Data Mining and Knowledge Discovery: Practice Notes

Petra Kralj Novak

Petra.Kralj.Novak@ijs.si

6.12.2018

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Discussion 1

1. Can KNN be used for classification tasks?
2. Compare KNN and Naïve Bayes.
3. Compare decision trees and regression trees.
4. Consider a dataset with a target variable with five possible values:
    1. non sufficient
    2. sufficient
    3. good
    4. very good
    5. excellent

    1. Is this a classification or a numeric prediction problem?
    2. What if such a variable is an attribute, is it nominal or numeric?

KNN for classification?

- Yes.
- A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

Discussion

1. Can KNN be used for classification tasks?
2. Compare KNN and Naïve Bayes.
3. Compare decision trees and regression trees.
4. Consider a dataset with a target variable with five possible values:
    1. non sufficient
    2. sufficient
    3. good
    4. very good
    5. excellent

    1. Is this a classification or a numeric prediction problem?
    2. What if such a variable is an attribute, is it nominal or numeric?

Comparison of KNN and naïve Bayes

	Naïve Bayes	**KNN**
Used for		
Handle categorical data		
Handle numeric data		
Model interpretability		
Lazy classification		
Evaluation		
Parameter tuning		

Comparison of KNN and naïve Bayes

	Naïve Bayes	**KNN**
Used for	Classification	Classification and numeric prediction
Handle categorical data	Yes	Proper distance function needed
Handle numeric data	Discretization needed	Yes
Model interpretability	Limited	No
Lazy classification	Partial	Yes
Evaluation	Cross validation,…	Cross validation,…
Parameter tuning	No	No

Discussion

1. Can KNN be used for classification tasks?
2. Compare KNN and Naïve Bayes.
3. Compare decision trees and regression trees.
4. Consider a dataset with a target variable with five possible values:
    1. non sufficient
    2. sufficient
    3. good
    4. very good
    5. excellent

    1. Is this a classification or a numeric prediction problem?
    2. What if such a variable is an attribute, is it nominal or numeric?

Comparison of regression and decision trees

1. Data
2. Target variable
3. Evaluation
4. Error
5. Algorithm
6. Heuristic
7. Stopping criterion

Comparison of regression and decision trees

Regression trees	Decision trees
Data: attribute-value description	
Target variable: Continuous	**Target variable**: Categorical (nominal)
Evaluation: cross validation, separate test set, …	
Error: MSE, MAE, RMSE, …	**Error**: 1-accuracy
Algorithm: Top down induction, shortsighted method	
Heuristic: Standard deviation	**Heuristic** : Information gain
Stopping criterion: Standard deviation< threshold	**Stopping criterion:** Pure leafs (entropy=0)

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Discussion

1. Can KNN be used for classification tasks?
2. Compare KNN and Naïve Bayes.
3. Compare decision trees and regression trees.
4. Consider a dataset with a target variable with five possible values:
    1. non sufficient
    2. sufficient
    3. good
    4. very good
    5. excellent

    1. Is this a classification or a numeric prediction problem?
    2. What if such a variable is an attribute, is it nominal or numeric?

Classification or a numeric prediction problem?

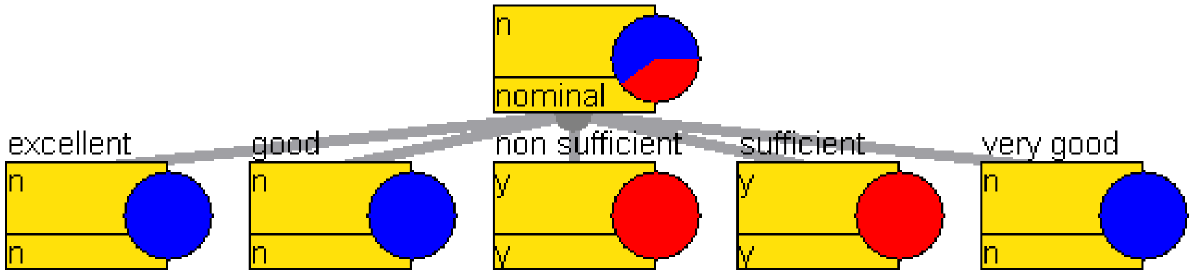
- Target variable with five possible values:
    1. non sufficient
    2. sufficient
    3. good
    4. very good
    5. excellent
- Classification: the **misclassification cost** is the same if "non sufficient" is classified as "sufficient" or if it is classified as "very good"
- Numeric prediction: The error of predicting "2" when it should be "1" is 1, while the error of predicting "5" instead of "1" is 4.
- If we have a variable with ordered values,

it should be considered numeric.

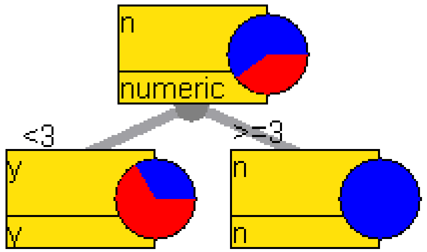
Nominal or numeric attribute?

- A variable with five possible values:
  1. non sufficient
  2. sufficient
  3. good
  4. very good
  5. Excellent

Nominal:



Numeric:

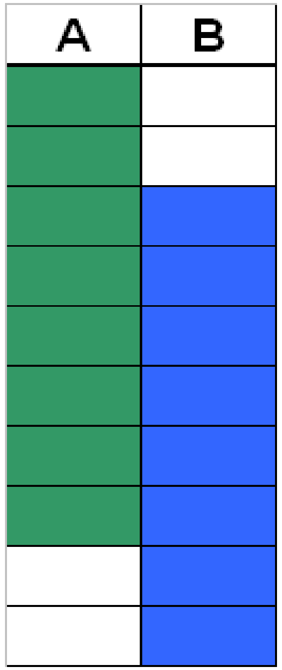


- If we have a variable with **ordered** values, it should be considered numeric.

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Discussion 2

- Transformation of an attribute-value dataset to a transaction dataset.
- What are the benefits of a transaction dataset?
- What would be the association rules for a dataset with two items A and B, each of them with support 80% and appearing in the same transactions as rarely as possible?
  - minSupport = 50%, min conf = 70%
  - minSupport = 20%, min conf = 70%
- What if we had 4 items: A, ¬A, B, ¬ B
- Compare decision trees and association rules regarding handling an attribute like “PersonID”. What about attributes that have many values (eg. Month of year)



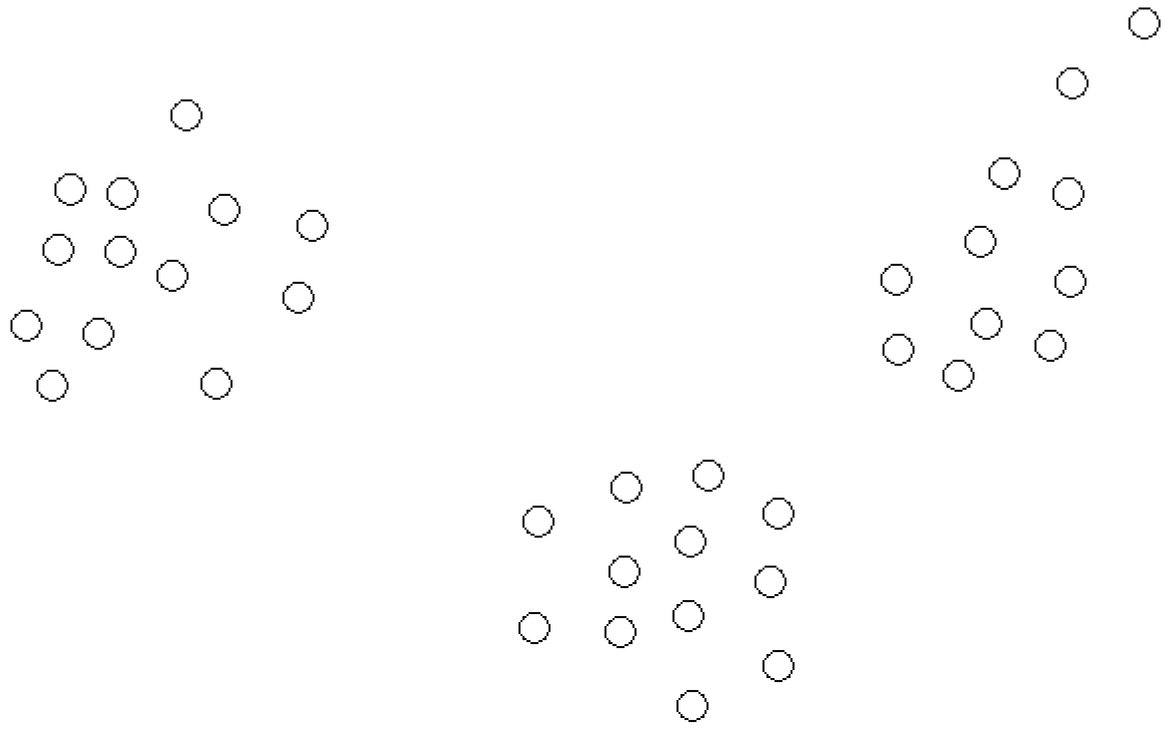
Clustering

Clustering

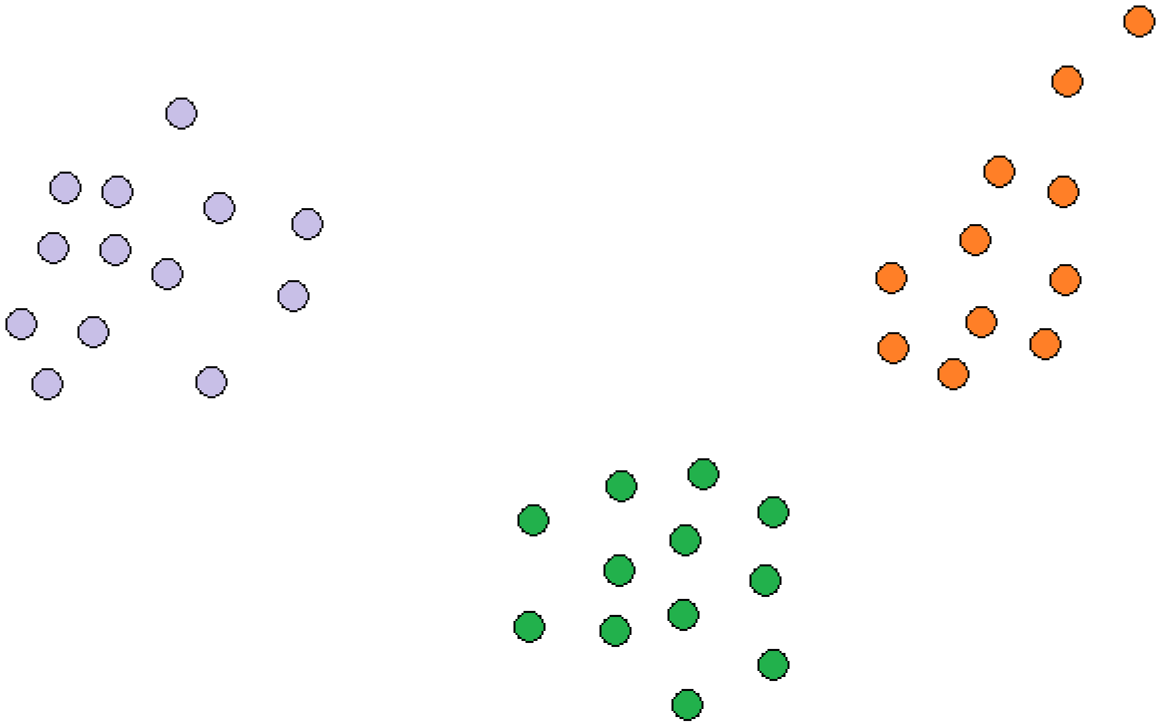
- … is the process of grouping the data instances into clusters so that objects within a cluster have high similarity but are very dissimilar to objects in other clusters.

- Wish list:
  - Identity clusters irrespective of their shapes
  - Scalability,
  - Ability to deal with noisy data,
  - Insensitivity to the order of input records.

Clustering



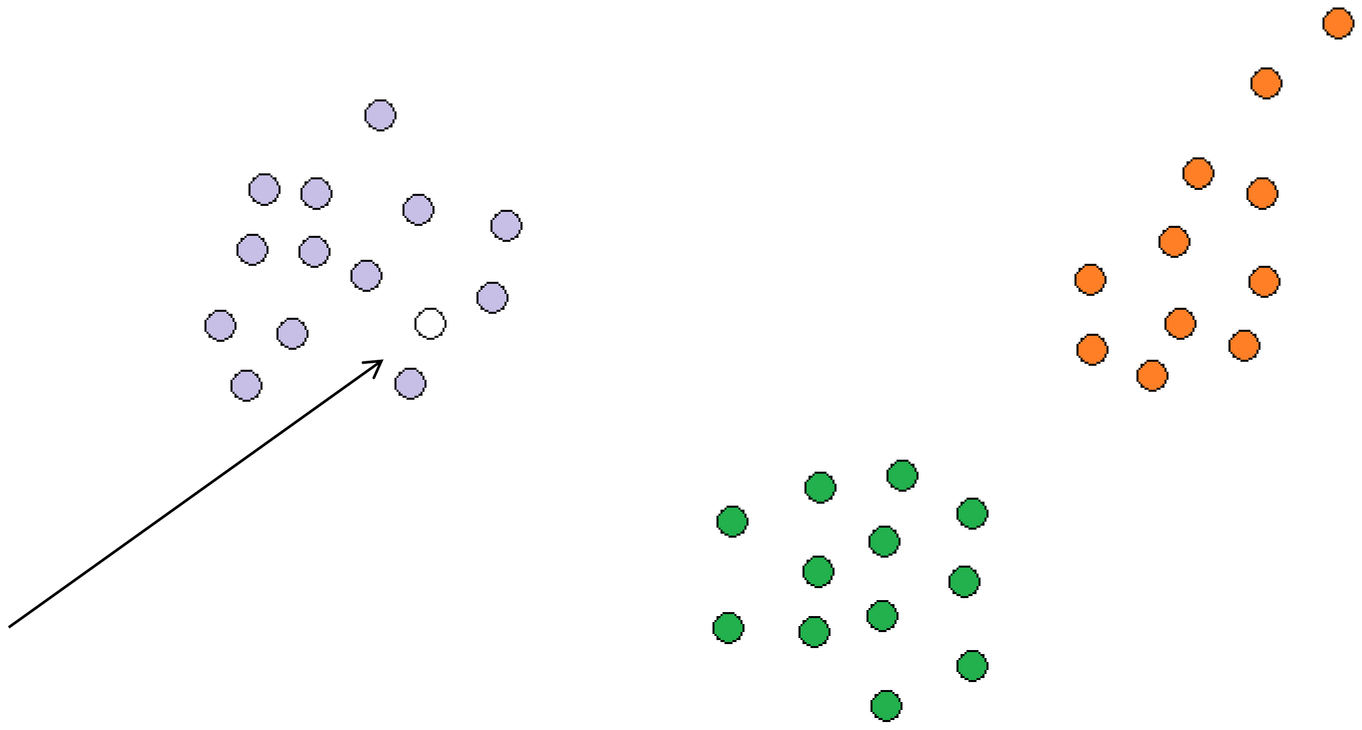
Clustering



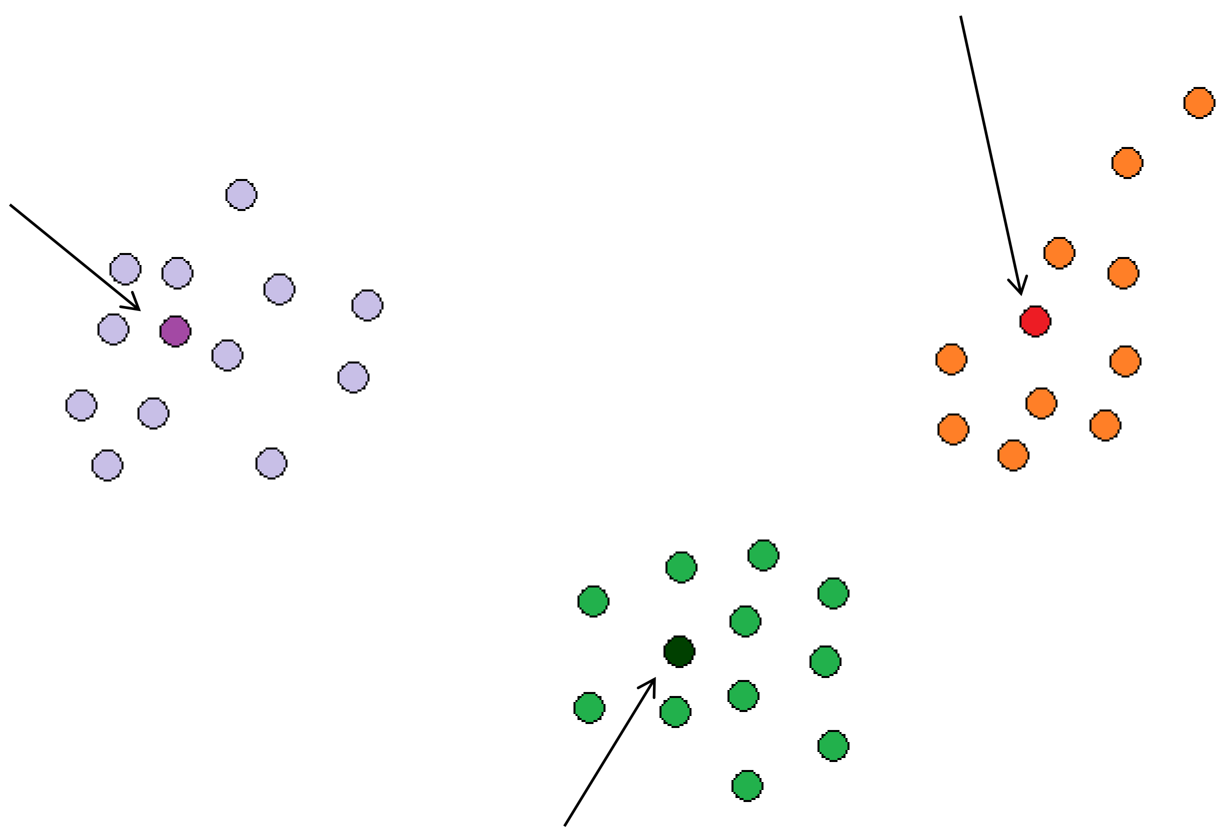
Applications

- Data mining
    - Unsupervised classification
    - Data summarization
    - Outlier analysis
    - …
- Custumer segmentation and collaborative filtering
- Text applications
- Social network analysis

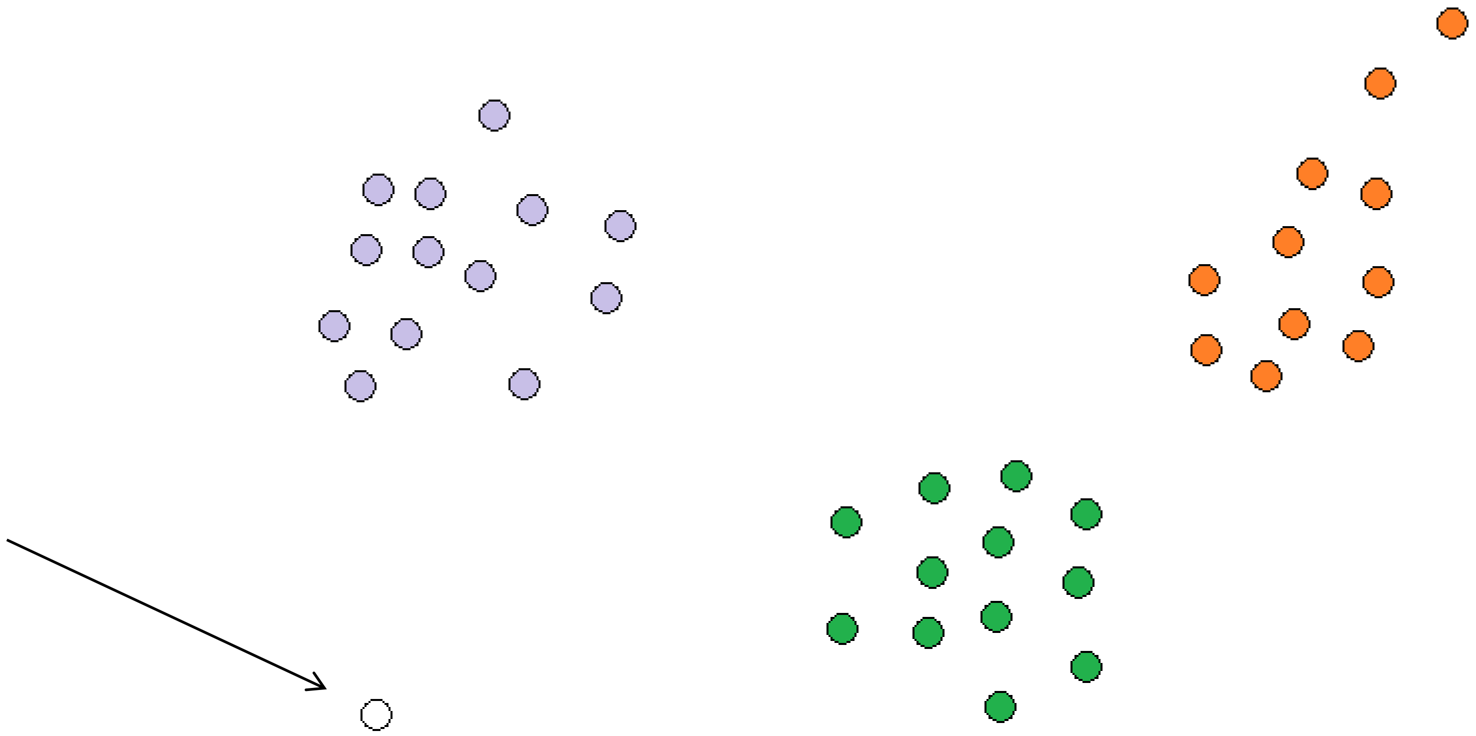
Unsupervised classification



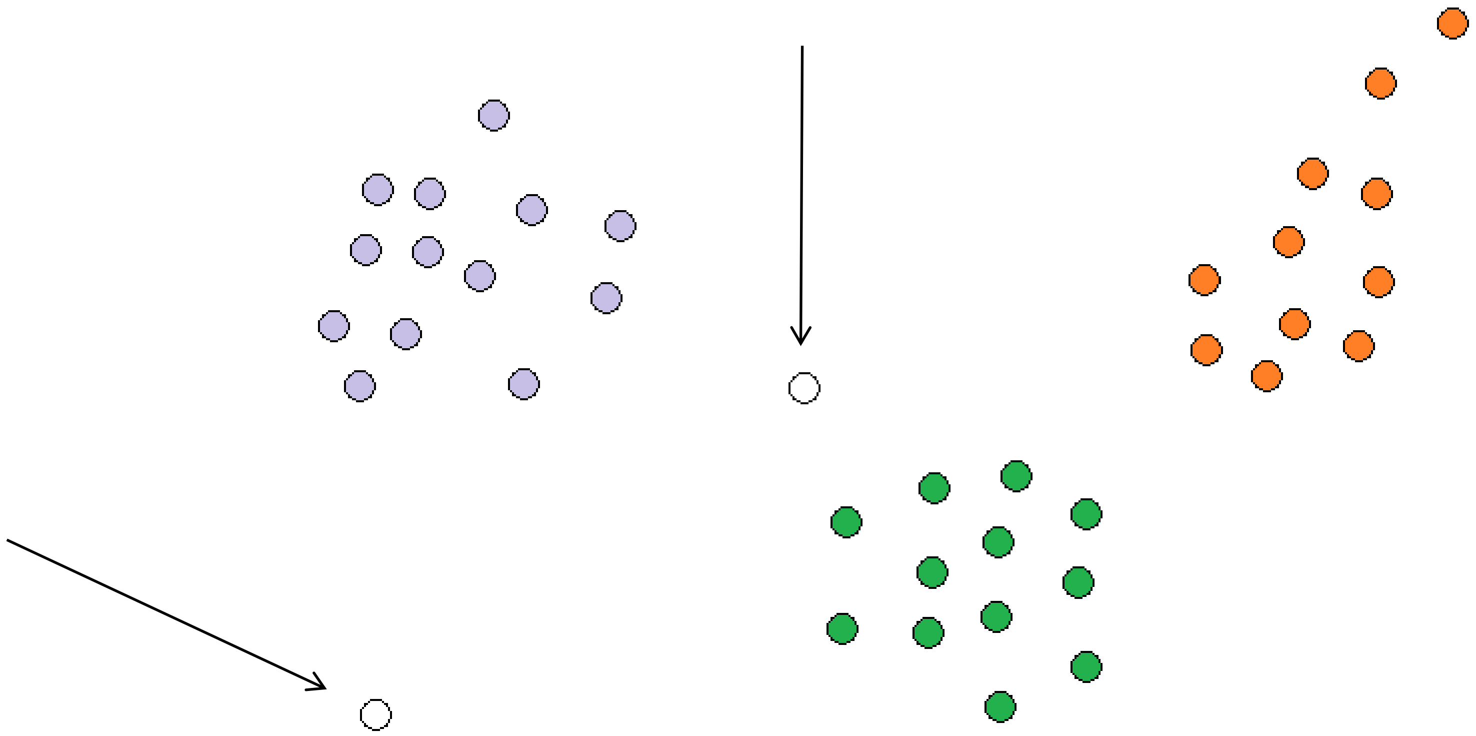
Data summarization



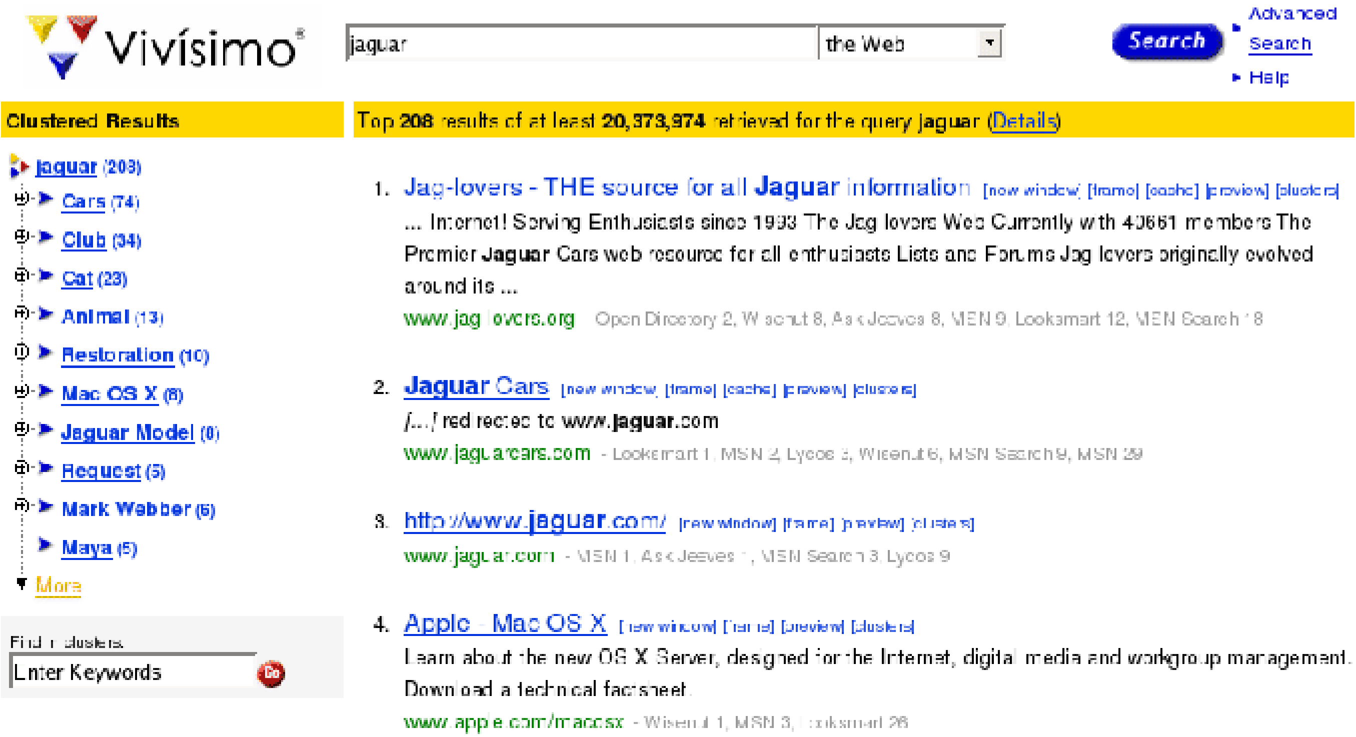
Outlier detection



Outlier detection



Text applications



Clustering types

- Partitioning
  - k-means, k-medoids, k-modes
- Hierarchical
  - Agglomerative
- Grid-based
  - Multi-resolution grid structure
  - Efficient and scalable
- Density-based
  - A cluster is a dense region of points, which is separated by low density regions, from other regions of high density
  - Algorithms: DBSCAN, OPTICS, DenClue

K-means

1. Choose **k** random instances as cluster centers
2. Assign each instance to its closest cluster center
3. Recompute cluster centers by computing the average (aka *centroid*) of the instances pertaining to each cluster
4. If cluster centers have moved, go back to Step 2

(Equivalent termination criterion: stop when assignment of instances to cluster centers has not changed)

Alternatives: K-medoids, K-modes

- Might get stuck in local minia
- Silhuette for finding the optimal K

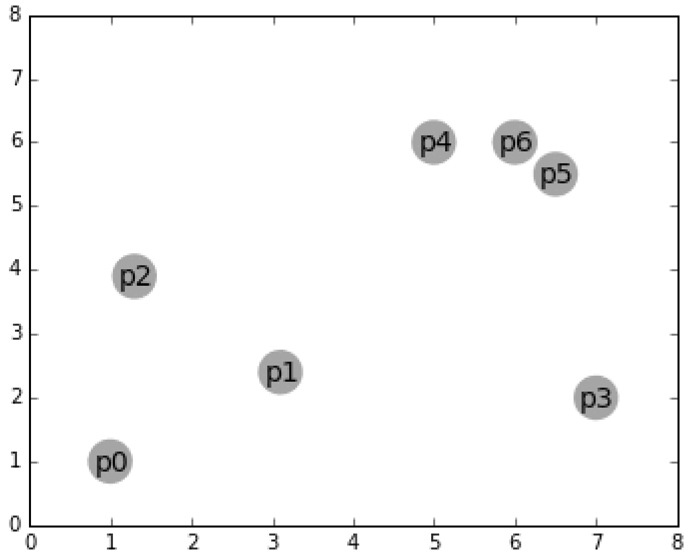
Agglomerative clustering

1. Start with a collection **C** of **n** singleton clusters
   - Each cluster contains one data point $\mathbf{c_i = \{x_i\}}$
2. Repeat until only one cluster is left:
   1. Find a pair of clusters that is closest: min $\mathbf{D(c_i, c_j)}$
   2. Merge the clusters $\mathbf{c_i}$ and $\mathbf{c_j}$ into $\mathbf{c_{i+j}}$
   3. Remove $\mathbf{c_i}$ and $\mathbf{c_j}$ from the collection **C**, add $\mathbf{c_{i+j}}$

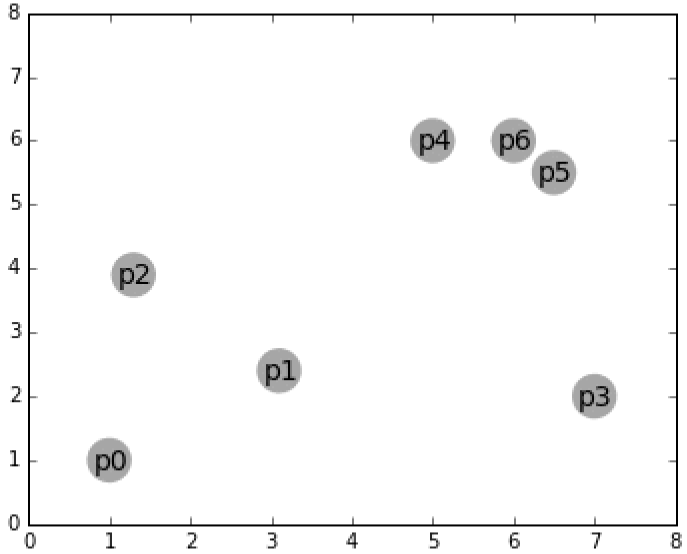
Some new index, not a sum

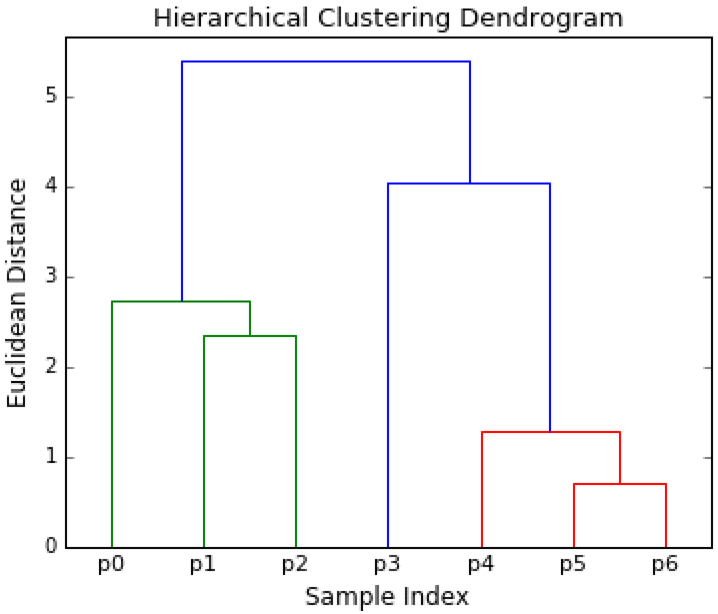
- Time and space complexity
- Sensitive to noisy data

Agglomerative clustering - example

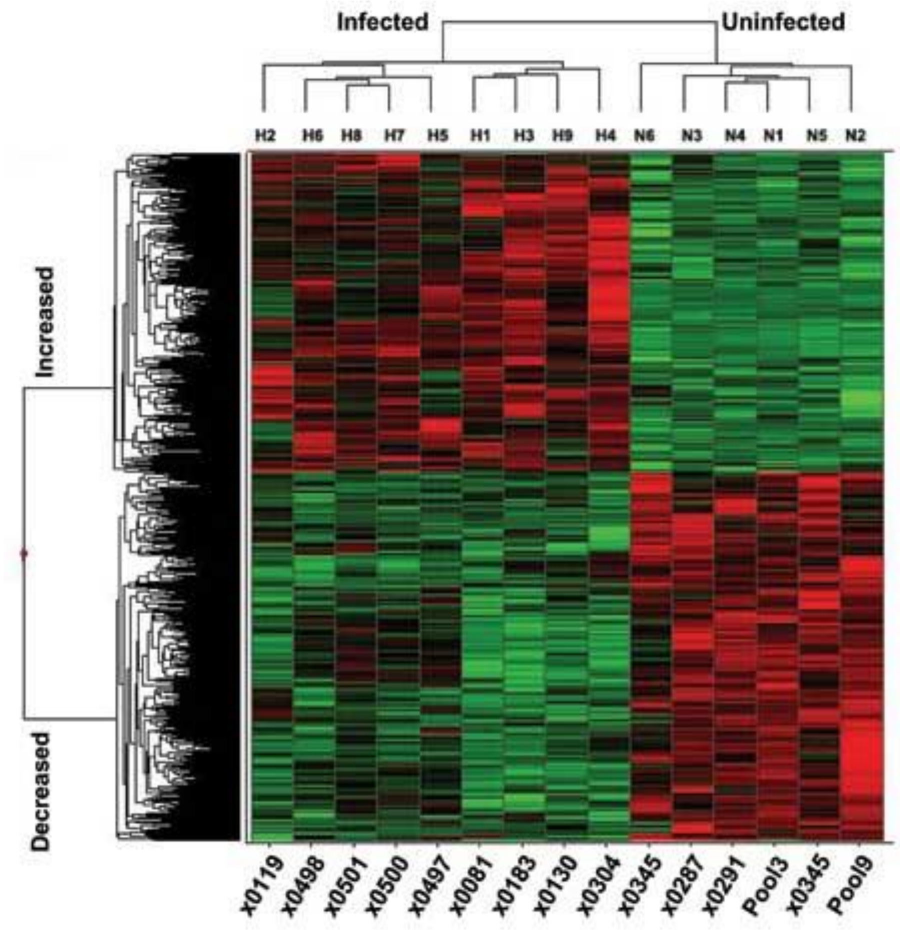


Agglomerative clustering - dendrogram





Example: Hierarchical clustering of genes

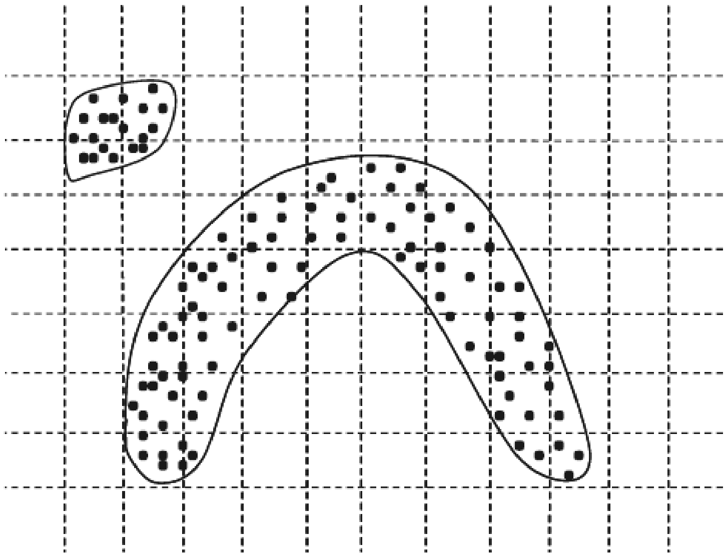


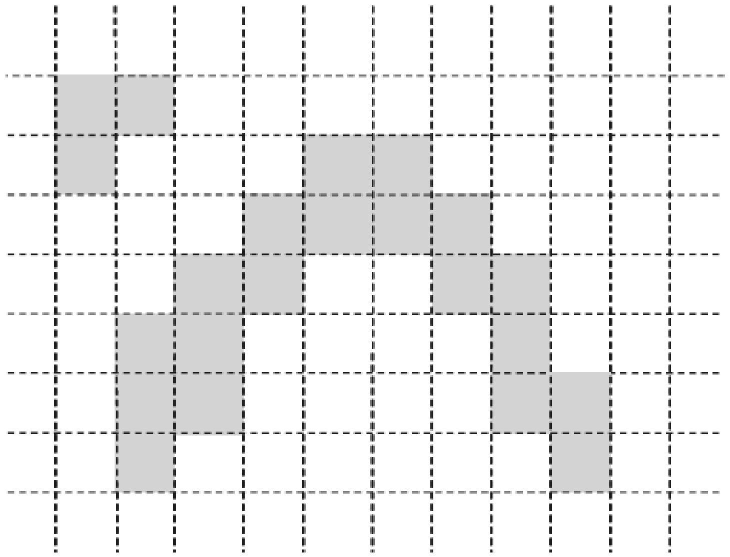
Grid-based (parameters **p** and **τ**)

1. Discretize each dimension of **D** into **p** ranges
2. Determine dense grid cells at level **τ**
3. Create graph where dense grid cells are connected if they are adjacent
4. Determine connected components of graph
5. Return: points in each connected component as a cluster

Grid-based (parameters **p** and **τ**)

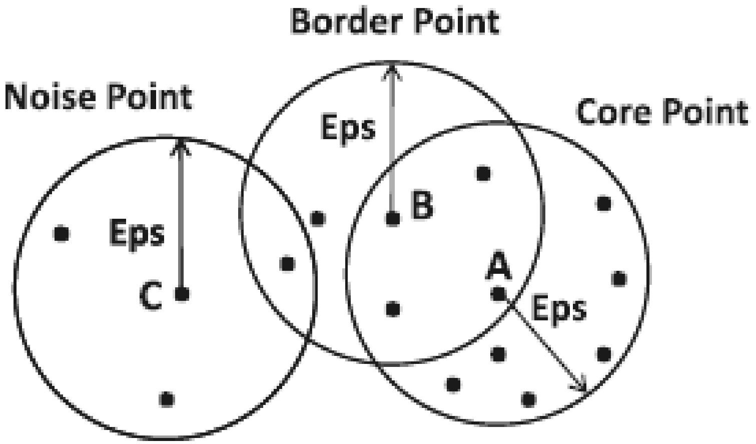
1. Discretize each dimension of **D** into **p** ranges
2. Determine dense grid cells at level **τ**
3. Create graph where dense grid cells are connected if they are adjacent
4. Determine connected components of graph
5. Return: points in each connected component as a cluster





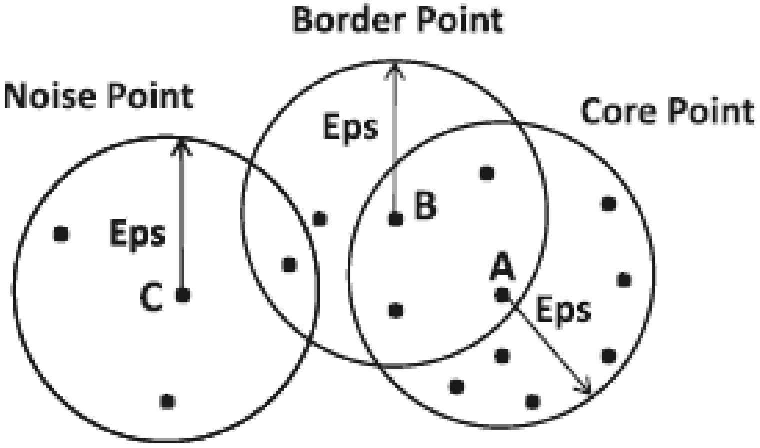
Density based clustering *DBSCAN*(Data: *D*, Radius: *Eps*, Density: τ)

- *Core point:* A data point is defined as a *core* point, if it contains at least τ data points within a radius *Eps* within a radius *Eps*.
- *Border point:* A data point is defined as a *border* point, if it contains less than τ points, but it also contains at least one core point within a radius *Eps*.
- *Noise point:* A data point that is neither a core point nor a border point is defined as a *noise* point.



Density based clustering
DBSCAN(Data: *D*, Radius: *Eps*, Density: τ)

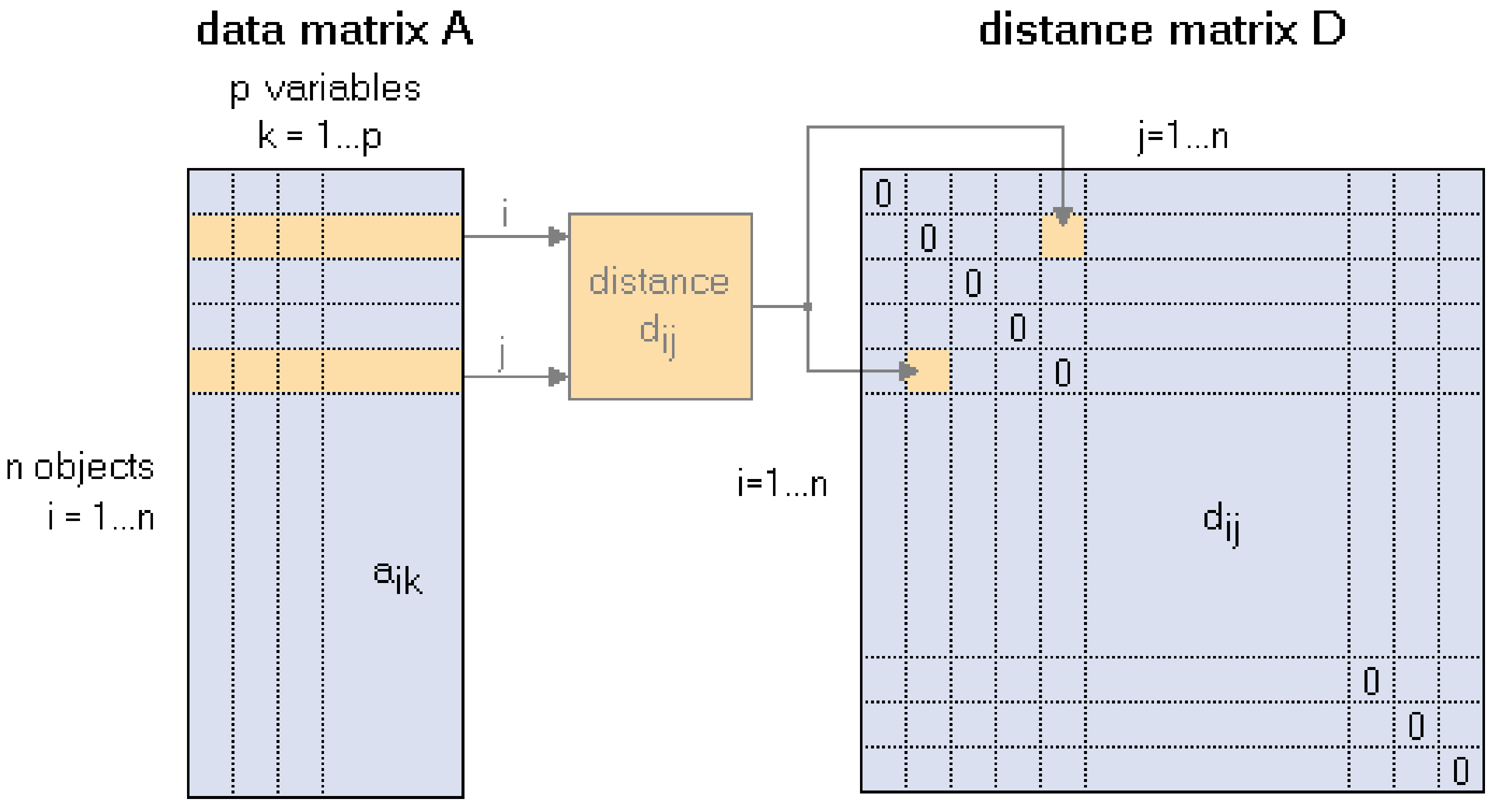
1. Determine core, border and noise points of *D* at level (*Eps, τ*);
2. Create graph in which core points are connected if they are within *Eps* of one another;
3. Determine connected components in graph;
4. Assign each border point to connected component with which it is best connected;
5. **Return** points in each connected component as a cluster;



Similarity / distance measures

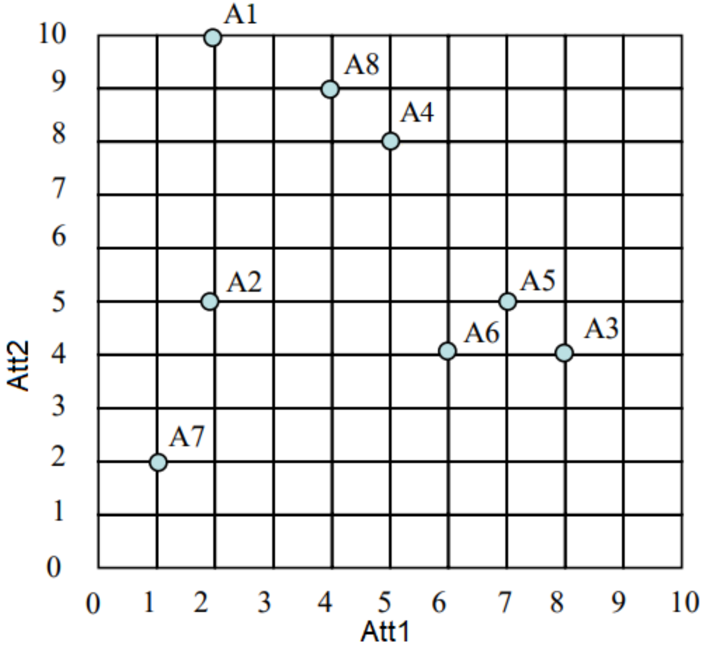
- The similarity measure depends on characteristics of the input data:
  - Attribute type: binary, categorical, continuous
  - Sparseness
  - Dimensionality
  - Type of proximity

Distance matrix



Distance matrix example

	Att1	Att2
A1	2	10
A2	2	5
A3	8	4
A4	5	8
A5	7	5
A6	6	4
A7	1	2
A8	4	9



	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Euclidian $\rightarrow$ $Dist(A,B) = \sqrt[2]{(Att1(A) - Att1(B))^2 + (Att2(A) - Att2(B))^2}$

Distance measures

Euclidean	$d(x,y)=\sqrt{\sum(x_i-y_i)^2}$
Squared Euclidean	$d(x,y)=\sum(x_i-y_i)^2$
Manhattan	$d(x,y)=\sum\lvert(x_i-y_i)\rvert$
Canberra	$d(x,y)=\sum\frac{\lvert x_i-y_i\rvert}{\lvert x_i+y_i\rvert}$
Chebychev	$d(x,y)=\max(\lvert x_i-y_i\rvert)$
Bray Curtis	$d(x,y)=\frac{\sum\lvert x_i-y_i\rvert}{\sum x_i+y_i}$
Cosine Correlation	$d(x,y)=\frac{\sum(x_iy_i)}{\sqrt{\sum(x_i)^2\sum(y_i)^2}}$
Pearson Correlation	$d(x,y)=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(y_i-\bar{y})^2}\sqrt{\sum(y_i-\bar{y})^2}}$
Uncentered Peason Correlation	$d(x,y)=\frac{\sum x_iy_i}{\sqrt{\sum(y_i-\bar{y})^2}\sqrt{\sum(y_i-\bar{y})^2}}$
Euclidean Nullweighted	Same as Euclidean, but only the indexes where both x and y have a value (not NULL) are used, and the result is weighted by the number of values calculated. Nulls must be replaced by the missing value calculator (in dataloader).

Minkowski distance

$$D(X,Y)=\left(\sum_{i=1}^{n}|x_i-y_i|^p\right)^{1/p}$$

Aggarwal, C. C. (2015). *Data mining: the textbook*. Springer. (Chapter 3)

Evaluation of clustering

- Objective functions in clustering formalize the goal of attaining high intra-cluster similarity and low inter-cluster similarity.
- Internal evaluation:
  - Sum of square distances to centroid
  - Intracluster to intercluster distance ratio
  - Silhuette coefficient
  - -biased towards algorithms
- External evaluation: we can use a set of classes in an evaluation benchmark (gold standard, ground truth)

Discussion

- Similarity vs. distance
- List algorithms that are based on distance/similarity

….. 19.12.2018

- Written exam
  - 60 minutes of time
  - 4 tasks:
    - 2 computational (60%),
    - 2 theoretical (40%)
  - Literature is not allowed
  - Each student can bring
    - **one hand-written A4 sheet of paper**,
    - and **a hand calculator**
- Data mining seminar proposal
  - One page seminar proposal on **paper**