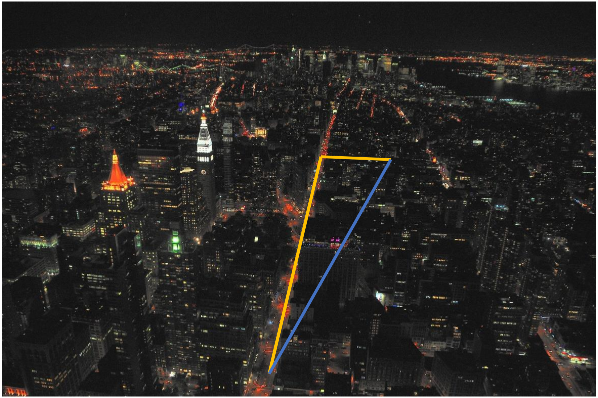
Data Mining, Lecture 2: Distance & Similarity Part I

S. Nõmm

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

September 14, 2017

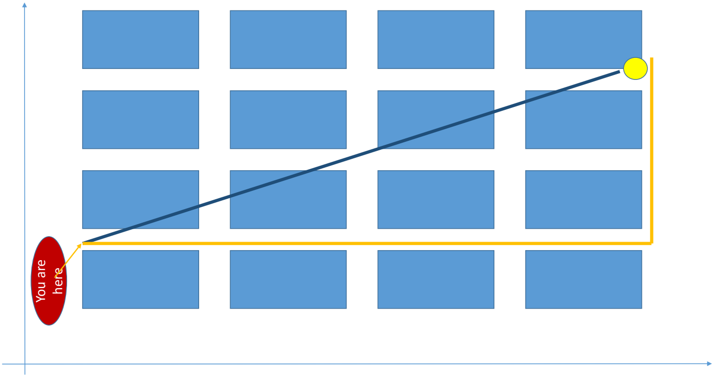
Distance ?



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*

Metric (some times referred as distance function)

Definition

A function $d : X \times X \to \mathbb{R}$ 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 indiscernibles, or coincidence axiom

$$S(x, y) = 0 \Leftrightarrow x = y$$

3. Symmetry

$$S(x, y) = S(y, x)$$

4. Subadditivity or triangle inequality

$$S(x, z) \leq S(x, y) + S(y, z)$$

Examples 1

- 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

$$S(x,y) = \lim_{k\to\infty}\Big(\sum_{i=1}^{n} \mid x_i - y_i \mid^k\Big)^{\frac{1}{k}} = \max_i\big(\mid x_i - y_i \mid\big)$$

Examples 2

- 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.

- 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

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

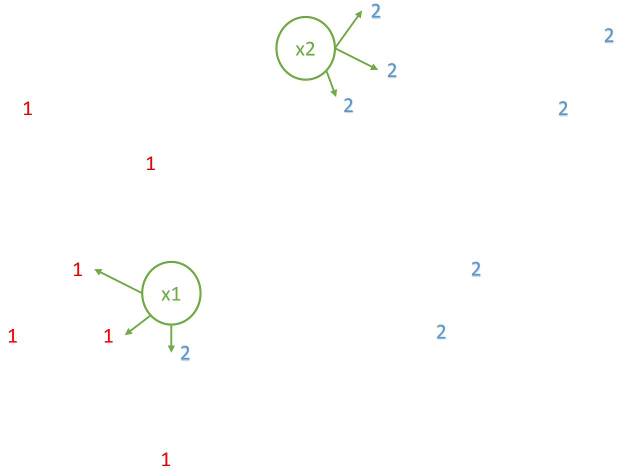
Examples 3: Distances between strings

- 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.

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*

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*

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