Lecture: Machine Learning for Data Science

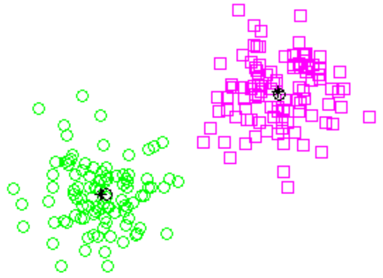
Winter semester 2021/22

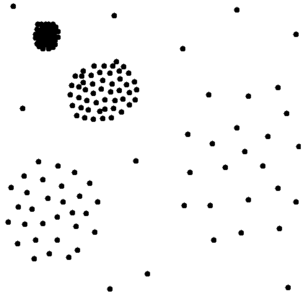
Lecture 13: Unsupervised learning – Clustering evaluation

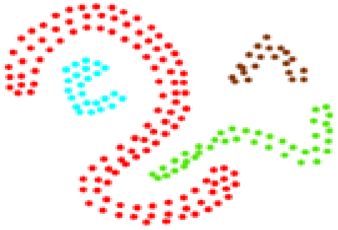
Prof. Dr. Eirini Ntoutsi

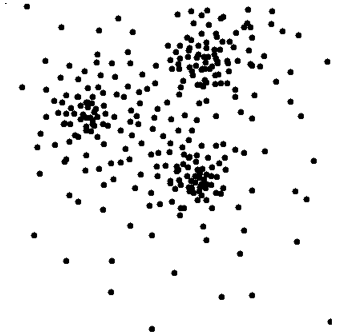
Clustering topics covered in this lecture

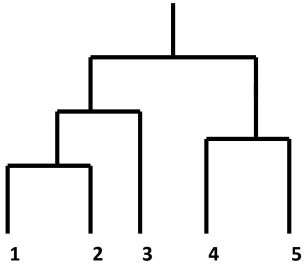
- Partitioning-based clustering
  - k-Means, k-Medoids
- Hierarchical clustering
- Density-based clustering
- Grid-based clustering
- Soft clustering
- Clustering evaluation

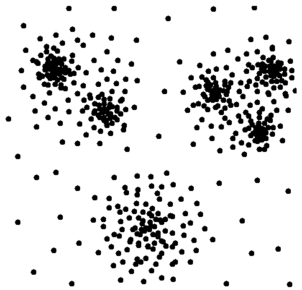


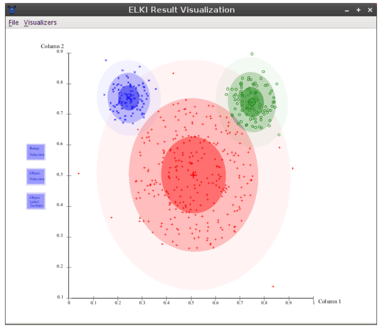












Outline

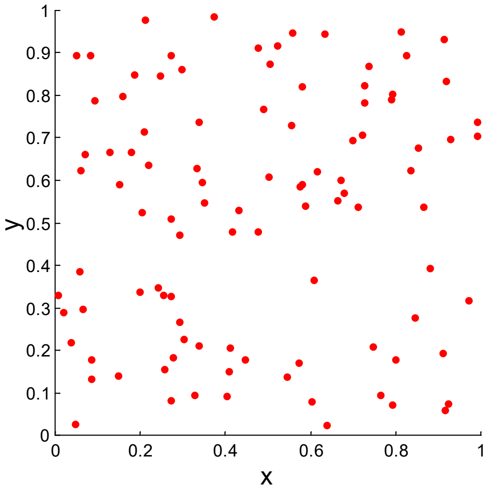
- Intro to clustering evaluation
- Internal measures
- External measures
- Things you should know from this lecture & reading material

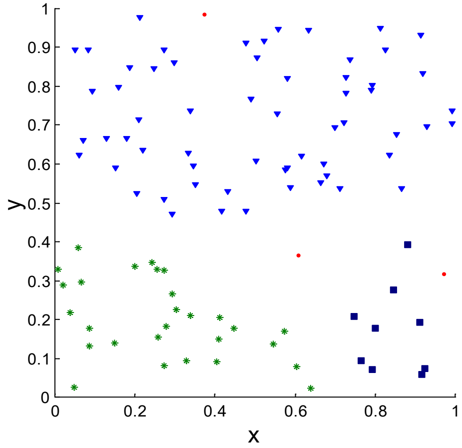
Cluster Validity

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

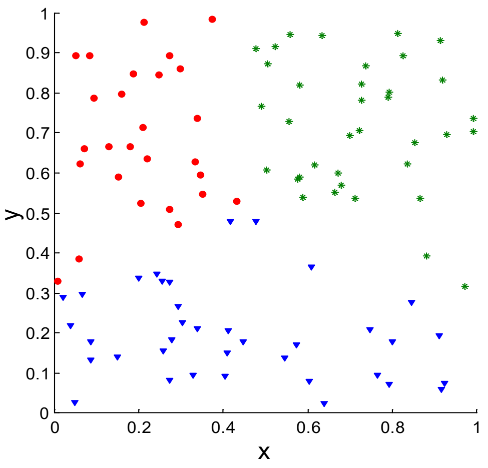
Random Points

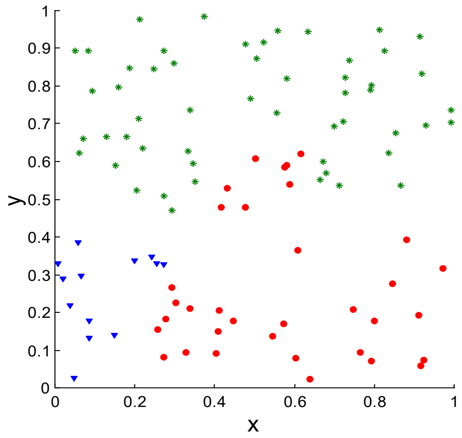




DBSCAN

K-means





Complete Link

Different Aspects of Cluster Validation

- Cluster validation has different goals:
  - Determining the clustering tendency of a dataset, 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 (and other input parameters).

- 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/Validity measures: Used to measure the goodness of a clustering structure without any external information.
    - Sum of Squared Error (SSE)
  - External Indices/Criteria/Validity measures: Used to measure the extent to which cluster labels match externally supplied class labels.
    - Entropy
  - Relative Indices/Criteria/Validity measures: Used to compare two different clusterings or clusters.
    - Often an external or internal index is used for this function, e.g., SSE or entropy

Outline

- Intro to clustering evaluation
- Internal measures
- External measures
- Things you should know from this lecture & reading material

Internal measures of cluster validity

(see also k-Means)

- Rely on cluster-member characteristics, no external information is available
- Examples: cohesion and separation
- Cluster Cohesion: How closely related are objects in a cluster
  - Cohesion is measured by the within cluster sum of squares (SSE)

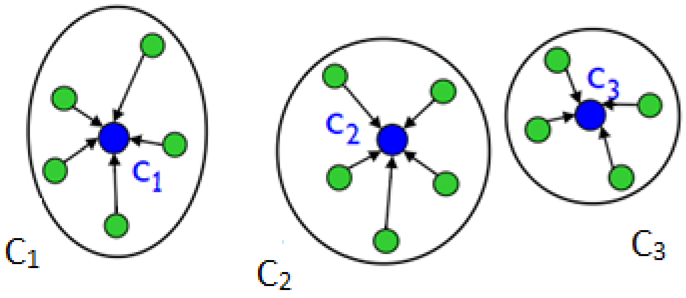
$$WSS = \sum_{i=1}^{k} \sum_{x \in C_i} (x - c_i)^2$$

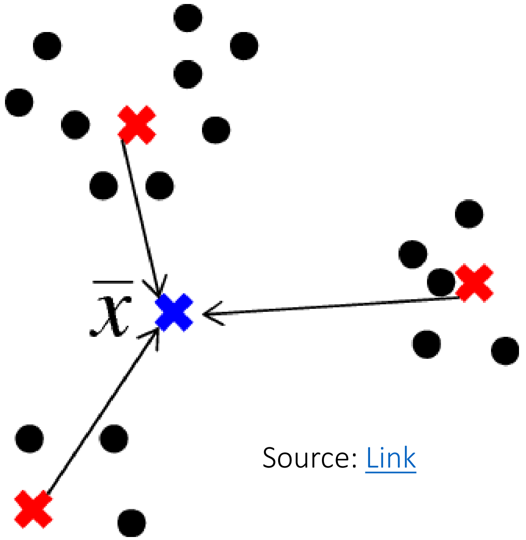
    - c_i: centroid of cluster C_i ; $|C_i|$: cardinality of cluster C_i

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

$$BSS = \sum_{i} |C_i| (c - c_i)^2$$

    - c is the overall mean of all data points

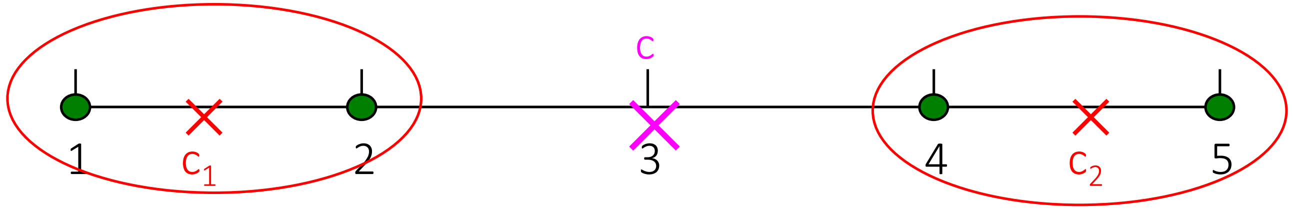




Source: Link

Example

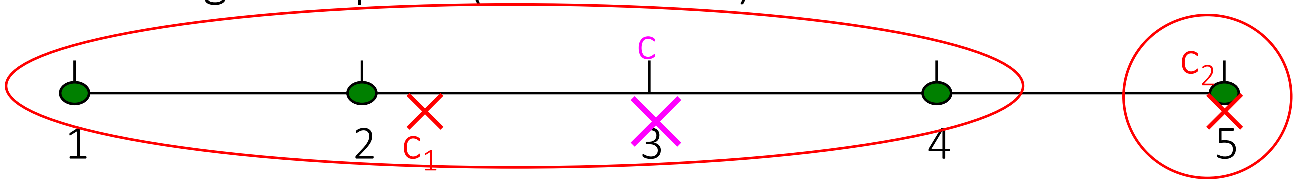
- Compute cluster cohesion and cluster separation for the following example (*k*=2 clusters)



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

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

- What about the following example? (*k*=2 clusters)

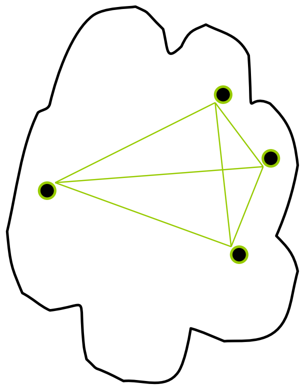


$$WSS = (1-2.3)^2 + (2-2.3)^2 + (3-2.3)^2 + (5-5)^2 = 1{,}87$$

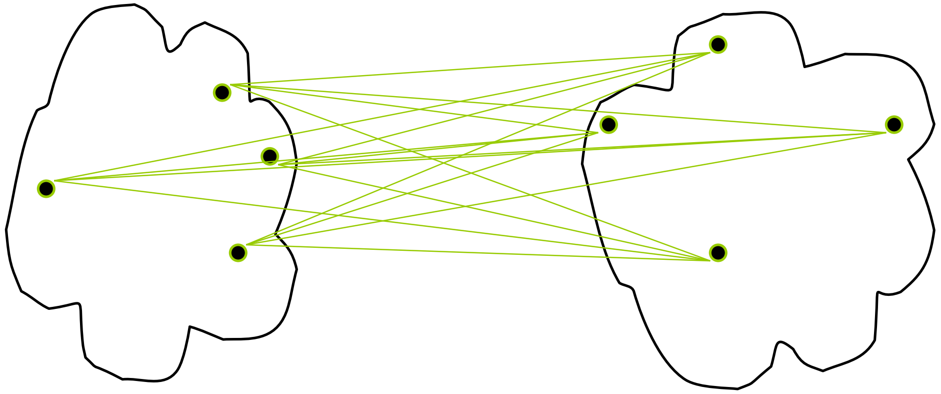
$$BSS = 3\times(3-2.3) + 1\times(3-5)^2 = 6{,}1$$

Internal measures of cluster validity

- A proximity graph based approach can also be used for defining 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

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

- Silhouette Coefficient combines ideas of cohesion and separation, for individual points, as well as for clusters and clusterings
- Silhouette coefficient of an object *i* (Kaufman & Rousseeuw 1990)
  - Let *A* be the cluster to which *i* belongs
  - Let *a(i)* the distance of object *i* to *A* (the so-called best first cluster distance)

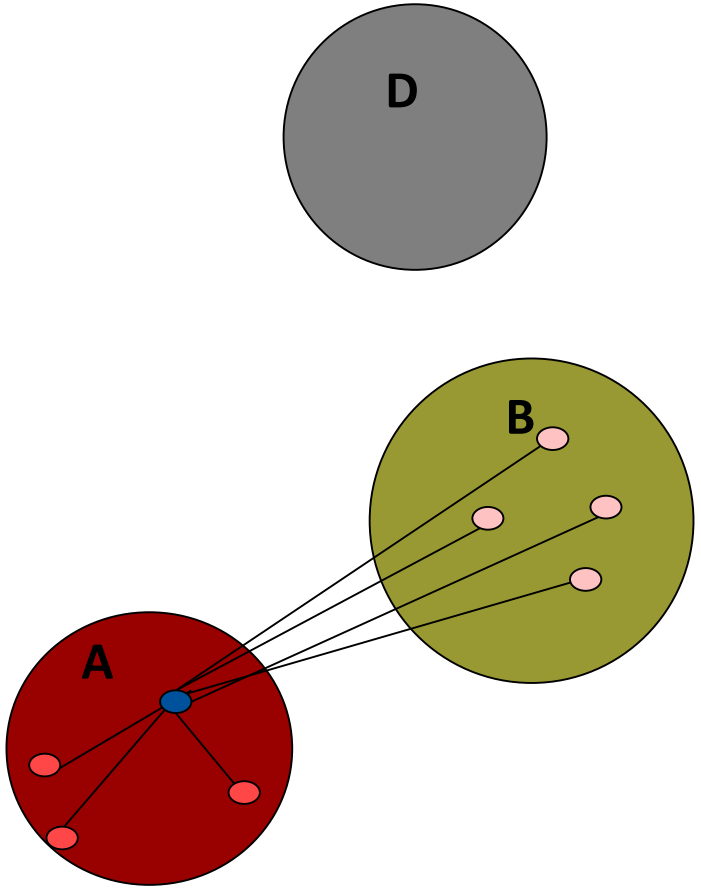
$$\begin{aligned} a(i) &:= \frac{1}{|A|-1} \sum_{j \in A, j \neq i} d(i,j) \\ &= \text{average dissimilarity of } i \text{ to all other objects of } A. \end{aligned}$$

  - Let *b(i)* be the distance of *i* to its second best cluster (we denote it by *B*)

$$b(i) := \min_{C \neq A} d(i, C).$$

where

$$\begin{aligned} d(i,C) &:= \frac{1}{|C|} \sum_{j \in C} d(i,j) \\ &= \text{average dissimilarity of } i \text{ to all objects of } C. \end{aligned}$$



What is the right number of clusters

- The Silhouette value *s(i)* of the object *i* is given by:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

- Interpreting Silhouette values
  - The closer to 1 the best

$$-1 \leq s(i) \leq +1$$

s(i) ~ -1 / 0 / +1 : bad / indifferent /good assignment

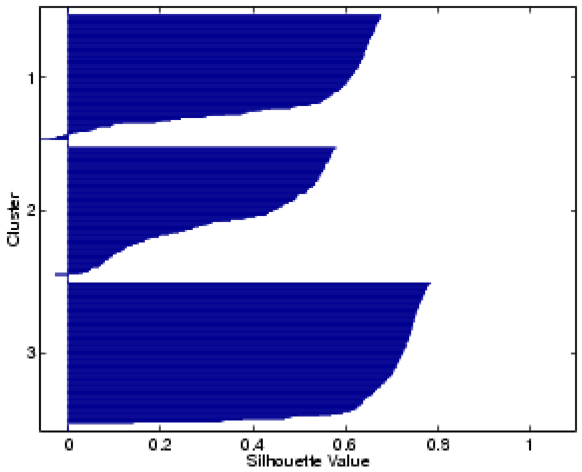
  - *s(i)~1* → *a(i) < < b(i)*. Small *a(i)* means it is well matched to its own cluster *A*. Large *b(i)* means is badly matched to its neighboring cluster *B* → good assignment
  - *s(i)~-1* → the neighbor cluster *B* seems more appropriate → bad assignment
  - s(i)~0 → in the border between the two natural clusters *A, B* → indifferent assignment

What is the right number of clusters

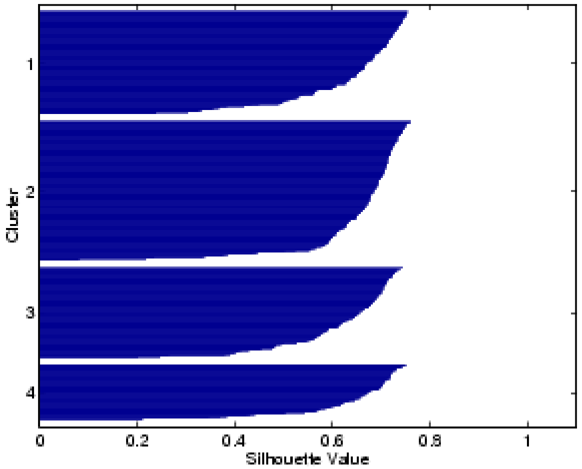
- We can computer the Silhouette value of a cluster or a clustering
- The Silhouette coefficient of a cluster is the avg silhouette of all its objects
  - Is a measure of how tightly grouped all the data in the cluster are.
- The Silhouette coefficient of a clustering is the avg silhouette of all objects
  - is a measure of how appropriately the dataset has been clustered
- How to interpret the Silhouette values?
  - As before, the closer to 1.0 the best
  - > 0,7: strong structure, > 0,5: usable structure ….

Evaluating cluster and clustering quality using Silhouette plots

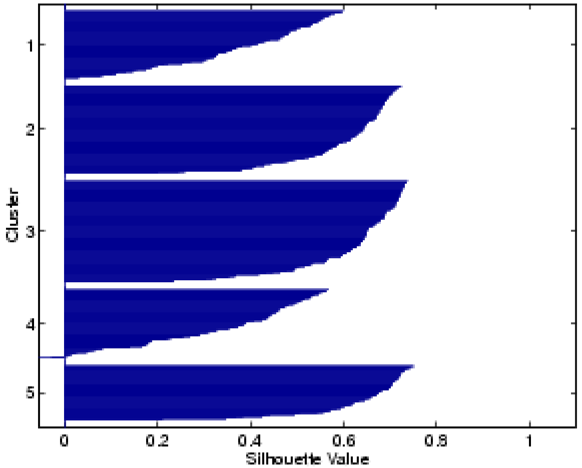
- The silhouette plot of a cluster A consists all its s(i) ranked in decreasing order.
- The entire silhouette plot of a clustering shows the silhouettes of all clusters below each other, so the quality of the clusters can be compared:



K=3



K=4



K=5

Outline

- Intro to clustering evaluation
- Internal measures
- External measures
- Things you should know from this lecture & reading material

External measures of cluster validity

- Idea: Measure the extent to which discovered clusters match externally supplied class labels.

fruit	length	width	weight
fruit 1	165	38	172
fruit 2	218	39	230
fruit 3	76	80	145
fruit 4	145	35	150
fruit 5	90	88	160
...			
fruit n	...	...	...

Unlabeled dataset for clustering

label
Banana
Banana
Orange
Banana
Orange
...

External labels used only for evaluation

- Typical measures: entropy, purity

External measures of cluster validity

- Idea: Measure the extent to which discovered clusters match externally supplied class labels.
- In our example below, for each cluster (1-6), the distribution of its instances in the different classes (Entertainment, Financial, Foreign, Metro, National, Sports) is provided.
- Intuition: Clusters should be "pure" in terms of classes

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

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports
1	3	5	40	506	96	27
2	4	7	280	29	39	2
3	1	1	1	7	4	671
4	10	162	3	119	73	2
5	331	22	5	70	13	23
6	5	358	12	212	48	13
Total	354	555	341	943	273	738

Cluster

Class distribution

External measures of cluster validity: Entropy of a cluster/ clustering

- Cluster/Clustering purity is measures in terms of entropy

Recall the detailed discussion on entropy In classification – decision trees

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

- Entropy of a cluster *j*: how pure in terms of classes a cluster is:
  - p_{ij}: the probability of observing class *i* in cluster *j*.
- Entropy of a clustering:

$$e_j = -\sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$$

$$p_{ij} = m_{ij}/m_j$$

$$e = \sum_{j=1}^{k} \frac{m_j}{m} e_j$$

External measures of cluster validity: purity

- Purity focuses on the most likely class in a 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

Outline

- Intro to clustering evaluation
- Internal measures
- External measures
- Things you should know from this lecture & reading material

Overview and Reading

- Clustering evaluation
- Internal measures
- External measures

- Reading
  - Tan P.-N., Steinbach M., Kumar V book, Chapter 8.
  - Data Clustering: A Review, https://www.cs.rutgers.edu/~mlittman/courses/lightai03/jain99data.pdf

Hands on experience

- See Project 2

Thank you

Questions/Feedback/Wishes?

Acknowledgements

- 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)
  - Introduction to Data Mining book slides at http://www-users.cs.umn.edu/~kumar/dmbook/
  - Thank you to all TAs contributing to their improvement, namely Vasileios Iosifidis, Damianos Melidis, Tai Le Quy, Han Tran.