Vector Space Model: Distance Functions

(Teaching Inspired by Research)

Prof. Dr. Marcin Grzegorzek and
the Medical Data Science Team

UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MEDIZINISCHE INFORMATIK

1. Introduction
2. Distance Functions for Real Points
3. Distance Functions for Binary Points
4. Distance Functions for Sequences
5. Conclusion

1 Introduction

2 Distance Functions for Real Points

3 Distance Functions for Binary Points

4 Distance Functions for Sequences

5 Conclusion

Contents of the Course

Week	Lecture	Practical Exercises
1	**(05/04)** Introduction to Medical Information Retrieval (MIR)	**(05/04)** Introduction to Python
2	**(12/04)** Main Components and Classification of MIR Systems	**(12/04)** Introduction to Python
3	**(19/04)** Metadata in Medical Information Retrieval Systems	**(19/04)** CBIR in Medical Applications
4	***(26/04)*** *No Lecture due to a Business Trip*	**(26/04)** CBIR in Medical Applications
5	**(03/05)** Set Theoretic Model: Boolean Retrieval	**(03/05)** CBIR in Medical Applications
6	**(10/05)** Set Theoretic Model: Fuzzy Retrieval	**(10/05)** Flask Tutorial
7	**(17/05)** Vector Space Model: Similarity Measures	**(17/05)** Flask Tutorial

Contents of the Course

8	**(24/05)** Vector Space Model: Distance Functions	**(24/05)** HTML
9	**(31/05)** Vector Space Model: Latent Semantic Indexing	**(31/05)** HTML
10	**(07/06)** Probabilistic Model	**(07/06)** HTML
11	**(14/06)** Text-based Retrieval of Medical Information	**(14/06)** Deep Learning
12	**(21/06)** Audio-based Retrieval of Medical Information	**(21/06)** Deep Learning
13	**(28/06)** Image-based Retrieval of Medical Information	**(28/06)** Relevance Feedback
14	**(05/07)** Demonstrators from Current Research Projects	**(05/07)** Relevance Feedback
15	**(12/07)** Summary and Conclusions	**(12/07)** Evaluation

Generally about Distance Functions

A metric on a set $\mathbb{R}^l$ is a distance function

$$d : \mathbb{R}^l \times \mathbb{R}^l \longrightarrow [0, \infty) \quad ,$$

if for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^l$ all of the following conditions are satisfied:

(1) $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0$;

(2) $d(\boldsymbol{x}, \boldsymbol{y}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$;

(3) $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$;

(4) $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$.

1. Introduction
2. Distance Functions for Real Points
3. Distance Functions for Binary Points
4. Distance Functions for Sequences
5. Conclusion

Minkowski Distance – General Form

A popular metric extensively used for IR is the Minkowski distance:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{l} \omega_i |x_i - y_i|^p \right)^{\frac{1}{p}} \quad .$$

Minkowski Distance – Examples

Selected examples of the Minkowski distance for different values of p and $\omega_{i=1,\dots,l} = 1$:

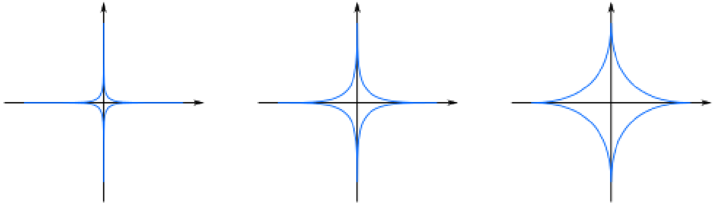
$$p = 1 \quad \Rightarrow \quad \sum_{i=1}^{l} |x_i - y_i|$$

$$\lim_{p \to \infty} \left(\sum_{i=1}^{l} |x_i - y_i|^p \right)^{\frac{1}{p}} = \max_{i=1,\dots,l} |x_i - y_i|$$

$$\lim_{p \to -\infty} \left(\sum_{i=1}^{l} |x_i - y_i|^p \right)^{\frac{1}{p}} = \min_{i=1,\dots,l} |x_i - y_i|$$

Minkowski Distance – Unit Circles

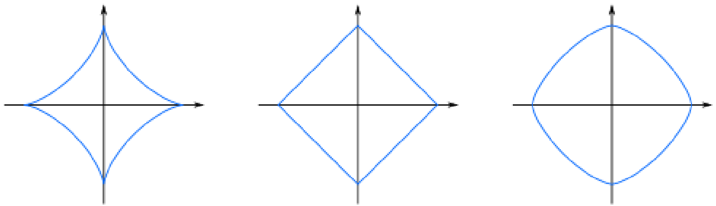
Assuming $\omega_{i=1,\ldots,l} = 1$:



$p = 2^{-2} = 0.25$ $\quad$ $p = 2^{-1.5} = 0.354$ $\quad$ $p = 2^{-1} = 0.5$

Minkowski Distance – Unit Circles

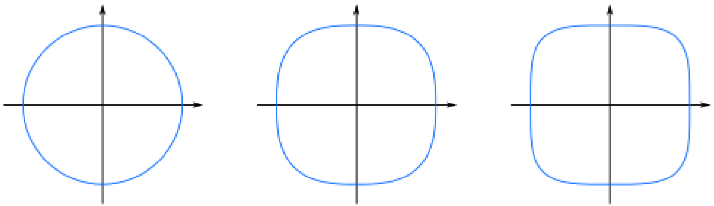
Assuming $\omega_{i=1,\ldots,l} = 1$:



$p = 2^{-0.5} = 0.707$ $\quad$ $p = 2^{0} = 1$ $\quad$ $p = 2^{0.5} = 1.414$

Minkowski Distance – Unit Circles

Assuming $\omega_{i=1,\ldots,l} = 1$:

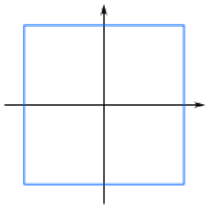


$p = 2^1 = 2$ | $p = 2^{1.5} = 2.828$ | $p = 2^2 = 4$

Minkowski Distance – Unit Circles

Assuming $\omega_{i=1,\ldots,l} = 1$:

$\ldots$



$p = 2^{\infty}$
$= \infty$

1. Introduction
2. Distance Functions for Real Points
3. Distance Functions for Binary Points
4. Distance Functions for Sequences
5. Conclusion

General Information

- Binary points store information about fulfilling or not fulfilling of properties.
- Graphically, such points can be represented as corner of a hypercube.

Comparison of Properties

$p \in P$	p fulfilled for $\boldsymbol{x}_1$	p not fulfilled for $\boldsymbol{x}_1$
p fulfilled for $\boldsymbol{x}_2$	$n_{1/1}$	$n_{0/1}$
p not fulfilled for $\boldsymbol{x}_2$	$n_{1/0}$	$n_{0/0}$

Example:

$$\boldsymbol{x}_1 = (0,0,0,0,1,1,1,1)^{\mathrm{T}} \quad \boldsymbol{x}_2 = (1,1,0,1,1,1,0,0)^{\mathrm{T}}$$

$$\Downarrow$$

$$n_{0/0} = 1 \quad n_{0/1} = 3 \quad n_{1/0} = 2 \quad n_{1/1} = 2$$

Minkowski Distance for Binary Points

General form assuming $\omega_{i=1,\ldots,l} = 1$:

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \left(\sum_{i=1}^{l} |x_{1,i} - x_{2,i}|^p \right)^{\frac{1}{p}} \quad .$$

For binary points:

$$d = (n_{1/0} + n_{0/1})^{1/p}$$

1 Introduction

2 Distance Functions for Real Points

3 Distance Functions for Binary Points

4 Distance Functions for Sequences

5 Conclusion

General Information

- A sequence is just a list of data elements of the same type.
- The number of data elements describing a document can vary.

Earth Mover's Distance – Data Format

- The following data type is considered here:
  $\mathrm{tuple}(\boldsymbol{x}_i \;:\; \mathrm{array}[1 \ldots l](\mathrm{real}) \;;\; w_{\boldsymbol{x}_i} \;:\; \mathrm{real})$.
- It describes the i-th element of the sequence $\boldsymbol{X}$.

Earth Mover's Distance – Story Behind

- Computing the distance between the sequence $\boldsymbol{X}$ with m elements and the sequence $\boldsymbol{Y}$ with n elements, we consider the elements of $\boldsymbol{X}$ to be mounds and the elements of $\boldsymbol{Y}$ to be holes in the ground.
- The points $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ can be interpreted as the positions of the mounds and the holes in a l-dimensional space.
- The volumes of the mounds/holes are given by $w_{\boldsymbol{x}_i}$ and $w_{\boldsymbol{y}_j}$ respectively.
- The distance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined as the minimum cost of transporting the earth from the mounds to the holes.

Earth Mover's Distance – Minimisation of Costs

- Thus, the goal is to minimise the transportation costs.
- A particular constellation of the whole earth transportation process between mounds and holes can be described by a matrix $F = [f_{ij}]$ with f_{ij} standing for the earth volume moved from the mound $\boldsymbol{x}_i$ into the hole $\boldsymbol{y}_j$.
- The overall transportation costs can be now computed as follows:

$$T_{\mathrm{costs}}(\boldsymbol{X}, \boldsymbol{Y}, F) = \sum_{i=1}^{m} \sum_{j=1}^{n} d(\boldsymbol{x}_i, \boldsymbol{y}_j) f_{ij} \qquad .$$

Earth Mover's Distance – Final Result

The final distance value:

$$d_{\text{EMD}}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\min\limits_{|f_{ij}|} \left(\sum\limits_{i=1}^{m} \sum\limits_{j=1}^{n} d(\boldsymbol{x}_i, \boldsymbol{y}_j) f_{ij} \right)}{\sum\limits_{i=1}^{m} \sum\limits_{j=1}^{n} f_{ij}} .$$

Earth Mover’s Distance

1. Introduction
2. Distance Functions for Real Points
3. Distance Functions for Binary Points
4. Distance Functions for Sequences
5. Conclusion

Final Statements

- Most frequently, documents to be compared with each other for retrieval are described by feature vectors of the same dimensionality. In this case, well-known distance functions defined for real points, e.g., the Minkowski distance, can be applied.

- In more heterogeneous IR scenarios, the documents are represented by data structures with less consistency. In this case other methods (e.g., distance functions for sequences) are necessary for their comparison.