major approaches to find optimal number of clusters:  
(1) Domain knowledge  
(2) Data driven approach  
  
**Domain knowledge**— Domain knowledge might give some prior knowledge on finding number of clusters. For example, in case of clustering iris data set, if we have the prior knowledge of species (*sertosa, virginica, versicolor*) , then k = 3. Domain knowledge driven *k value*gives more relevant insights.  
  
**Data driven approach**— If the domain knowledge is not available, mathematical methods help in finding out right number of clusters.

*Empirical Method:-*

A simple empirical method of finding number of clusters is Square root of N/2 where N is total number of data points, so that each cluster contains square root of 2 \* N  
  
*Elbow method:-*

Within-cluster variance is a measure of compactness of the cluster. Lower the value of within cluster variance, higher the compactness of cluster formed.

Sum of within-cluster variance, *W*, is calculated for clustering analyses done with different values of k. *W* is a cumulative measure how good the points are clustered in the analysis. Plotting the *k* values and their corresponding sum of within-cluster variance helps in finding the number of clusters.



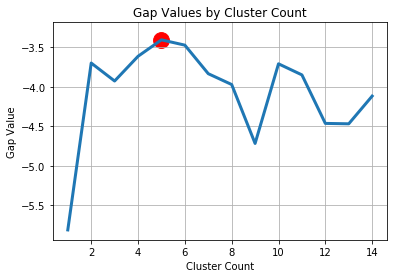
Plot of Sum of within cluster distance vs Number of clusters in Elbow method

Plot shows that number of optimal clusters = 4.  
Initially, Error measure (within-cluster variance) decreases with increase in cluster number. After a particular point, k=4 , Error measure starts flattening. Cluster number corresponding to that particular point, k=4, should be considered as optimal number of clusters.  
  
*Statistical approach:-*

Gap statistic is a powerful statistical method to find the optimal number of clusters, *k*.

Similar to Elbow method, sum of within-cluster (intra-cluster) variance is calculated for different values of k.  
Then Random data points from reference null distribution are generated and Sum of within-cluster variance is calculated for the clustering done for different values of k.

In Simpler words, Sum-of-within-Cluster variance of original data set for different values of k to Sum-of-within-cluster variance of reference data set (null reference data set of uniform distribution) of corresponding values of k is compared to find the ideal *k* value where ‘deviation’ or ‘Gap’ between two is highest. As Gap statistic quantifies this deviation, More the Gap statistic means more the deviation.



Gap statistic Value-Cluster number (k)

Cluster number with maximum Gap statistic value corresponds to optimal number of cluster.

Clustering quality

Once clustering is done, how well the clustering has performed can be quantified by a number of metrics. Ideal clustering is characterised by minimal intra cluster distance and maximal inter cluster distance.

There are majorly two types of measures to assess the clustering performance.

(i)*Extrinsic Measures* which require ground truth labels. Examples are Adjusted Rand index, Fowlkes-Mallows scores, Mutual information based scores, Homogeneity, Completeness and V-measure.

(ii) *Intrinsic Measures*that does not require ground truth labels. Some of the clustering performance measures are Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index etc.