Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Eirini Ntoutsi
Erich Schubert

Knowledge Discovery in Databases

SS 2012

Übungsblatt 3: Distance functions and Evaluation of Classifiers

Aufgabe 3-1 *Distance functions*

Distance functions can be classified into the following categories:

$d : S \times S \to \mathbb{R}_0^+$ $x, y, z \in S$:	reflexive reflexiv $x = y \Rightarrow d(x, y) = 0$	symmetric symmetrisch $d(x, y) = d(y, x)$	strict strikt $d(x, y) = 0 \Rightarrow x = y$	Triangle inequality Dreiecksungleichung $d(x, z) \leq d(x, y) + d(y, z)$
Dissimilarity function Unähnlichkeitsfunktion	×			
(Symmetric) Pre-metric (Symmetrische) Prämetrik	×	×		
Semi-metric, Ultra-metric Semimetrik, Ultrametrik	×	×	×	
Pseudo-metric Pseudometrik	×	×		×
Metric Metrik	×	×	×	×

So if a distance measure satisfies $d : S \times S \to \mathbb{R}_0^+$ and for any vector $x, y, z \in S$: is reflexive, symmetric and strict and also satisfies the triangle inequality, then it is a metric.

As you can see, a pre-metric does not necessarily need to be *strictly* reflexive. Make sure you understand the difference between reflexivity and strictness!

Note: these terms as well as “distance function” are used inconsistently in literature. In mathematics, “distance function” is commonly used synonymous with “metric”. In a database (and thus data mining) context, strictness is often not relevant at all, and a “distance function” usually refers to a pseudo-metric, pre-metric or even dissimilarity function. Do not rely on Wikipedia, it uses multiple definitions within itself!

Decide for each of the following functions $d(\mathbb{R}^n, \mathbb{R}^n)$, whether they are a distance, and if so of which type.

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

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

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

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

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

Aufgabe 3-2 *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. $\sim$.

- Which properties does the distance function $d^\sim : A^\sim \times A^\sim \to \mathbb{R}_0^+$ with $d^\sim(x^\sim, y^\sim) := d(x, y)$ have?
- Given a database similar to this, what properties does the following distance function have?

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

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

Explain which records are considered equivalent by this distance function, and discuss whether this is sensible in a database and data mining context to have pseudo-metric distance functions.

Aufgabe 3-3 *Evaluation of classifiers*

Given a data set with known class labels of the objects. In order to evaluate the quality of a classifier K, each object is additionally classified using K. The results are given in the table (all three columns) below.

ID	Object class	$K(o)$
O_1	A	A
O_2	B	A
O_3	A	C
O_4	C	C
O_5	C	B

ID	Object class	$K(o)$
O_6	B	B
O_7	A	A
O_8	A	A
O_9	A	A
O_{10}	B	C

ID	Object class	$K(o)$
O_{11}	B	A
O_{12}	C	A
O_{13}	C	C
O_{14}	C	C
O_{15}	B	B

- Using the table (all three columns) above, compute precision and recall for each class.
- To get a complete measure for the quality of the classification with respect to a single class, the F_1-measure (the harmonic mean of precision and recall) is commonly used. It is defined as follows:

$$F_1(K, i) = \frac{2 \cdot \text{Recall}(K, i) \cdot \text{Precision}(K, i)}{\text{Recall}(K, i) + \text{Precision}(K, i)}$$

  Compute the F_1-measure for all classes.

  Note: "F_1-measure" may refer to the same formula but computed using a different precision and different recall in other applications. It is a specialization of F_β with equal weighting of precision and recall.

- So far, the F_1-measure is only defined for classes and not yet useful to get an overview of the overall performance of the classifiers. For this, one commonly takes the average over all classes using one of the following two approaches:
  - Micro Average F_1-Measure: The values of TP, FP and FN are added up over all classes. Then precision, recall and F_1-measure are computed using these sums.
  - Macro Average F_1-Measure: Precision and recall are computed for each class individually, afterwards the average precision and average recall are used to compute the F_1-measure.

  Compute the Micro- and Macro-Average F_1-measures for the example above. What do you observe?