Unsupervised Clustering Evaluation Methods

Reference:

Javier B´ejar, AMLT - 2016/2017

Cluster Validation, Kent state university

Why do we need evaluation methods?

We can use such evaluation methods to:

- Avoid finding patterns in noise
- Compare different clustering algorithms
- Compare different sets of parameters used in the same algorithm
- Compare sets of clusters

Supervised Classification vs. Unsupervised Clustering

- When dealing with supervised classification problems we have a variety of measures to evaluate our models:
  - Accuracy
  - Precision
  - Recall

How can we evaluate the “goodness” of a clustering algorithm result?

Are there any real patterns in the data?

- Test the hypothesis of the existence of clusters in the data against a uniformly homogeneously distributed dataset
  - Hopkins Statistics:
    - Sample n points (p_i) from the dataset (D) uniformly and compute the distance from each point to its nearest neighbor in D ($d(p_i)$)
    - Generate n points (q_i) uniformly distributed in the space of the dataset D, and compute the distance from each generated point to its nearest neighbor in the dataset ($d(q_i)$)
    - Compute the Hopkins quotient:

$$H = \frac{\sum_{i=1}^{n} d(p_i)}{\sum_{i=1}^{n} d(p_i) + \sum_{i=1}^{n} d(q_i)}$$

    - If points are uniformly distributed then the value of H should be around 0.5 (because the densities in the generated point's areas should be similar to those of the sampled point's areas)

How “good” are the clusters we found?

- There are 3 types of numerical measures used to judge the validity of the resulting clusters (called “criteria” or “indices”):
  - External Index: Used to measure the extent to which cluster match externally supplied labels (REQUIRES LABELS! – wont be discussed)
  - Internal Index: Used to measure the quality of the resulting partitioning without any external data, can also be used to estimate the “real” number of clusters
  - Relative Index: Compare two different clustering structures (using the same distance measure) by comparing an internal index value computed from the 2 structures

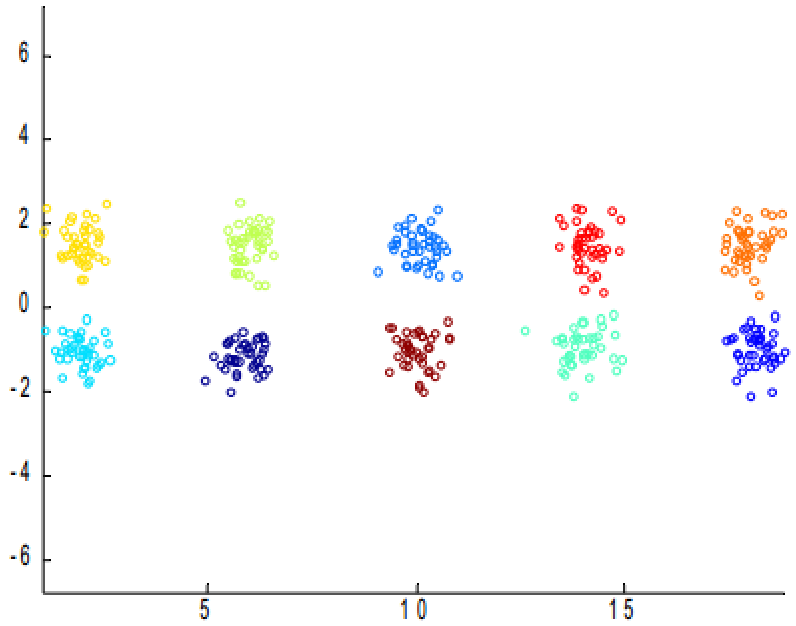
Internal Indices

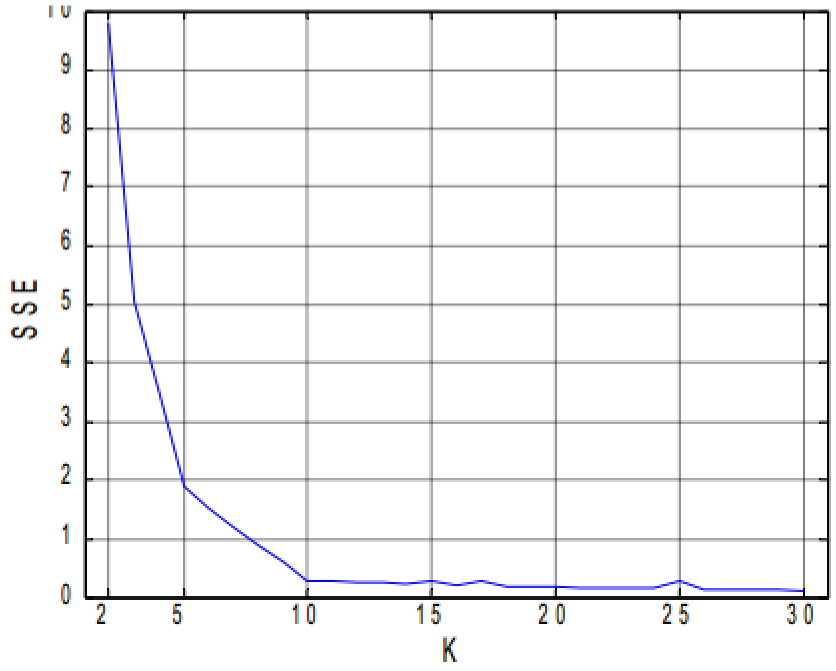
- Aims to measure 2 things:
  - “compactness” of the clusters
  - “separation” of the clusters
- Based on the model used
- Based on statistical properties of the attributes of the model:
  - Values distribution
  - Distances distribution

SSE Internal Index

We can use the Sum of Squared Errors to estimate the number of clusters

- For example by using distances from each point to its cluster centroid





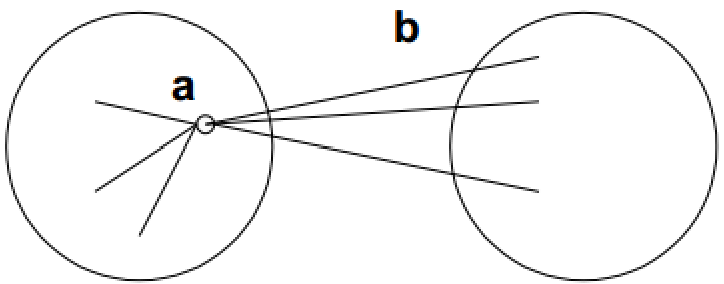
Correlation Internal Index

- Define 2 matrices:
  - Proximity matrix (n by n matrix, with the distance from point i to point j listed in cell I, j)
  - Incidence matrix (n by n matrix with cell i, j marked “1” if points i and j belong to the same cluster or “0” otherwise)
- Compute the correlation between the 2 matrices, with a high correlation indicating that points belonging to the same cluster are close to each other

Silhouette Internal Index

- For an individual point, I
  - Calculate a_i = average distance of i to the points in its cluster
  - Calculate b_i = min (average distance of i to points in another cluster)

$$S = \frac{1}{N}\sum_{i=0}^{N}\frac{b_i - a_i}{max(a_i, b_i)}$$



  - Results are between -1 and 1, with results closer to 1 indicating a better clustering pattern