University of Southern Denmark
IMADA
Rolf Fagerberg
Arthur Zimek

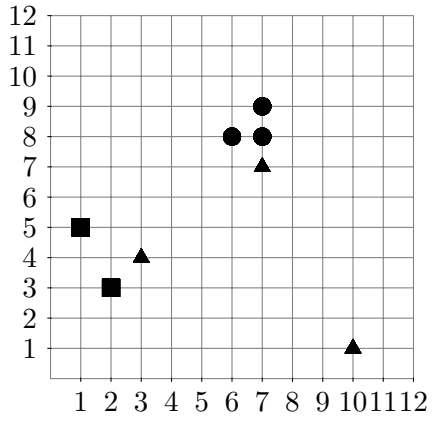
DM 534: Introduction to Computer Science

Autumn term 2016

Exercise 44: Clustering, Color Histograms

Exercise 44-1 k-means, choice of k, and compactness

Given the following data set with 8 objects (in $\mathbb{R}^2$) as in the lecture:



Compute a complete partitioning of the data set into $k = 3$ clusters using the basic k-means algorithm (due to Forgy and Lloyd). The initial assignment of objects to clusters is given using the triangle, square, and circle markers.

Objects x are assigned to the cluster with the least increase in squared deviations $SSQ(x, c)$ where c is the cluster center.

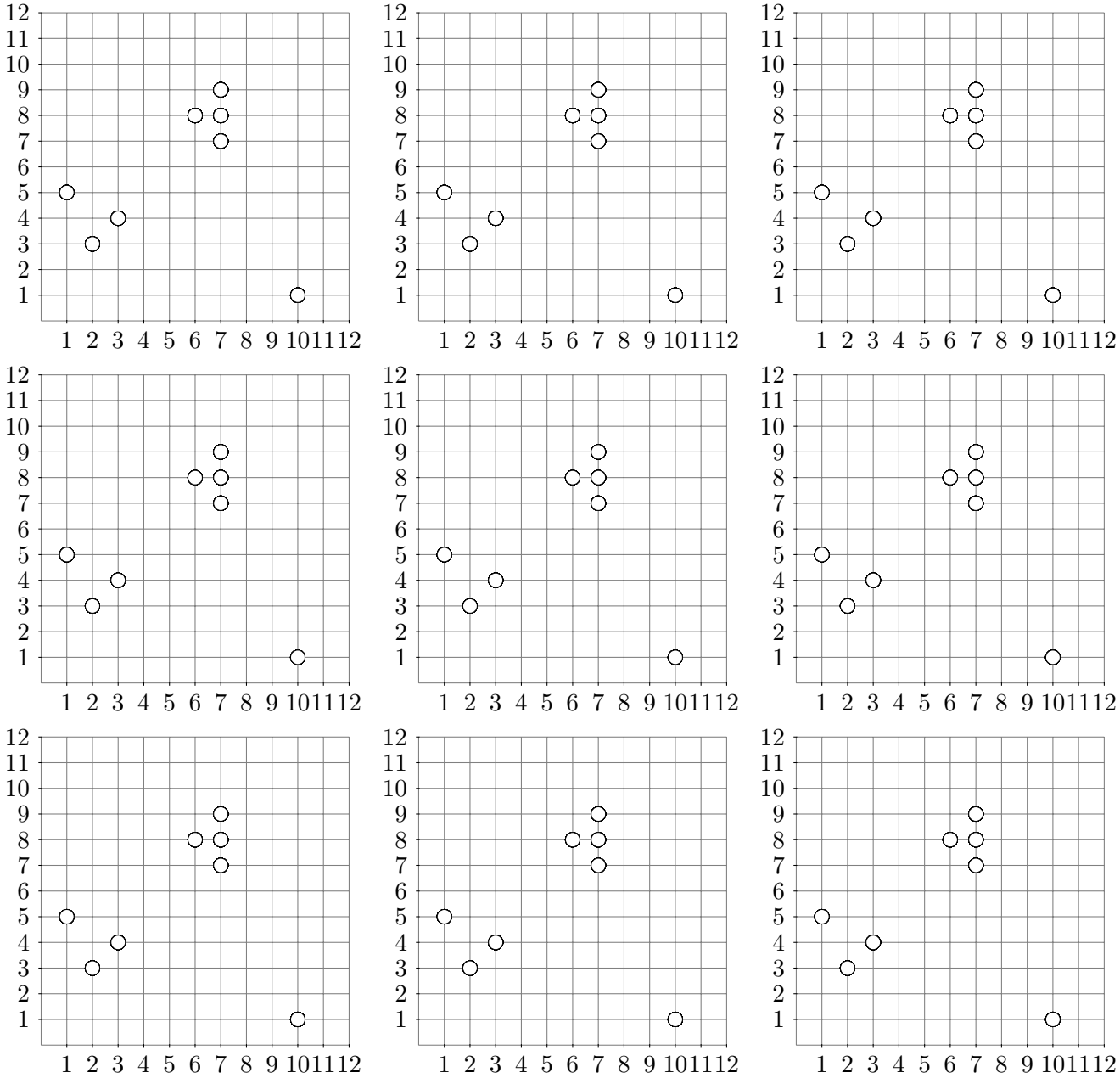
$$SSQ(x, c) = \sum_{i=1}^{d} |x_i - c_i|^2$$

Start with computing the initial centroids, and draw the cluster assignments after each step and explain the step. Remember to use the least squares assignment!

You can use the data set sketches on the next page.

Give the final quality of the clustering (TD^2). How does it compare with the solutions for $k = 2$ discussed in the lecture? Can we conclude on $k = 3$ or $k = 2$ being the better parameter choice on this data set?

Also compute solutions with $k = 4$, $k = 5$, starting from some random initial assignments of objects to clusters. What do you observe in terms of the TD^2 measure?



Exercise 44-2 Color-Histograms and Distancefunctions

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.

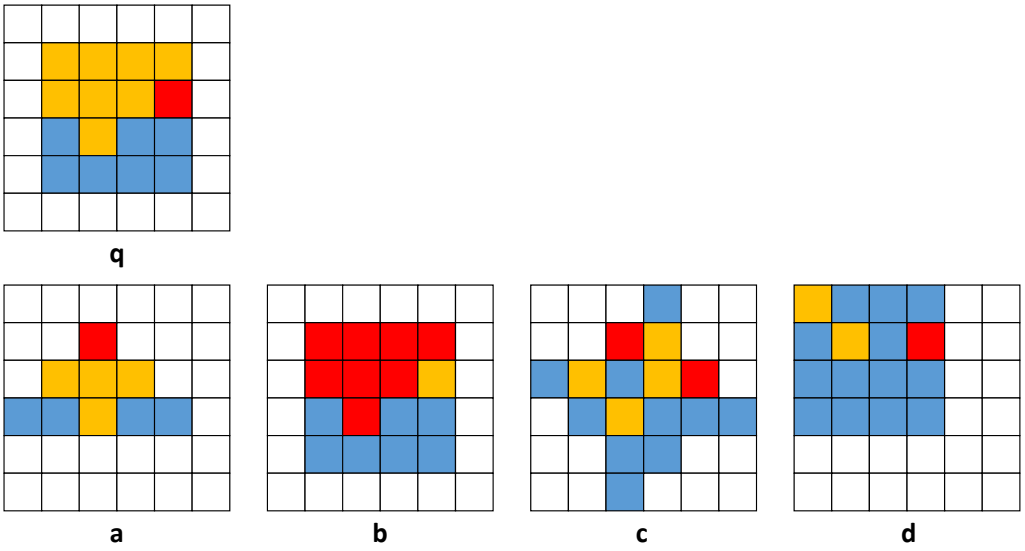


Figure 1: 6×6 pixel pictures

(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.