Data Mining, Lecture 2

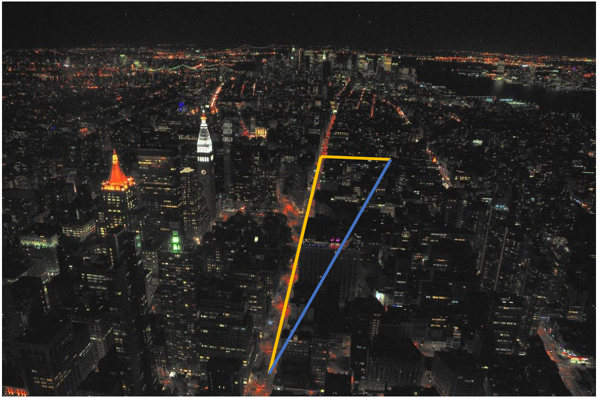
Similarity & Distance

S. Nõmm

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

10.09.2019

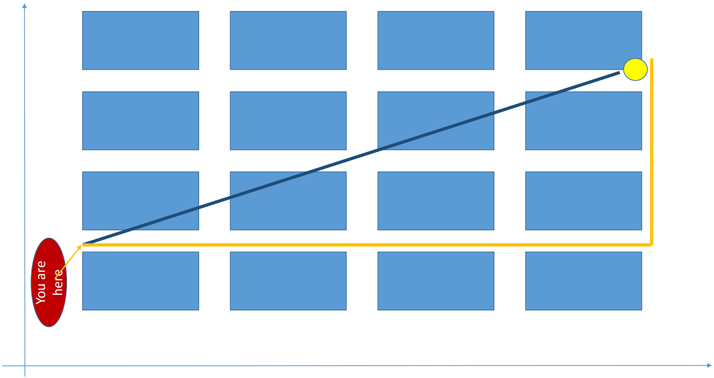
Distance ?



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

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

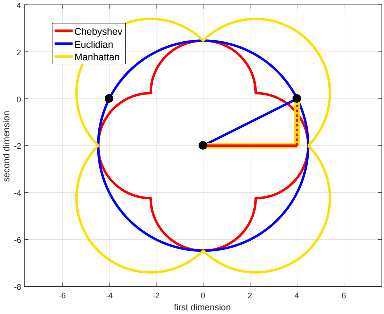
- Manhattan distance also referred as city block distance or taxicab distance

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

- 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

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

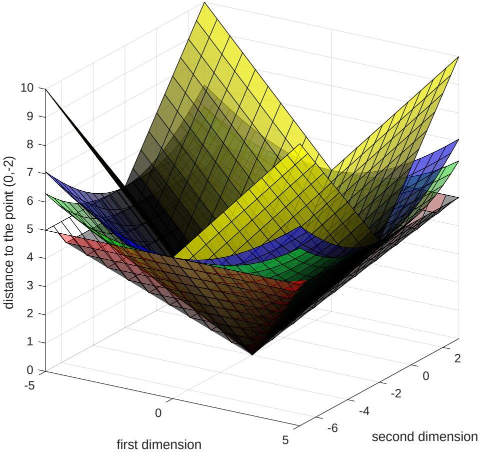
Distance function:Examples 3 Minkowsky distance

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

- $p < 1$ 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.

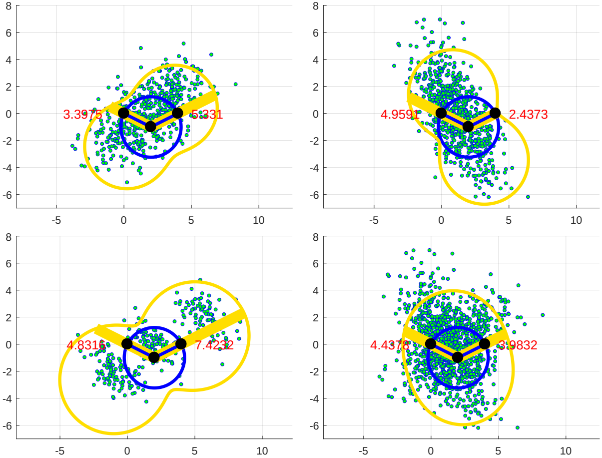


Distance function: Examples 5

Mahalanobis distance

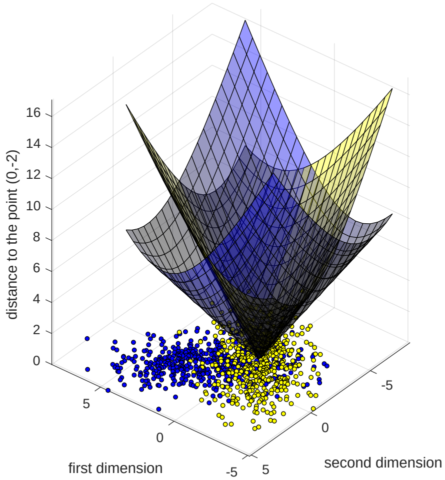
$$S(x, y) = \sqrt{(x - y)^T C^{-1} (x - y)}$$

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*

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

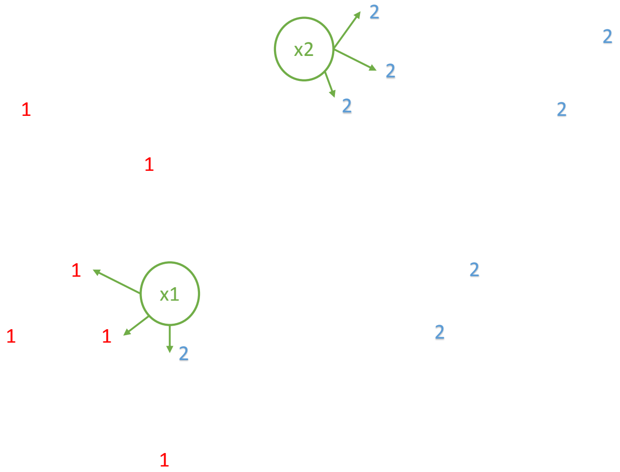
k-nearest neighbour (k-NN) classification

- 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.

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

This distance is frequently referred as *Minkowski distance*