Cluster Validation

Done By Venkatesh E
Priyadarshan

For cluster analysis, the question is how to evaluate the "goodness" of the resulting clusters?

Then why do we want to evaluate them?

—>To avoid finding patterns in noise.

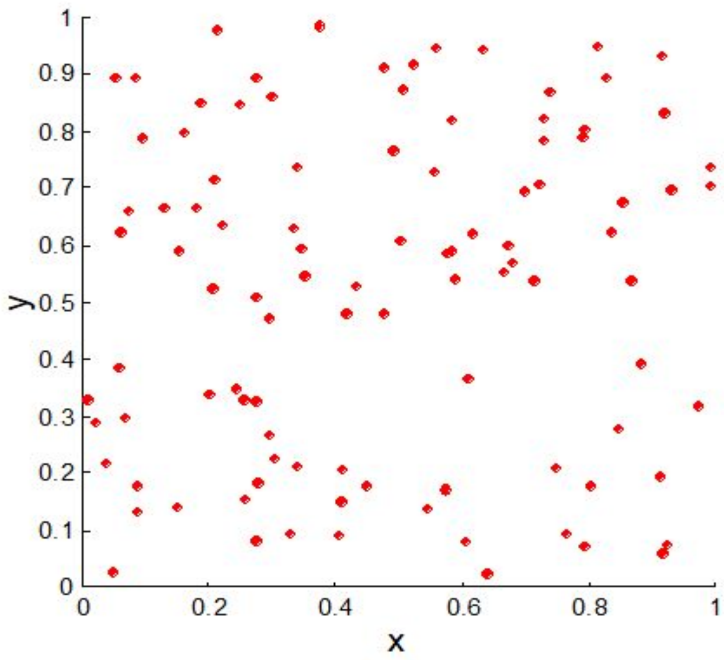
—>To compare clustering algorithms.

—>To compare two sets of clusters.

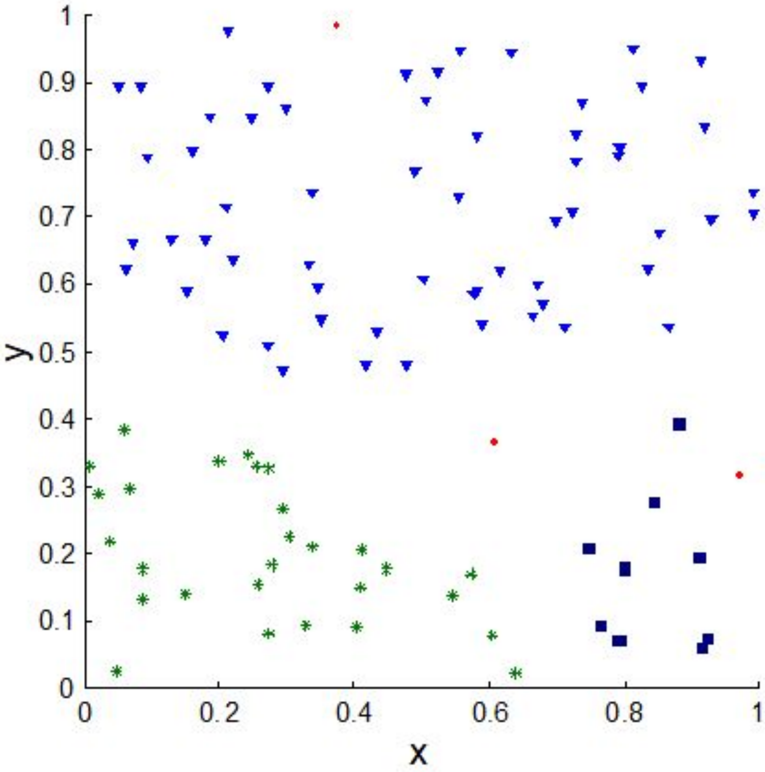
—>To compare two clusters.

Some clusters with random datas

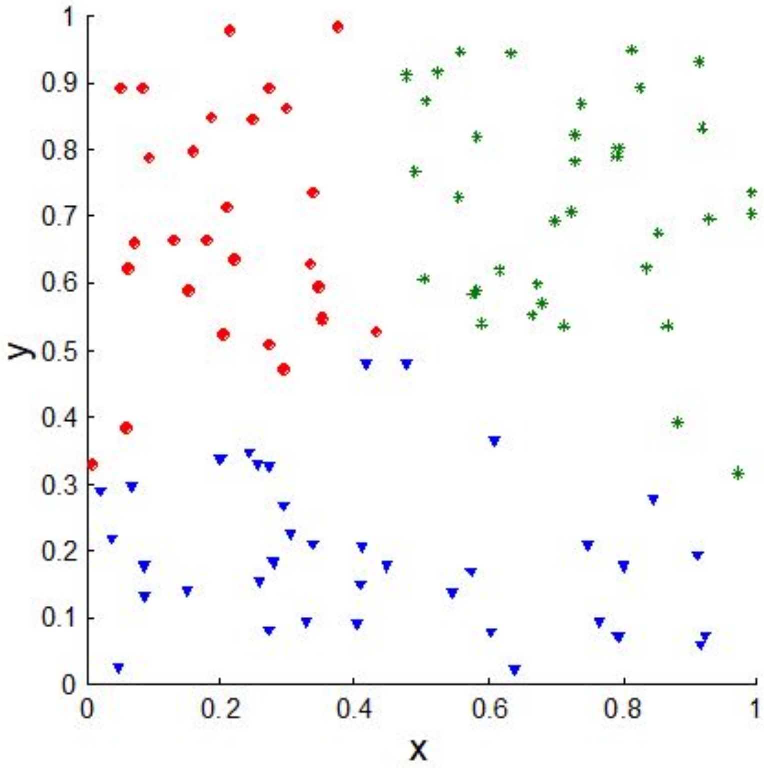
Random Points



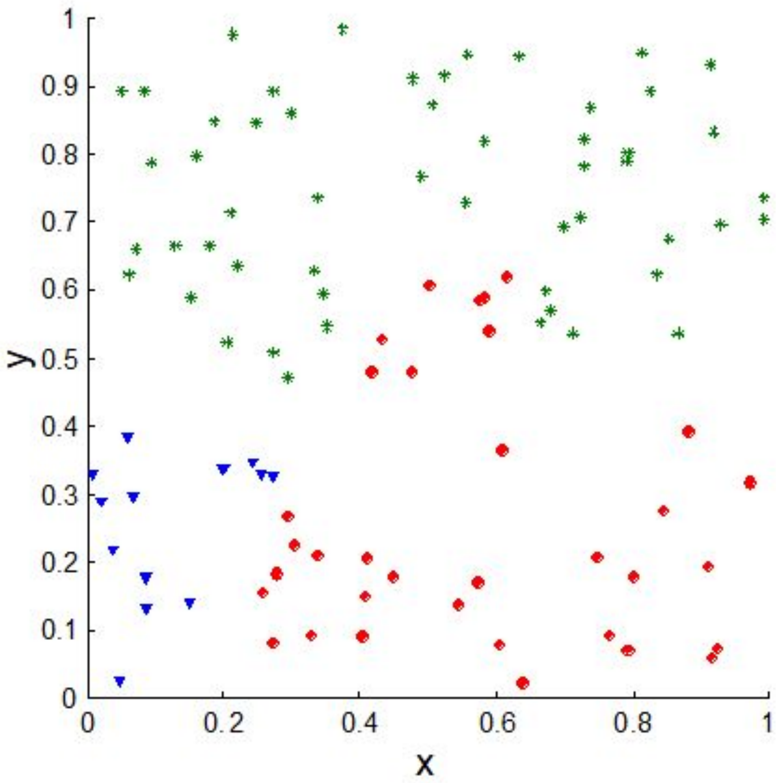
DBSCAN



K-Means



Complete Link



Different aspects of cluster validation

Aspect - 1

Determining the clustering tendency of a set of data, i.e., distinguishing whether non-random structure actually exists in the data.

Clustering tendency

Clustering tendency assessment **determines whether a given dataset contains meaningful clusters**

Aspect - 2

Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels.

Aspect - 3

Evaluating how well the results of a cluster analysis fit the data *without* reference to external information.We should use only the data given.

Aspect - 3

Evaluating how well the results of a cluster analysis fit the data *without* reference to external information.We should use only the data given.

Aspect - 4

Comparing the results of two different sets of cluster analyses to determine which is better.

Aspect - 5

Determining the ‘correct’ number of clusters.

For 2, 3, and 4, we can further distinguish whether we want to evaluate the entire clustering or just individual clusters.

Framework for cluster validation

Need a framework to interpret any measures

For example,

If our measure of evaluation has the value, 10,

is that good, fair, or poor?

Statistics provide a framework for cluster validation

– The more "atypical" a clustering result is, the more likely it represents valid structure in the data

– Can compare the values of an index that result from random data or clusterings to those of a clustering result.

- If the value of the index is unlikely, then the cluster results are valid

– These approaches are more complicated and harder to understand.

For comparing clusters framework is less necessary

However, there is the question of whether the difference between two index values is significant

Measure of cluster validity

- **Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.**

  – External Index: Used to measure the extent to which cluster labels match externally supplied class labels.

  – Internal Index: Used to measure the goodness of a clustering structure *without* respect to external information.

  – Relative Index: Used to compare two different clusterings or clusters.

- Often an external or internal index is used for this function, e.g., SSE or entropy

- **Sometimes these are referred to as criteria instead of indices**

  – However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

External Validation

Algorithm 21.4: Algorithm for matching partitions and clusters

MatchPartitionCluster **(**P**,**C**,**$match$**)**:

foreach $p \in P$ **do**

  $match(p) \leftarrow \emptyset$

  foreach $c \in C$ **do**

    $overlap(p, c) \leftarrow \frac{|p \cap c|}{|p|}$

while $overlap \neq \emptyset$ **do**

  $(p_{max}, c_{max}) \leftarrow GetMaxOverlap(overlap)$

  $match(p_{max}) \leftarrow c_{max}$

  $overlap \leftarrow overlap - \{overlap(p_{max}, *), overlap(*, c_{max})\}$

Correlation Measures

Is a statistical technique which determines how one variables

moves/changes in relation with the other variable.

Huberts tau statistics

$$\Gamma = \frac{1}{m}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} X_P(i,j)X_C(i,j)$$

Normalized tau statistics

$$\hat{\Gamma} = \frac{\frac{1}{m}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(X_P(i,j) - \mu_P)(X_C(i,j)\mu_C)}{\sigma_P \sigma_C}$$

where μ_P and μ_C are the means and σ_P and σ_C are the variances of the matrices X_C and X_P.

Measuring cluster validity via correlations

Two matrices

Proximity Matrix

"Incidence" Matrix

- One row and one column for each data point
- An entry is 1 if the associated pair of points belong to the same cluster
- An entry is 0 if the associated pair of points belongs to different clusters

Compute the correlation b/w 2 matrices

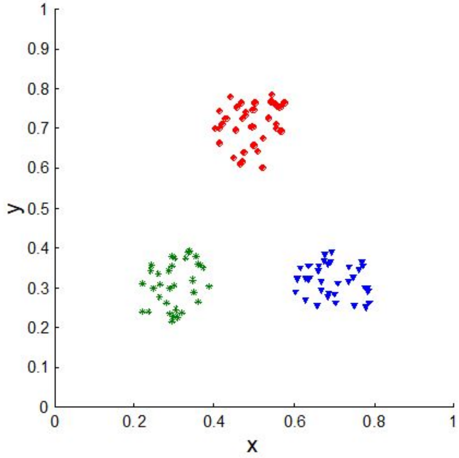
Since the matrices are symmetric, only the correlation between

$n(n-1) / 2$ entries needs to be calculated.

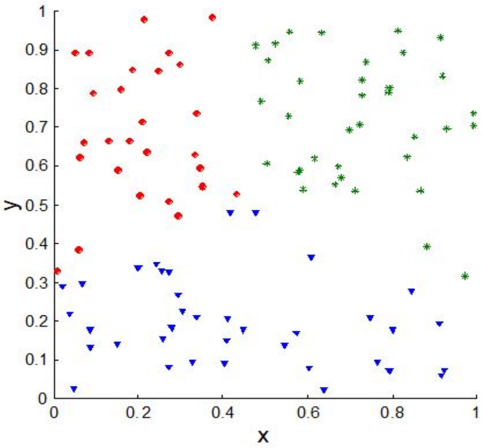
Higher correlation means points are close to each other in same cluster.

Not a good measure for some density or contiguity based clusters.

Correlation of incidence and proximity matrices for k-means clustering of the following 2 data sets



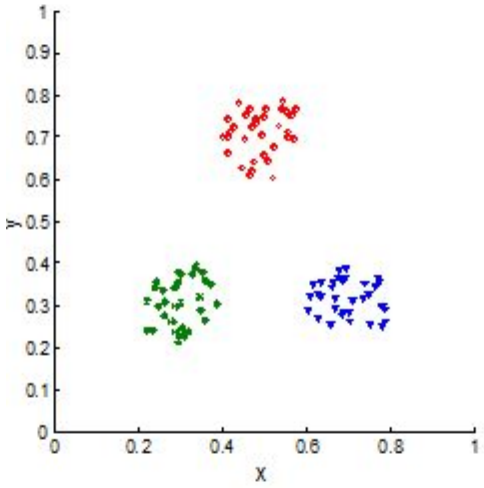
Corr=-0.9235

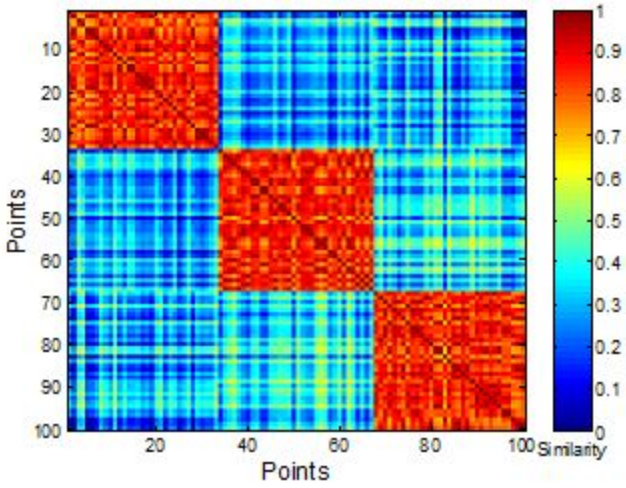


Corr=-0.5810

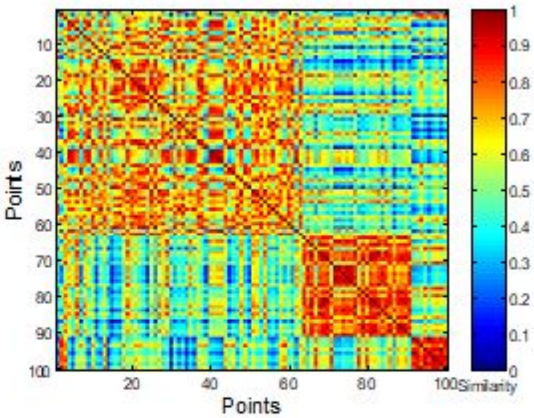
Using similarity matrix for cluster validation

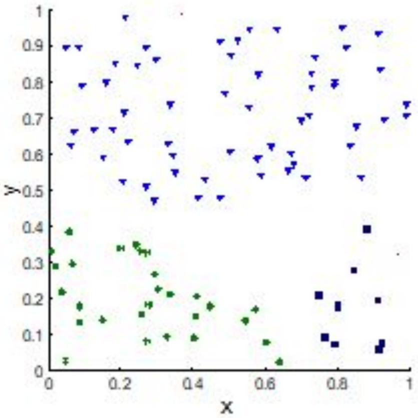
Order the similarity matrix with respect to cluster labels and inspect visually.





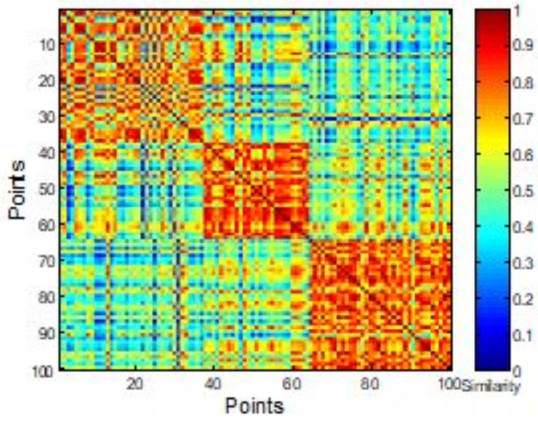
Clusters in random data are not so crisp

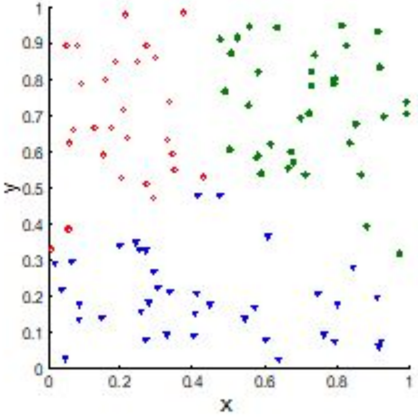




DBSCAN

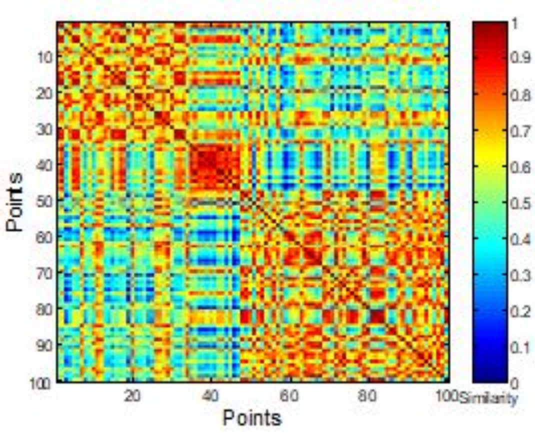
Clusters in random data are not so crisp

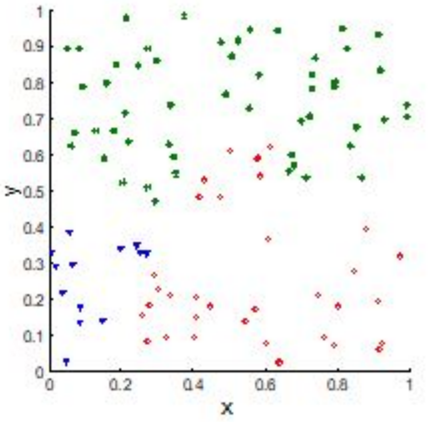




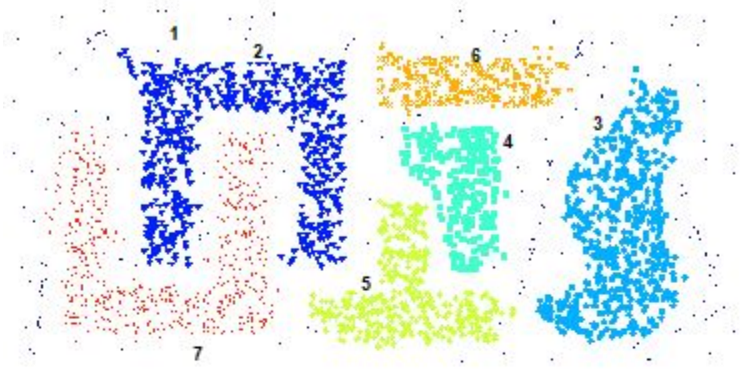
K-MEANS

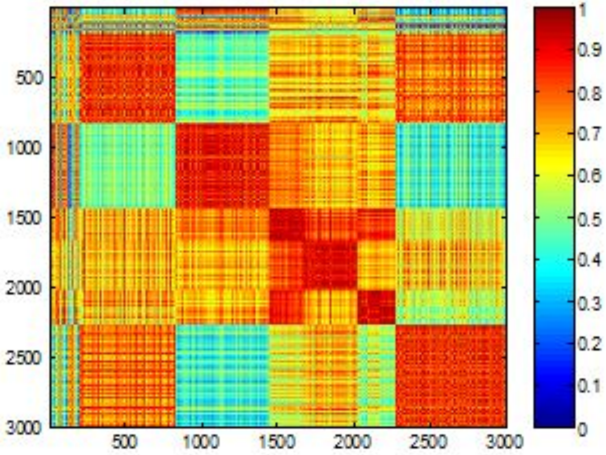
Clusters in random data are not so crisp





COMPLETE LINK


1
2
6
4
3
5
7

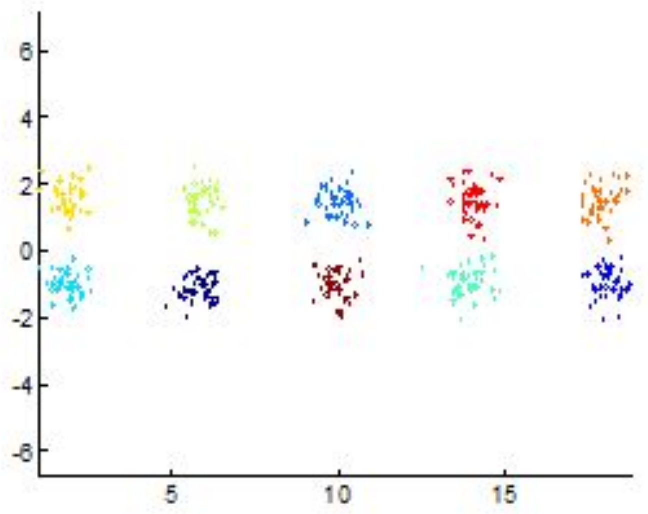

500
1000
1500
2000
2500
3000
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

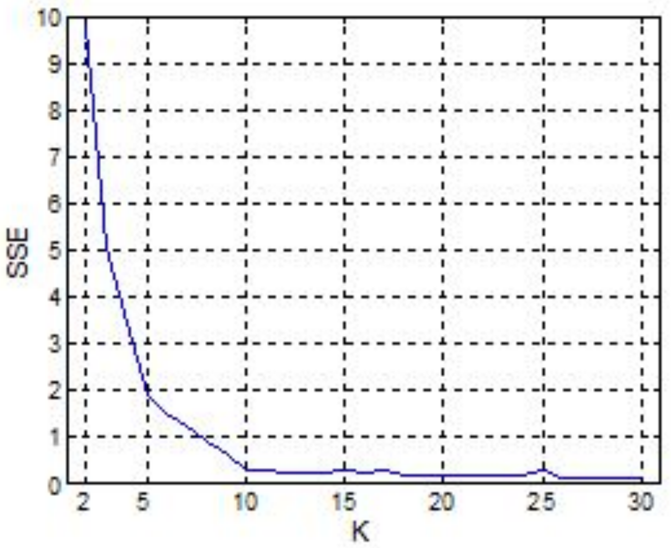
DBSCAN

Internal measures : SSE

- Clusters in more complicated figures aren’t well separated.
- Internal Index: Used to measure the goodness of a clustering structure without respect to external information.
- SSE is good for comparing two clusterings or two clusters (average SSE).
- Can also be used to estimate the number of clusters

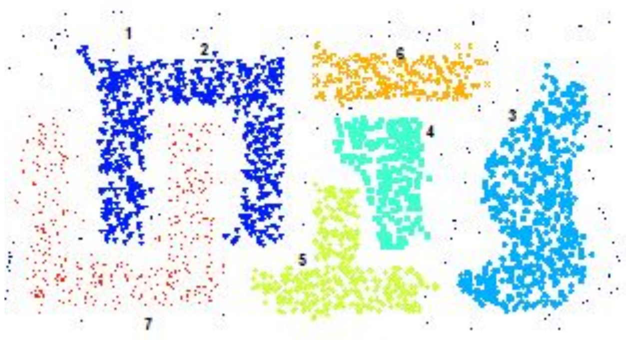
Number of clusters can be determined

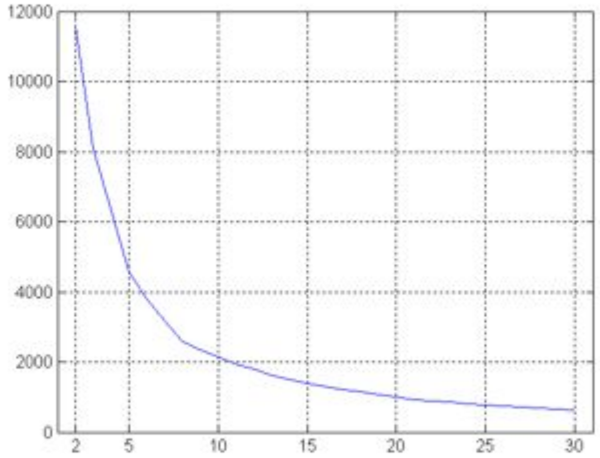




Internal measures : SSE

SSE curve for a more complicated data set.





SSE Cluster found using K-Means

Internal measure : cohesion & separation

Cluster Cohesion : Measures how closely related are objects in a cluster

$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

Cluster Separation : Measure how distinct or well-separated a cluster is from other clusters

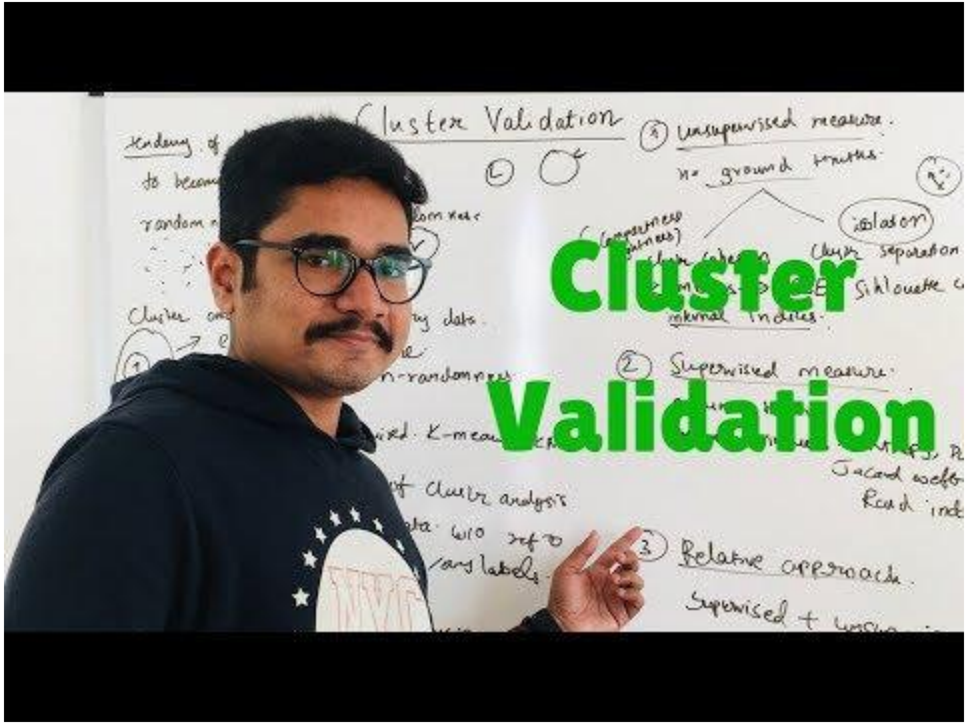
$$BSS = \sum_i |C_i| (m - m_i)^2$$

Comment made by Jain & Dube in book (Algorithms of clustering data)

“The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage.”

For more details :



Thank you