Data Mining, Lecture 1

Introduction & Distance Function

S. Nõmm

[1]Department of Software Science, Tallinn University of Technology TalTech

05.09.2023

Course organisation

- Mode of studies: in class attendance. Lectures are recorded and shared via MS Teams environment. No hybrid mode is offered.
- Closed book tests will be conducted in class only!!
- During first few practices the students will be given a lot of guidance in solving the exercises. While the course progresses amount of guidance will be decreased and practices will be used to overview students assignments, answer the questions and organize small competitions.
- Exception: there may be a few online lecture and practices.
- When joining online, please keep the microphone muted. **Only teacher or lecturer may initiate meeting and start recording.** If you wish to ask a question, use the chat option. It is recommended to attend lectures and practices in class.
- It is recommended to install MS Teams as stand alone application.
- It is mandatory to use "R" for all the computational exercises. The students are encouraged to install "R-studio".

Course organization: (administrative part)

- For all correspondence concerning the course use email sven.nomm@ttu.ee Please avoid using phone.
- Grading:
  - 2x mandatory closed book tests. Each test gives 10% of the final grade. For each test one make-up attempt will be given. Tests are performed in class only.
  - 3x mandatory home assignments (Computational assignment + short write up.) 10% of the final grade each. Assignments are accepted up to one week after the deadline with the penalty of 10% for each day except Saturday and Sunday. Upon submission the assignment each student will be asked to defend it.
  - final exam (gives 50 % of the final grade): Computational assignment and written report on assigned topic + discussion with lecturer. Note examination date will be announced in the end of November - beginning of December. Prerequisites:
    - all 2 closed book tests are accepted (graded as 51 or higher)
    - all 3 home assignments are accepted (graded as 51 or higher)
  - In addition to the mandatory tests the lecturer may give grading points to the students active during the lectures and practices. Such grading points are usually assigned based on non-mandatory short tests given

Course organization: Tentative program

- 05.09.23 Introduction and Distance function.
- 12.09.23 Cluster analysis I.
- 19.09.23 Cluster analysis II. (EM algorithm)
- 26.09.23 Anomaly and outlier analysis.
- 03.10.23 Classification I. (05.10 Home assignment I defense.)
- 10.10.23 Classification II.
- 17.10.23 Regression analysis.
- 24.10.23 Association pattern mining.
- 31.10.23 Closed book test I. (02.11 Home assignment II defense.)
- 14.11.23 Time series mining.
- 21.11.23 Data streams mining.
- 28.11.23 Text data mining.
- 06.12.23 Graph data mining and Social networks analysis.
- 13.12.23 Privacy preserving data mining.
- 20.12.23 Closed book test II (22.12 Home Assignment III

Course organization: Grading *vs.* expected knowledge

- **Excellent 91 -100** Able to apply all the methods and techniques, thought during the course, on practice. Interpret the results and explain theoretical foundations of the applied techniques Discuss achieved results with respect of possible further analysis. Able to learn new techniques independently and apply them on practice.
- **Very Good 81 -90** Able to apply all the methods and techniques, thought during the course, on practice. Interpret the results and explain theoretical foundations of the applied techniques. Discuss achieved results with respect of possible further studies.
- **Good 71 -80** Able to apply all the methods and techniques, thought during the course, on practice, interpret the results and explain theoretical foundations of the applied techniques.
- **Satisfactory 61-70** Able to apply all the methods and techniques, thought during the course, on practice. Interpret the results.
- **Acceptable 51-60** Able to apply all core methods and techniques, thought during the course, on practice. Interpret the results.

References

The structure of the present course, main notations and definitions are inherited from [1]. [7] provides basic knowledge of "R" for the data mining assignments. Implementation of different data mining algorithms in "R" is discussed by [3]. Some data mining methods are borrowed from the neighbouring fields of research, such as Machine Learning [2], Pattern Recognition [6] and Feature Extraction [4]. Lectures related to the networked data mining are based on [5].

[1] C.C. Aggarwal. *Data Mining: The Textbook*. Springer International Publishing, 2015.

[2] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley, 2013.

[3] P. Cichosz. *Data Mining Algorithms: Explained Using R*. Wiley, 2015.

[4] I. Guyon, S. Gunn, M. Nikravesh, and L.A. Zadeh. *Feature Extraction: Foundations and Applications*. Studies in Fuzziness and Soft Computing. Springer Berlin Heidelberg, 2008.

[5] E.D. Kolaczyk and G. Csárdi. *Statistical Analysis of Network Data with R*. Use R! Springer New York, 2014.

[6] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Elsevier Science, 2008.

[7] G. Williams. *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*. Use R! Springer New York, 2011.

Course organization: administrative part (continued)

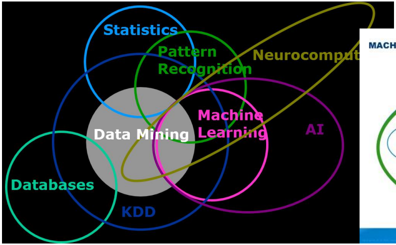
- You are expected to attend the lectures and practices. Lecture slides do not contain all the information. Also, this is the place where you can gain experience!
- Consultations: By appointment only! Please do not hesitate to ask if you need consultation.
- It is advisable to write your own notes!
- Mind academic 15 min!
- Many concepts introduced during the course require understanding of the probability theory and statistics.
- "R" and some related packages will be used to perform computational part of the assignments.
- **No Plagiarism in any of assignments and final project!!!**. You should cite all the references, including software and extra libraries. The student should be able to explain the meaning of all the computations performed, interpret and present the results.
- If you feel unwell or have any symptoms of infection diseases please do not come to the class!!!

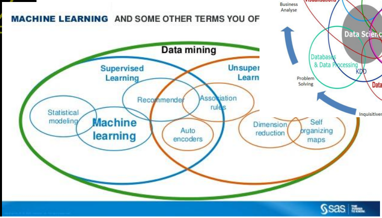
It is advisable to refresh your knowledge of:

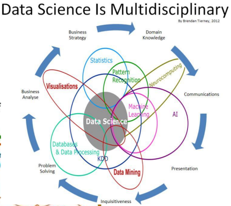
- Mathematics (calculus and linear algebra).
- Statistics.
- Programming.

What data mining is?

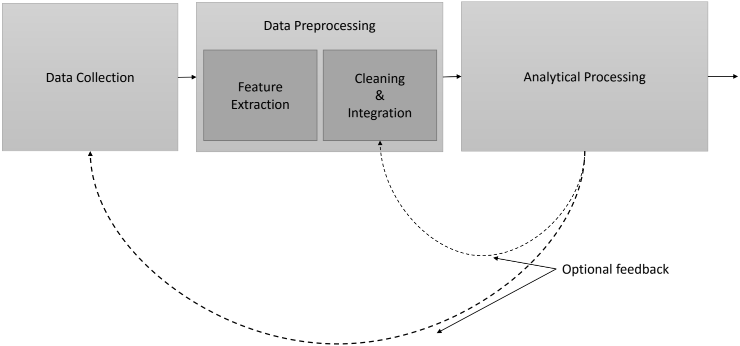
- Aggarwal: "Data mining is the study of collecting, cleaning, processing, analyzing and gaining useful insides from the data."
- Williams: "Data mining is the art and science of intelligent data analysis."







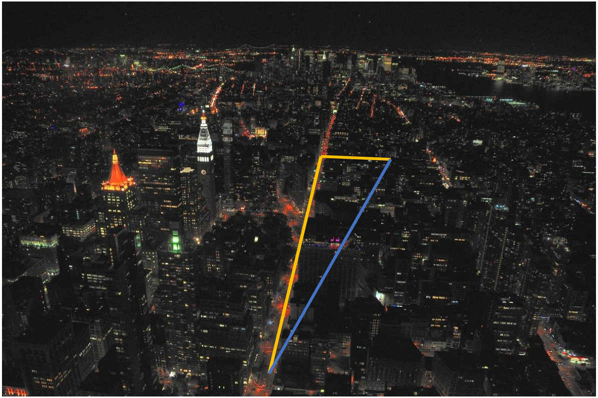
The Data Mining Process



Attribute, Feature, Dimensionality

- **Widely used explanation** Different measured properties of the process are referred as *features*, *attributes* or *dimensions*.
- In order to avoid confusion, here and after, single measured property of the process will be referred as *attribute*, sets or tuples of attributes will be referred as *features*. Note! That feature may contain just one attribute therefore attribute is always a feature but not vice verse! *Dimensionality* is the property of the process describing number of attributes.

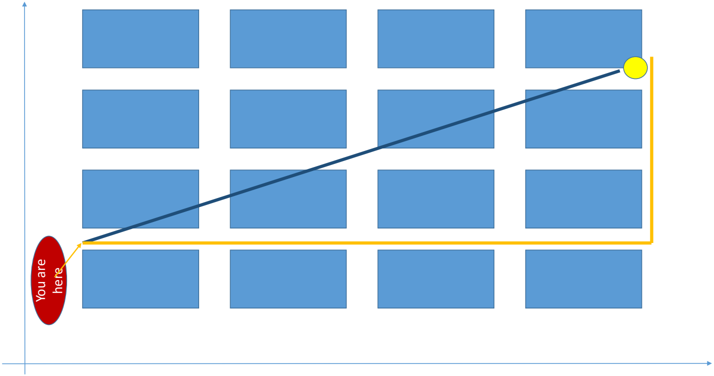
Distance ?



This is the distance used to compute the price of a taxi ride

Actual distance between the starting end ending points of your journey

Distance ?



Real world distances

- Euclidean distance

$$S(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- Manhattan distance also referred as city block distance or taxicab distance

$$S(x, y) = \sum_{i=1}^{n} \mid x_i - y_i \mid$$

Let us suppose that $(2, 3)$ are the coordinates of the starting point and (11,14) are the coordinates of the destination. Then Euclidian distance between the starting point and destination is: 14.21. At the same time Manhattan distance is 20.

Similarity or Distance

Problem statement: *Given two objects* $\mathcal{O}_1$ *and* $\mathcal{O}_2$*, determine a value of the similarity between two objects*

Distance function

Distance function is one of most fundamental notions in Machine learning and Data mining. Formally defined in pure mathematics as *metric* function. It provides measure of similarity or distance between two elements.

Definition

A function $S : X \times X \to \mathbb{R}$ is called metric if for any elements x, y and z of X the following conditions are satisfied.

1. Non-negativity or separation axiom
$$S(x, y) \geq 0$$
2. Identity of indiscernible, or coincidence axiom
$$S(x, y) = 0 \Leftrightarrow x = y$$
3. Symmetry
$$S(x, y) = S(y, x)$$
4. Subadditivity or triangle inequality
$$S(x, z) \leq S(x, y) + S(y, z)$$

Distance function: Examples 1 (Most common distance functions)

- Euclidean distance

$$S(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

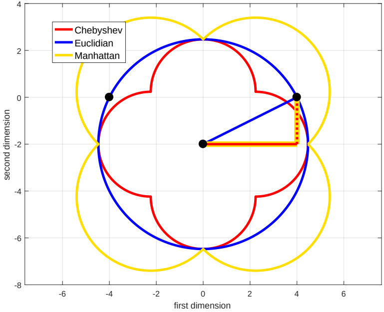
- Manhattan distance also referred as city block distance or taxicab distance

$$S(x,y) = \sum_{i=1}^{n} \mid x_i - y_i \mid$$

- Chebyshev distance

$$S(x,y) = \max_i \Big(\mid x_i - y_i \mid \Big)$$

Distance function: Examples 2



Euclidean distance

$$S(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Manhattan distance

$$S(x,y) = \sum_{i=1}^{n} \mid x_i - y_i \mid$$

Chebyshev distance

$$S(x,y) = \max_i \left(\mid x_i - y_i \mid \right)$$

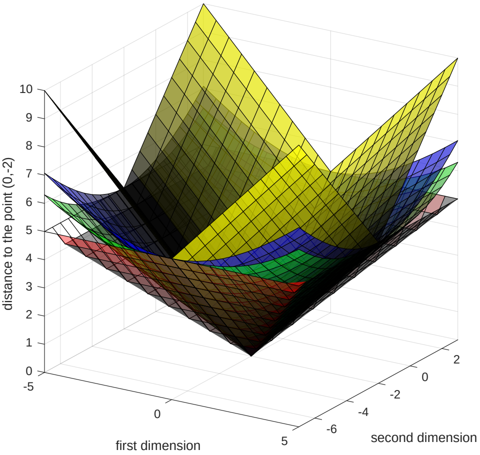
Distance function:Examples 3 Minkowsky distance

$$S(x, y) = \Big(\sum_{i=1}^{d} \mid x_i - y_i \mid^p\Big)^{1/p}$$

- $p < 1$ triangle inequality is violated, therefore for the values of p smaller than one, equation above is not a distance function.
- $p = 1$ case of Manhattan distance.
- $p = 2$ case of Euclidian distance.
- $p \to \infty$ case of Chebyshev distance.

Distance function: Examples 4

3D representation of the Minkovski distances for different values of parameter p. $p = 1$ - yellow surface, Manhattan; $p = 2$ - blue surface, Euclidean,; $p = 3$ - green surface; $p \to \infty$ - red surface, Chebyshev.

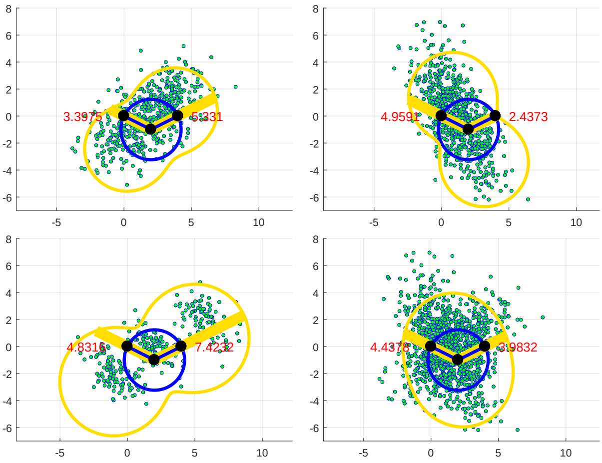


Distance function: Examples 5

Mahalanobis distance

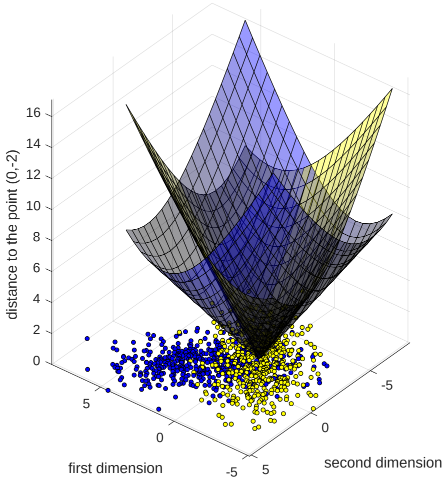
$$S(x, y) = \sqrt{(x-y)^T C^{-1} (x-y)}$$

where C is the covariance matrix. Takes into account impact of data distribution.



Distance function: Examples 6

- Impact of the rotation of underlying data set.



Distance function: Examples 7

- Canberra distance

$$S(x, y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

weighted version of Manhattan distance.

- Cosine distance Cosine similarity is the measure of the angle between two vectors

$$S_c(x, y) = \frac{x \cdot y}{\|x\|\|y\|}$$

Usually used in high dimensional positive spaces, ranges from -1 to 1. Cosine distance is defined as follows

$$S_C(x, y) = 1 - S_c(x, y)$$

Distance function: Examples 8

- Levenshtein or SED distance. SED - minimal number of single -charter edits required to change one string into another. Edit operations are as follows:
  - insertions
  - deletions
  - substitutions
- SED(delta, delata)$=1$ delete "a" or SED(kitten,sitting)$=3$: substitute "k" with "s",substitute "e" with "i", insert "g".
- Hamming distance Similar to Levenshtein but with substitution operation only. Frequently used with categorical and binary data.
- Specialized similarity measures Distance and similarity functions applicable to the graphs, temporal data etc. These topics are left outside of the framework of the present course.

Impact of High Dimensionality (Curse of Dimensionality)

Curse of dimensionality - term introduced by Richard Bellman. Referred to the phenomenon of efficiency loss by distance based data-mining methods. Let us consider the following example.

- Consider the unit cube in d - dimensional space, with one corner at the origin.
- What is the Manhattan distance from the arbitrary chosen point inside the cube to the origin?

$$S(\bar{0}, \bar{Y}) = \sum_{i=1}^{d} (Y_i - 0)$$

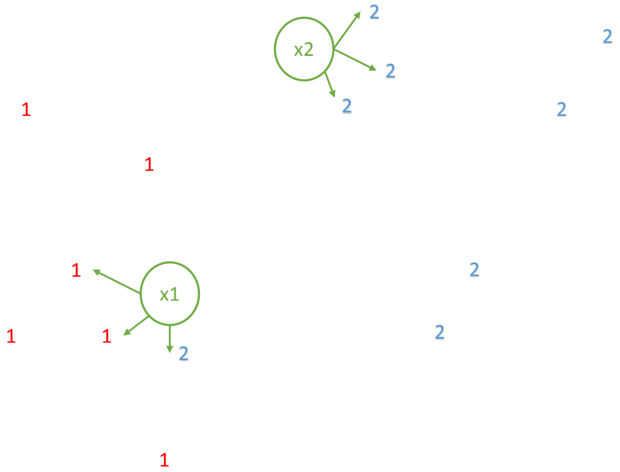
  Note that Y_i is random variable in $[0, 1]$
- The result is random variable with a mean $\mu = d/2$ and standard deviation $\sigma = \sqrt{d/12}$
- The ratio of the variation in the distances to the mean value is referred as *contrast*

$$G(d) = \frac{S_{max} - S_{min}}{\mu} = \sqrt{\frac{12}{d}}$$

k-nearest neighbour (k-NN) classification

- Let N be a labeled set of points belonging to c different classes such that
$$\sum_{i=1}^{c} N_i = N$$
- Classification of a given point x
  - Find k - nearest points to the point x.
  - Assign x the majority label of neighbouring (k-nearest) points

Example



L_p norms

- The real valued function f defined in a vector space V over the subfield F is called a norm if for any $a \in F$ and all $u, v \in V$ it satisfies following three conditions
  - $f(av) = \mid a \mid f(v)$
  - $f(u + v) \leq f(u) + f(v)$
  - $f(v) = 0 \Rightarrow v = 0$
- L_p is defined as follows

$$S(\bar{X}\bar{Y}) = \left(\sum_{i=1}^{d} \mid x_i - y_i \mid^p\right)^{\frac{1}{p}}$$

- In case of $p = 1$ we are dealing with already known to you Manhattan distance. In case of $p = 2$ Euclidean.

Impact of Domain-Specific Relevance

There are cases when some features are more important than the others. Generalized L_p distance is most suitable in such cases.

$$S(x, y) = \Big(\sum_{i=1}^{d} a_i \mid x_i - y_i \mid^p\Big)^{1/p}$$

Computational exercises:

1. Program in "R" your own distance functions: Euclidean, Manhattan, Chebyshev.
2. Minkowsky for different p values.
3. Mahalanobis.
4. Propose your own implementation to replicate figure from slide 20.
5. Propose your own implementation to replicate figures from slide 21.