Data Mining I

Summer semester 2019

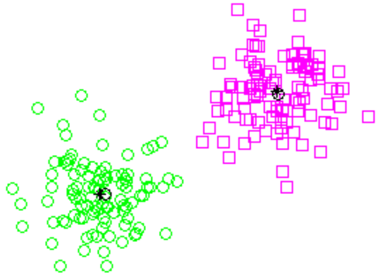
Lecture 12.a: Clustering – 4: Evaluation

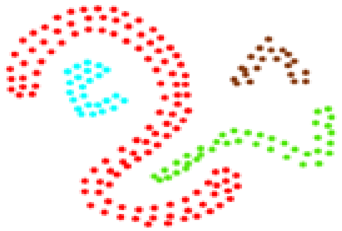
Lectures: Prof. Dr. Eirini Ntoutsi

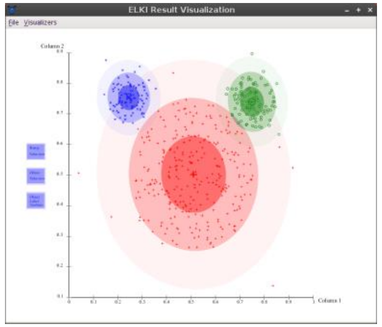
TAs: Tai Le Quy, Vasileios Iosifidis, Maximilian Idahl, Shaheer Asghar

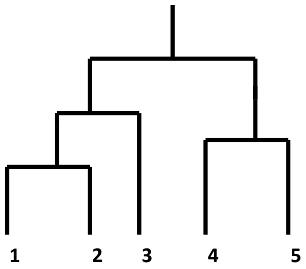
Clustering topics covered in DM1

1. Partitioning-based clustering
   - kMeans, kMedoids
2. Density-based clustering
   - DBSCAN
3. Grid-based clustering
4. Hierarchical clustering
   1. Diana, Agnes, BIRCH, ROCK, CHAMELEON
5. Clustering evaluation









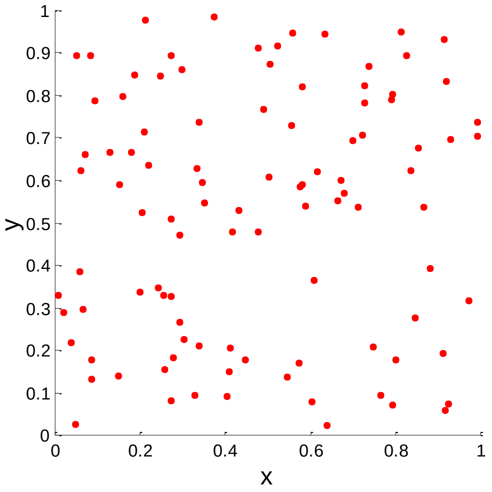
Cluster Validity

- In supervised learning, there is a variety of measures to evaluate how good a classifier is
  - accuracy, precision, recall, …

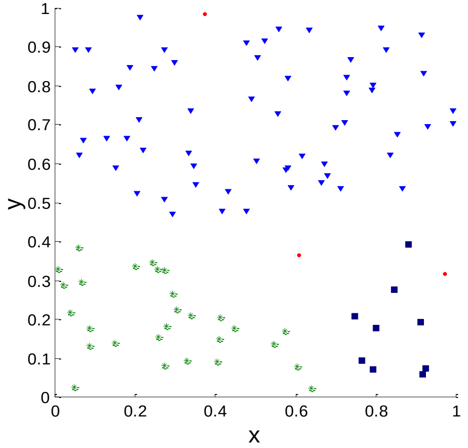
- For cluster analysis, the analogous question is how to evaluate the “goodness” of the resulting clusters?
  - That is a tricky question as “clusters are in the eye of the beholder”!

Clusters found in random data

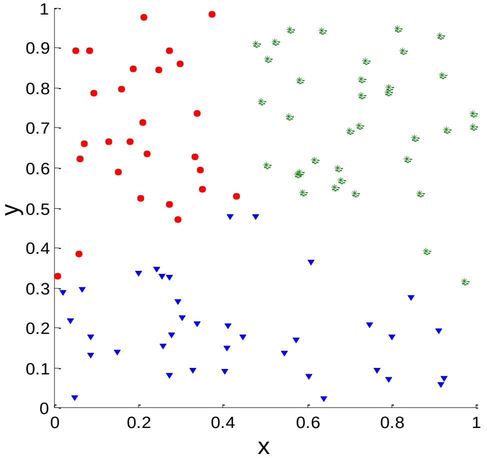
Random Points



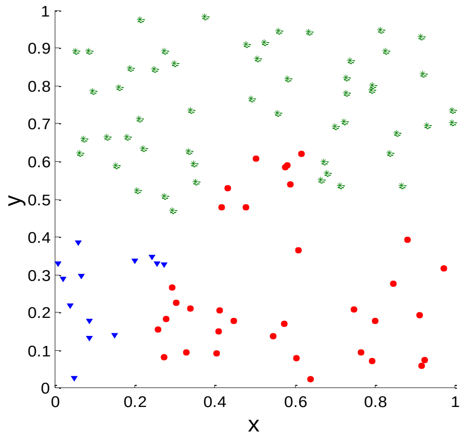
DBSCAN



K-means



Complete Link



Different Aspects of Cluster Validation

- Cluster validation has different goals:
  - Determining the clustering tendency of a set of data, i.e., distinguishing whether non-random structure actually exists in the data.
  - Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels.
  - Evaluating how well the results of a cluster analysis fit the data *without* reference to external information.

    - Use only the data
  - Comparing the results of two different sets of cluster analyses to determine which is better.
  - Determining the ‘correct’ number of clusters.

- Another aspect: Do we want to evaluate the entire clustering or just individual clusters?

Measures of Cluster Validity

- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types:
  - Internal Indices/Criteria: Used to measure the goodness of a clustering structure *without* any external information.
    - Sum of Squared Error (SSE)
  - External Indices/Criteria: Used to measure the extent to which cluster labels match *externally supplied class labels*.
    - Entropy
  - Relative Indices/Criteria: Used to compare two different clusterings or clusters.
    - Often an external or internal index is used for this function, e.g., SSE or entropy

Internal measures of cluster validity

- Idea: Check cluster characteristics, <u>do not</u> rely on external information
- Examples: cohesion and separation
- Cluster Cohesion: Measures how closely related are objects in a cluster
  - Cohesion is measured by the within cluster sum of squares (SSE)

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

(already discussed in the context of *k*-Means

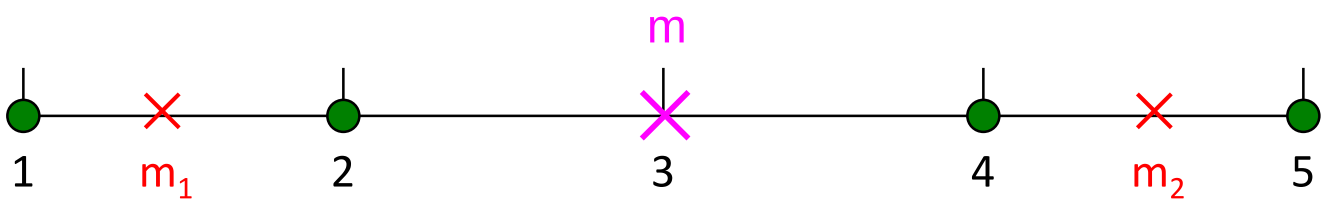
Check those slides for details

- Cluster separation: Measures how distinct or well-separated a cluster is from other clusters
  - Separation is measured by the between clusters sum of squares

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

    - where $|C_i|$ is the size of cluster *i* and *m* is the overall mean of all data points

Example



K=2 clusters:

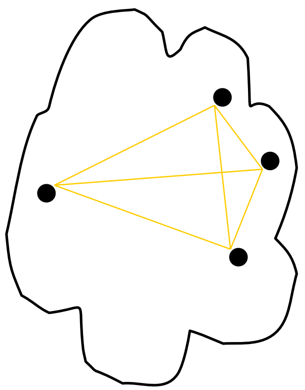
$$WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$

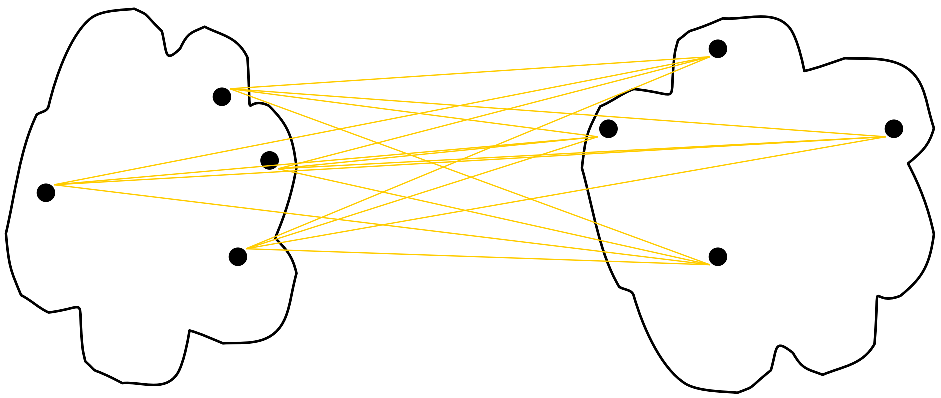
$$Total = 1 + 9 = 10$$

Internal measures of cluster validity

- A proximity graph based approach can also be used for definining cohesion and separation.
  - Cluster cohesion is the sum of the weight of all links within a cluster.
  - Cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster.



cohesion



separation

Internal Measures: Silhouette Coefficient

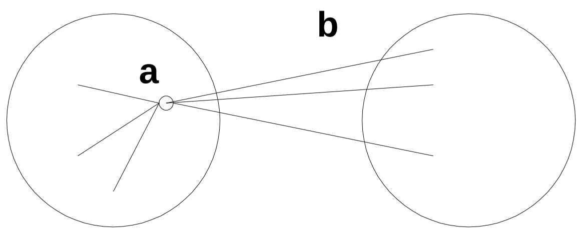
- Silhouette Coefficient combine ideas of both cohesion and separation, but for individual points, as well as clusters and clusterings
- For an individual point, *i*
  - Calculate ***a*** = average distance of *i* to the points in its cluster
  - Calculate ***b*** = min (average distance of *i* to points in another cluster)
  - The silhouette coefficient for a point is then given by

    s = (b-a)/max(a,b)

  - Typically between 0 and 1.
  - The closer to 1 the better.

(already discussed in the context of *k*-Means

Check those slides for details



- Can calculate the Average Silhouette width for a cluster or a clustering

External measures of cluster validity

- Idea: Measure the extent to which cluster labels match externally supplied class labels.
- Examples: entropy, purity

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

Cluster

Class distribution

Recall the detailed discussion on entropy In classification – decision trees

- Entropy of a cluster j: how pure in terms of the classes a cluster is: $e_j = -\sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$
  - p_{ij}: the probability of observing class i in cluster j. $p_{ij} = m_{ij}/m_j$
- Entropy of a clustering: $e = \sum_{j=1}^{k} \frac{m_j}{m} e_j$

External measures of cluster validity

- Purity focuses on the most likely class in the cluster

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

Cluster **Class distribution**

- Purity of cluster *j*: $purity_j = max\ p_{ij}$

- Purity of the clustering: $purity = \sum_{j=1}^{k} \frac{m_j}{m} purity_j$

A final note on cluster validity

"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."

Algorithms for Clustering Data, Jain and Dubes

Things you should know from this lecture

- Cluster validity measures
- Internal indices
- External indices

Acknowledgement

- The slides are based on
  - KDD I lecture at LMU Munich (Johannes Aßfalg, Christian Böhm, Karsten Borgwardt, Martin Ester, Eshref Januzaj, Karin Kailing, Peer Kröger, Eirini Ntoutsi, Jörg Sander, Matthias Schubert, Arthur Zimek, Andreas Züfle)
  - Thank you to all TAs contributing to their improvement, namely Vasileios Iosifidis, Damianos Melidis, Tai Le Quy, Han Tran.