University of Southern Denmark
IMADA
Arthur Zimek

DM868/DM870/DS804: Data Mining and Machine Learning

Spring term 2023

Exercise 4: Distance Measures, k-means

Exercise 4-1 Color-histograms and distance functions (1 point)

As a warm-up on distance measures: For each of the following distance measures (Euclidean, Manhattan, maximum, weighted Euclidean, quadratic form)

$$\begin{aligned}
\mathrm{dist}_2(p,q) &= \left(|p_1-q_1|^2+|p_2-q_2|^2+|p_3-q_3|^2\right)^{\frac{1}{2}}\\
\mathrm{dist}_1(p,q) &= |p_1-q_1|+|p_2-q_2|+|p_3-q_3|\\
\mathrm{dist}_\infty(p,q) &= \max(|p_1-q_1|,|p_2-q_2|,|p_3-q_3|)\\
\mathrm{dist}_w(p,q) &= \left(w_1|p_1-q_1|^2+w_2|p_2-q_2|^2+w_3|p_3-q_3|^2\right)^{\frac{1}{2}}\\
\mathrm{dist}_M(p,q) &= \left((p-q)M(p-q)^{\mathrm{T}}\right)^{\frac{1}{2}}
\end{aligned}$$

calculate the distance between $p=(2,3,5)$ and $q=(4,7,8)$. As w use $(1,1.5,2.5)$ and as M use both of the following:

$$M_1=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}\qquad M_2=\begin{pmatrix}1&0.9&0.7\\0.9&1&0.8\\0.7&0.8&1\end{pmatrix}$$

Given 5 pictures as in Figure 1 with 36 pixels each.

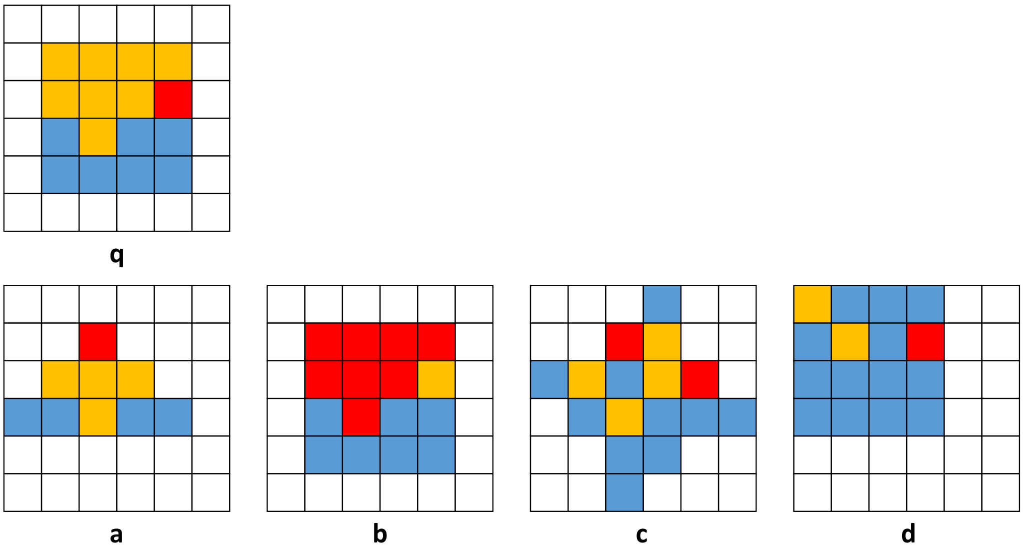

Figure 1: 6 × 6 pixel pictures

(a) Extract from each picture a color histogram with the bins *red*, *orange*, and *blue* (the white pixels are ignored).

(b) Which pictures are most similar to the query q, using Euclidean distance? Give a ranking according to similarity to q.

(c) The results are not entirely satisfactory. What could you change in the feature extraction or in the distance function to get better results? Report the improved feature extraction and features or the improved distance function.

Exercise 4-2 Distance functions (1 point)

Distance functions can be classified into the following categories:

$d : S \times S \to \mathbb{R}_0^+$	reflexive	symmetric	strict	triangle inequality
$x, y, z \in S$:	$x = y \Rightarrow d(x, y) = 0$	$d(x, y) = d(y, x)$	$d(x, y) = 0 \Rightarrow x = y$	$d(x, z) \leq d(x, y) + d(y, z)$
Dissimilarity function	×			
(Symmetric) Pre-metric	×	×		
Semi-metric, Ultra-metric	×	×	×	
Pseudo-metric	×	×		×
Metric	×	×	×	×

So if a distance measure satisfies $d : S \times S \to \mathbb{R}_0^+$ and $\forall x, y, z \in S$ it is reflexive, symmetric, and strict and it 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 the literature. In mathematics, "distance function" is commonly used synonymously with "metric". In a database and data mining context, strictness is often not relevant at all, and a "distance function" usually refers to a pseudo-metric, pre-metric, or even just to some 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, 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) $d(x, y) = \sum_{i=1}^{n} \begin{cases} 1 & \text{iff} \quad x_i \neq y_i \\ 0 & \text{iff} \quad x_i = y_i \end{cases}$

Exercise 4-3 Distances on a database (1 point)

Given a database similar to this one:

r	x	y
1	0	1
2	1	1
3	0	1

r	x	y
4	1	1
5	2	2
6	3	3

Which properties does the following distance function have?

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

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

Hint: What could be the nature of attribute r in a database context?

Exercise 4-4 Visualization of distance functions (1 point)

Brainstorm on how you could visualize the behavior of distance measures, and how you could implement a visualization tool. If you got some ideas: well, go and implement them!

The point is credited for demonstrating a tool that you implemented yourself.

Exercise 4-5 k-means 1-dimensional Example (1 point)

Given are the following 1-dimensional points: $\{2, 3, 4, 10, 11, 12, 20, 25, 30\}$. We set $k = 3$ and choose as initial means: $\mu_1 = 2$, $\mu_2 = 4$, and $\mu_3 = 6$.

Compute the new clusters after each iteration of k-means (Lloyd/Forgy) until convergence.