University of Southern Denmark
IMADA
Rolf Fagerberg
Richard Roettger based on the work of Arthur Zimek

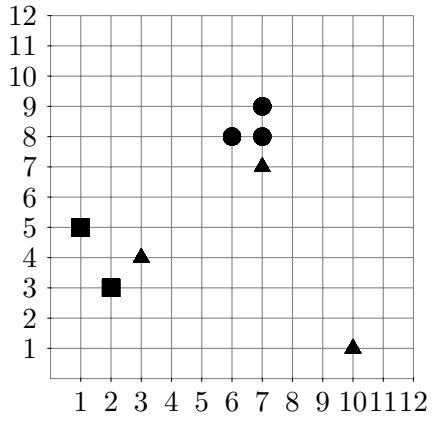
DM534: Introduction to Computer Science

Autumn term 2018

Exercise Clustering: Clustering, Color Histograms

Exercise Clustering-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.

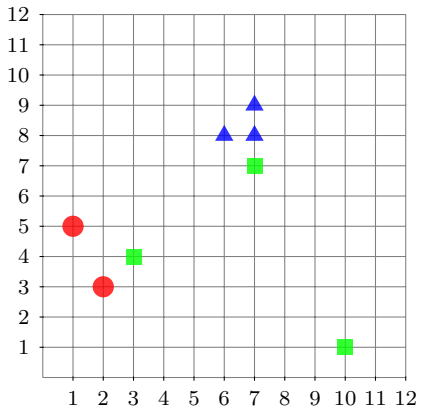
$$SSQ(x, c) = \sum_{i=1}^{d} |x_i - c_i|^2$$

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!

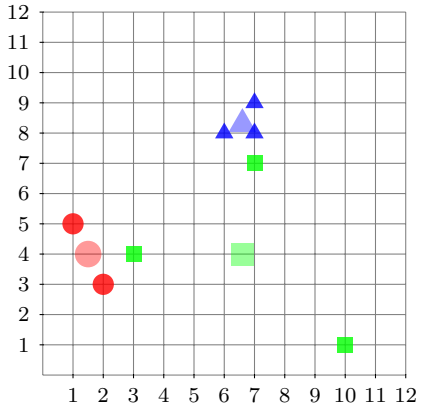
You can use the data set sketches on the next page.

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?

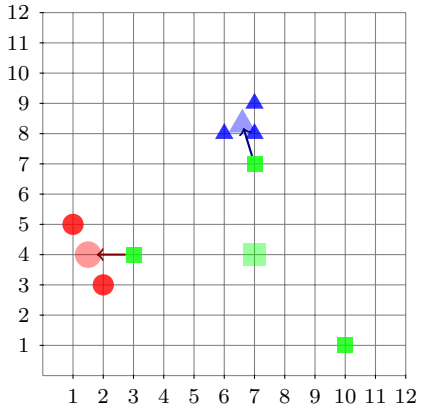


Initial clusters.

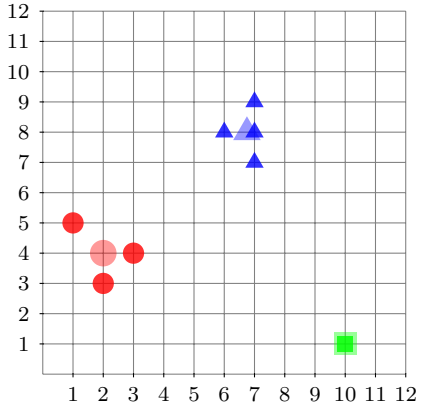


Compute centroids:

- $\mu = (1.5, 4)$
- $\mu \approx (6.6, 8.3)$
- $\mu \approx (6.6, 4)$



Reassign points to closest representant.



Recompute centroids:

- $\mu = (2, 4)$
- $\mu = (6.75, 8)$
- $\mu = (10, 1)$

Then reassignment of points: no change.
Algorithm terminates.

$SSQ(\mu_1, p_1) = |10 - 10|^2 + |1 - 1|^2 = 0$
$TD^2(C_1) = 0$

$SSQ(\mu_1, p_2) = |2 - 2|^2 + |4 - 3|^2 = 0 + 1 = 1$
$SSQ(\mu_1, p_3) = |2 - 3|^2 + |4 - 4|^2 = 1 + 0 = 1$
$SSQ(\mu_1, p_4) = |2 - 1|^2 + |4 - 5|^2 = 1 + 1 = 2$
$TD^2(C_2) = 4$

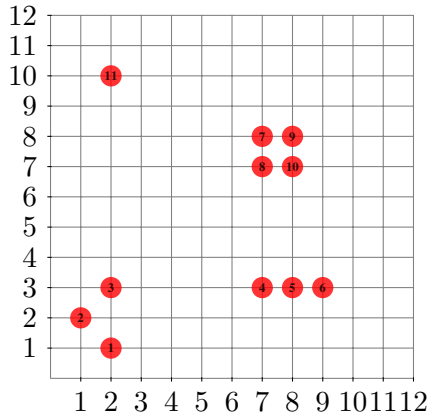
$SSQ(\mu_2, p_5) = |6.75 - 7|^2 + |8 - 7|^2 = \frac{1}{16} + 1 = 1\frac{1}{16}$
$SSQ(\mu_2, p_6) = |6.75 - 6|^2 + |8 - 8|^2 = \frac{9}{16} + 0 = \frac{9}{16}$
$SSQ(\mu_2, p_7) = |6.75 - 7|^2 + |8 - 8|^2 = \frac{1}{16} + 0 = \frac{1}{16}$
$SSQ(\mu_2, p_8) = |6.75 - 7|^2 + |8 - 9|^2 = \frac{1}{16} + 1 = 1\frac{1}{16}$
$TD^2(C_3) = 2\frac{3}{4}$ $TD^2 = 6\frac{3}{4}$

In terms of the compactness measure TD^2, this solution with $k = 3$ is much better than any solution with $k = 2$.

However, if we increase k further, the compactness will be even smaller. With $k = 8$, we could get a solution with $TD^2 = 0$, because each point will be identical with its cluster. Optimizing compactness alone is therefore not good enough to find the optimal number of clusters.

Exercise Clustering-2 Furthest First Initialization

Given the following data set with 11 objects (in $\mathbb{R}^2$):



Aim is now to perform a furthest-first initialization as seen in the lecture.

You should use the following distance measures in order to measure the distance between two points $p = (p_1, p_2)$ and $q = (q_1, q_2)$.

$$\begin{aligned} \text{dist}_2(p,q) &= \left(|p_1 - q_1|^2 + |p_2 - q_2|^2\right)^{\frac{1}{2}} \\ \text{dist}_1(p,q) &= |p_1 - q_1| + |p_2 - q_2| \\ \text{dist}_\infty(p,q) &= \max(|p_1 - q_1|, |p_2 - q_2|) \end{aligned}$$

It might help to fill out the similarity matrix noting all pair-wise distances between all points (note: only the upper triangle is required since the distance functions are symmetric). You find table sketches on the next page.

Let us choose point 3 as our first center. Define the next 3 centers according to the three different norms. (In case two or more points have the same distance, choose the point with the lower point number). Does the sequence of points differ between the norms?

L_2 norm:

	1	2	3	4	5	6	7	8	9	10	11
1		1.41	2.00	5.39	6.32	7.28	8.60	7.81	9.22	8.49	9.00
2			1.41	6.08	7.07	8.06	8.49	7.81	9.22	8.60	8.06
3				5.00	6.00	7.00	7.07	6.40	7.81	7.21	7.00
4					1.00	2.00	5.00	4.00	5.10	4.12	8.60
5						1.00	5.10	4.12	5.00	4.00	9.22
6							5.38	4.47	5.10	4.12	9.90
7								1.00	1.00	1.41	5.39
8									1.41	1.00	5.83
9										1.00	6.32
10											6.71
11											

L_1 norm:

	1	2	3	4	5	6	7	8	9	10	11
1		2	2	7	8	9	12	11	13	12	9
2			2	7	8	9	12	11	13	12	9
3				5	6	7	10	9	11	10	7
4					1	2	5	4	6	5	12
5						1	6	5	5	4	13
6							7	6	6	5	14
7								1	1	2	7
8									2	1	8
9										1	8
10											9
11											

L_∞ norm:

	1	2	3	4	5	6	7	8	9	10	11
1		1	2	5	6	7	7	6	7	6	9
2			1	6	7	8	6	6	7	7	8
3				5	6	7	5	5	6	6	7
4					1	2	5	4	5	4	7
5						1	5	4	5	4	7
6							5	4	5	4	7
7								1	1	1	5
8									1	1	5
9										1	6
10											6
11											

First points using L_2 norm: p3, p9, p11, p6
First points using L_1 norm: p3, p9, p11, p6
First points using L_∞ norm: p3, p6, p11, p7

Exercise Clustering-3 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:

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

$$\begin{aligned}
\mathrm{dist}_2(p,q) &= 5.3851\ldots \\
\mathrm{dist}_1(p,q) &= 9 \\
\mathrm{dist}_\infty(p,q) &= 4 \\
\mathrm{dist}_w(p,q) &= 7.1063 \\
\mathrm{dist}_{M_1}(p,q) &= 5.3851\ldots \\
\mathrm{dist}_{M_2}(p,q) &= 8.4261\ldots
\end{aligned}$$

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

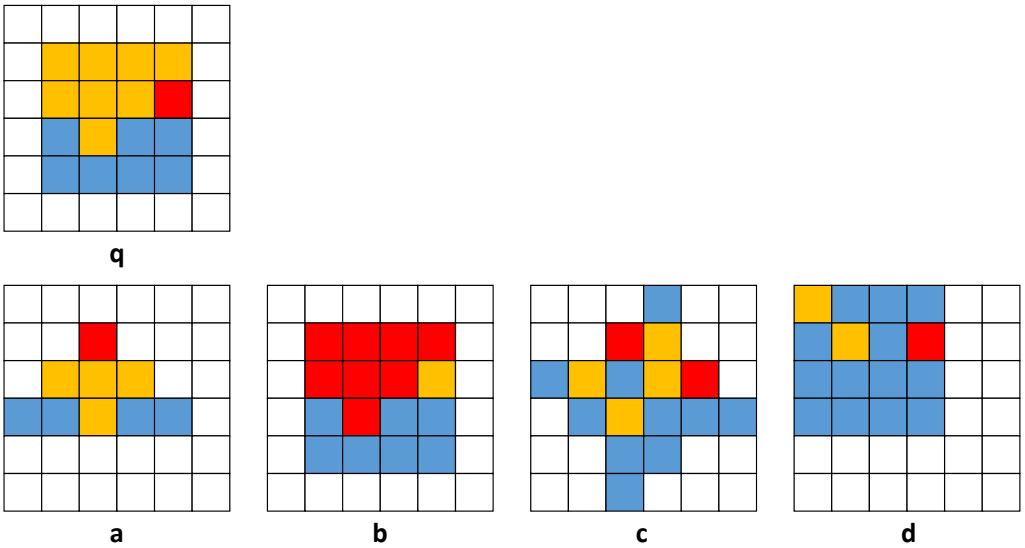


Figure 1: 6×6 pixel pictures

(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.

Color histograms (red, orange, blue); distance

$$
\begin{aligned}
q &= (1, 8, 7) \\
a &= (1, 4, 4); dist(q, a) = 5 \\
b &= (8, 1, 7); dist(q, b) = 9.9 \\
c &= (2, 4, 10); dist(q, c) = 5.1 \\
d &= (1, 2, 13); dist(q, d) = 8.5
\end{aligned}
$$

Ranking: a, c, d, b

(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.

Debatably, picture b is more similar to q than a or d are. The problem is that the Euclidean distance takes each color individually to compute the distance but does not take similarity between different colors (i.e., bins in the histogram) into account.

A solution would be to use the quadratic form (a.k.a. Mahalanobis-) distance. We need a similarity matrix to define the (subjective) similarity of bins with each other:

$$
A = \begin{pmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}
$$

$dist(q, a) = \sqrt{(q - a) \cdot A \cdot (q - a)^T} = 5$
$dist(q, b) = 3.1$
$dist(q, c) = 4.3$
$dist(q, d) = 8.5$