Data Mining Tutorial

Session 3: Distance functions homework

Erich Schubert, Eirini Ntoutsi

Ludwig-Maximilians-Universität München

2012-05-24 — KDD class tutorial

Distance functions

- Reflexive: $x = y \quad \Rightarrow \quad d(x, y) = 0$
  "Distance to self is 0"
- Symmetric: $d(x, y) = a \quad \Leftrightarrow \quad d(y, x) = a$
  "Order of arguments is irrelevant"
- Strict: $d(x, y) = 0 \quad \Rightarrow \quad x = y$
  "Only identical elements have distance 0"
- Triangle inequality: $d(x, y) \leq d(x, z) + d(z, y)$
  "Directly x to y is at least as short as a detour over z"

You will need to know these properties!

An important reminder

You cannot prove by example.

An important reminder

E. Schubert, E. Ntoutsi

You cannot prove by example.

… but you can disprove by example!

An important reminder

You cannot prove by example.

... but you can disprove by example!

Please, show that it holds for *all* situations, or give a counterexample. Do not give a positive example.

Distances Homework – difference sum

E. Schubert,
E. Ntoutsi

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

Distances Homework – difference sum

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

$d((0), (-1)) = -1$ – must not be negative!

Distances Homework – squared Euclidean

E. Schubert, E. Ntoutsi

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

Distances Homework – squared Euclidean

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

Reflexive, symmetric, strict: obvious. Triangle inequality?

Distances Homework – squared Euclidean

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

Reflexive, symmetric, strict: obvious. Triangle inequality?

How about $o = (0,0)$, $p = (1,0)$, $q = (2,0)$?

$$d(o,q) = 4 \qquad d(o,p) + d(p,q) = 1 + 1 = 2$$

Distances Homework – squared Euclidean

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

Reflexive, symmetric, strict: obvious. Triangle inequality?

How about $o = (0,0)$, $p = (1,0)$, $q = (2,0)$?

$$d(o,q) = 4 \quad \not\leq \quad d(o,p) + d(p,q) = 1 + 1 = 2$$

"Squared Euclidean distance" – not metrical.
(1 dimensional counter example: $0, 1, 2$)

Distances Homework – projected Euclidean

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

Distances Homework – projected Euclidean

E. Schubert,
E. Ntoutsi

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

Reflexive, symmetric: obvious. Triangle inequality requires some work.

Distances Homework – projected Euclidean

E. Schubert, E. Ntoutsi

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

Reflexive, symmetric: obvious. Triangle inequality requires some work.

But not strict: dimension n is ignored.

Distances Homework – concordance

$$d(x,y) = \sum_{i=1}^{n} \begin{cases} 1 & \text{iff} \quad x_i = y_i \\ 0 & \text{iff} \quad x_i \neq y_i \end{cases}$$

Distances Homework – concordance

$$d(x,y) = \sum_{i=1}^{n} \begin{cases} 1 & \text{iff} \quad x_i = y_i \\ 0 & \text{iff} \quad x_i \neq y_i \end{cases}$$

d is not reflexive – the other properties are irrelevant to us.

Distances Homework – Hamming distance

$$d(x,y) = \sum_{i=1}^{n} \begin{cases} 1 & \text{iff} \quad x_i \neq y_i \\ 0 & \text{iff} \quad x_i = y_i \end{cases}$$

Distances Homework – Hamming distance

$$d(x,y) = \sum_{i=1}^{n} \begin{cases} 1 & \text{iff} \quad x_i \neq y_i \\ 0 & \text{iff} \quad x_i = y_i \end{cases}$$

Discordance on binary vectors.
"Number of ones after an XOR of the two vectors".
Important metric from information theory.
Reflexivity, strictness, symmetry are obvious.

Distances Homework – Hamming distance

Proof of triangle inequality by case distinction on the individual positions (dimensions):

Distances Homework – Hamming distance

Proof of triangle inequality by case distinction on the individual positions (dimensions):

A) $x_i = y_i \land y_i = z_i$:

$$\begin{aligned} d(x_i, y_i) + d(y_i, z_i) &\geq d(x_i, z_i) \\ d(x_i, x_i) + d(y_i, x_i) &\geq d(x_i, x_i) \\ 0 + 0 &\geq 0 \end{aligned}$$

Proof of triangle inequality by case distinction on the individual positions (dimensions):

B) $x_i = y_i \wedge x_i \neq z_i$:

$$\begin{aligned} d(x_i, y_i) + d(y_i, z_i) &\geq d(x_i, z_i) \\ d(x_i, x_i) + d(x_i, z_i) &\geq d(x_i, z_i) \\ 0 + 1 &\geq 1 \end{aligned}$$

Distances Homework – Hamming distance

Proof of triangle inequality by case distinction on the individual positions (dimensions):

C) $x_i = z_i \wedge x_i \neq y_i$:

$$\begin{aligned} d(x_i, y_i) + d(y_i, z_i) &\geq d(x_i, z_i) \\ d(x_i, y_i) + d(y_i, x_i) &\geq d(x_i, x_i) \\ 1 + 1 &\geq 0 \end{aligned}$$

Distances Homework – Hamming distance

Proof of triangle inequality by case distinction on the individual positions (dimensions):

D) $x_i \neq y_i \wedge y_i = z_i$:

$$
\begin{aligned}
d(x_i, y_i) + d(y_i, z_i) &\geq d(x_i, z_i) \\
d(x_i, y_i) + d(y_i, y_i) &\geq d(x_i, y_i) \\
1 + 0 &\geq 1
\end{aligned}
$$

Proof of triangle inequality by case distinction on the individual positions (dimensions):
E) $x_i \neq y_i \wedge y_i \neq z_i \wedge x_i \neq z_i$:

$$\begin{aligned} d(x_i, y_i) + d(y_i, z_i) &\geq d(x_i, z_i) \\ 1 + 1 &\geq 1 \end{aligned}$$

Distances Homework – Hamming distance

Proof of triangle inequality by case distinction on the individual positions (dimensions):
Which implies:

$$
\begin{aligned}
d(x,y) + d(y,z) &= \sum_i^n d(x_i,y_i) + \sum_i^n d(y_i,z_i) \\
&= \sum_i^n \left(d(x_i,y_i) + d(y_i,z_i)\right) \\
&\geq \sum_i^n d(x_i,z_i) = d(x,z)
\end{aligned}
$$

(We have just shown the step line 2 to 3!)

Distances Homework – Other examples

E. Schubert, E. Ntoutsi

Other interesting distance functions (on sets $X, Y \subseteq \mathbb{R}^n$), for existing distance measures $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_0^+$:

- single-link$(X, Y) = \min_{x \in X, y \in Y} d(x, y)$
- average-link$(X, Y) = \frac{1}{|X| \cdot |Y|} \cdot \sum_{x \in X, y \in Y} d(x, y)$
- complete-link$(X, Y) = \max_{x \in X, y \in Y} d(x, y)$

They will be discussed in detail in the clustering chapter!

Distances Homework – Other examples

There are hundreds of distance functions.

- For time series: DTW, EDR, ERP, LCSS, ...
- For text: Cosine and normalizations
- For sets – based on intersection, union, ...
- For clusters (single-link etc.)
- For histograms: histogram intersection, "Earth movers distance", quadratic forms with color similarity
- With normalization: Canberra, ...
- Quadratic forms / bilinear forms: $d(x, y) := x^T M y$ for some positive (usually symmetric) definite matrix M.

Can be seen as a part of "preprocessing": choosing the appropriate distance function!

Induced metric

Given a pseudo-metric d on the set A: $d : A \times A \to \mathbb{R}_0^+$.

Define the equivalence relation $\sim$ such that
$x \sim y \Leftrightarrow d(x, y) = 0$.

Let $A^\sim$ be the set of equivalence classes of A wrt. $\sim$.

$$d^\sim : A^\sim \times A^\sim \to \mathbb{R}_0^+$$
$$\text{with } d^\sim(x^\sim, y^\sim) := d(x, y)$$

Properties?

Induced metric

Given a pseudo-metric d on the set A: $d : A \times A \to \mathbb{R}_0^+$.

Define the equivalence relation $\sim$ such that
$x \sim y \Leftrightarrow d(x, y) = 0$.

Let $A^\sim$ be the set of equivalence classes of A wrt. $\sim$.

$$d^\sim : A^\sim \times A^\sim \to \mathbb{R}_0^+$$
$$\text{with } d^\sim(x^\sim, y^\sim) := d(x, y)$$

Properties? Well defined?

Induced metric

Verify for any $z \in x^{\sim}$ and $w \in y^{\sim}$ that $d(z, w) = d(x, y)$.

Induced metric

Verify for any $z \in x^\sim$ and $w \in y^\sim$ that $d(z,w) = d(x,y)$.

Since $z \in x^\sim$ and $w \in y^\sim$ we have

$$
\begin{aligned}
z^\sim = x^\sim \text{ and } d(z,x) = 0\\
w^\sim = y^\sim \text{ and } d(w,y) = 0
\end{aligned}
$$

Induced metric

Verify for any $z \in x^\sim$ and $w \in y^\sim$ that $d(z,w) = d(x,y)$.

Since $z \in x^\sim$ and $w \in y^\sim$ we have

$$z^\sim = x^\sim \text{ and } d(z,x) = 0$$
$$w^\sim = y^\sim \text{ and } d(w,y) = 0$$

Use the triangle inequality twice:

$$d(z,w) \leq d(z,x) + d(x,y) + d(y,w) \leq d(x,y)$$

$$d(x,y) \leq d(x,z) + d(z,w) + d(w,y) \leq d(z,w)$$

Any element from the equivalence class gives the same distance for $d^\sim$. $\Rightarrow$ well defined on $A^\sim$.

Induced metric

Need to show:
$d^{\sim}(a^{\sim}, b^{\sim}) = 0 \Leftrightarrow a^{\sim} = b^{\sim}$

Induced metric

Need to show:

$$d^{\sim}(a^{\sim}, b^{\sim}) = 0 \Leftrightarrow a^{\sim} = b^{\sim}$$

$$d^{\sim}(a^{\sim}, b^{\sim}) = 0$$

$$\Rightarrow d(a, b) = 0$$

$$\Rightarrow a \sim b$$

$$\Rightarrow a^{\sim} = b^{\sim}$$

Induced metric

Need to show:

$d^{\sim}(a^{\sim}, b^{\sim}) = 0 \Leftrightarrow a^{\sim} = b^{\sim}$

$d^{\sim}(a^{\sim}, b^{\sim}) = 0$

$\Rightarrow d(a, b) = 0$

$\Rightarrow a \sim b$

$\Rightarrow a^{\sim} = b^{\sim}$

Symmetry, triangle inequality inherited from d!

Induced metric

$$\text{euclid}_{xy}((r_1, x_1, y_1), (r_2, x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Record number	x	y	Record number	x	y
1	0	1	4	1	1
2	1	1	5	2	2
3	0	1	6	3	3

Induced metric

$$\text{euclid}_{xy}((r_1, x_1, y_1), (r_2, x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Record number	x	y
1	0	1
2	1	1
3	0	1

Record number	x	y
4	1	1
5	2	2
6	3	3

Euclidean distance on $X \times Y$. Metric in $\mathbb{R}^2 \sim X \times Y$, but only a Pseudo-metric on Record number $\times X \times Y$. "Duplicate" records have a distance of 0.