Institut for Matematik og Datalogi
Syddansk Universitet, Odense

17. november 2020
RF

Opgaver DM534 uge 47/48

Husk at læse de relevante sider i slides før du/I forsøger at løse en opgave.

I: Løses i løbet af øvelsestimerne i uge 47

1. Repetér definitionen af en centroide for en cluster og beregn centroiden for en cluster C bestående af følgende tre punkter:

$$C = \{(2,3),(5,5),(4,1)\}.$$

2. Check beregningen af de to centroider i figuren på side 36 i Arthur Zimeks slides.

3. Løs **Exercise Clustering-1** fra de følgende sider.

   [Bemærk i øvrigt at $x^2+y^2 < a^2+b^2 \Leftrightarrow (x^2+y^2)^{1/2} < (a^2+b^2)^{1/2}$, dvs. at man behøver ikke tage kvadratroden, når man beregner afstande fra et punkt til alle centroider (og derefter vælger den nærmeste). Dette er grunden til at formuleringen med "least increase in squared deviations" midt på siden med opgaven blot er det samme som beskrivelsen af k-means algoritmen på slides.]

4. Repetér forskellen på Forgy-Lloyd og MacQueen udgaverne af k-means algoritmen. Giver de to udgaver altid samme resultat?

Den velkendte Euklidiske norm (længde) i $\mathbb{R}^2$ (planen) for et punkt $\vec{v} = (x, y)$ er givet ved $\sqrt{x^2+y^2} = (x^2+y^2)^{1/2}$. Dette kaldes også L_2-normen af $\vec{v}$.

Mere generelt er L_P-normen af et punkt $\vec{v} = (x_1, x_2, \ldots, x_k)$ i $\mathbb{R}^k$ givet ved $\sqrt[P]{\sum_{i=1}^k |x_i|^P} = (\sum_{i=1}^k |x_i|^P)^{1/P}$. Som eksempel er L_3-normen af $\vec{v} = (4, -7)$ givet ved $(4^3 + 7^3)^{1/3} = 7.410\ldots$.

Man definerer desuden L_∞-normen af $\vec{v} = (x_1, x_2, \ldots, x_k)$ som $\max_{i=1}^{k} |x_i|$.

En norm kan bruges til at definere afstanden mellem to punkter $\vec{p}$ og $\vec{q}$ som normen af $\vec{p} - \vec{q}$, dvs. $\text{dist}_P(\vec{p}, \vec{q})$ defineres som L_P-normen af $\vec{p} - \vec{q}$.

5. Beregn følgende:
   (a) L_3-normen af $\vec{v} = (-2, 5)$.
   (b) L_7-normen af $\vec{v} = (4.5, -3.2)$.
   (c) $L_{1.5}$-normen af $\vec{v} = (2, 3)$.
   (d) L_∞-normen af $\vec{v} = (4.5, -3.2)$.
   (e) Afstanden $\text{dist}_3(\vec{p}, \vec{q})$ i L_3-normen mellem $\vec{p}$ og $\vec{q}$, når $\vec{p} = (2, 3)$ og $\vec{q} = (-2, 5)$.
6. Forklar figuren midt på side 6 i Arthur Zimeks slides (som har $P = 1$ og $k = 2$). Dvs. forklar hvorfor mængden af alle punkter $\vec{q}$ i en given afstand r fra et punkt $\vec{q}$ (også kaldet kuglen om $\vec{q}$ med radius r) har en sådan facon.
7. Forklar figuren til højre på side 6 i Arthur Zimeks slides. Forklar hvorfor L_∞ er et godt navn, når man sammenligner med definitionen af L_P (hint: for $\vec{v} = (27, 2)$ beregn $(27^3 + 2^3)^{1/3}$ og $(27^{10} + 2^{10})^{1/10}$).

For at få en ide om udseendet af kugler i L_P-normer generelt, se evt. den engelske Wikipedia-side om superellipsen.

8. Løs **Exercise Clustering-3** fra de følgende sider.

II: Løses hjemme inden øvelsestimerne i uge 46

I k-means algoritmen har initialiseringen (dvs. det første valg af centroider) ofte betydning for slutresultatet. Én metode til initialisering er et tilfældigt valg. En mere struktureret metode er *Furthest First*, som er beskrevet på side 68–78 i disse slides: `https://imada.sdu.dk/~rolf/Edu/DM534/E19/dm534_clustering.pdf`.

1. Læs om Furthest First metoden i disse sider i ovenstående slides og løs derefter **Exercise Clustering-2** fra de følgende sider.

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

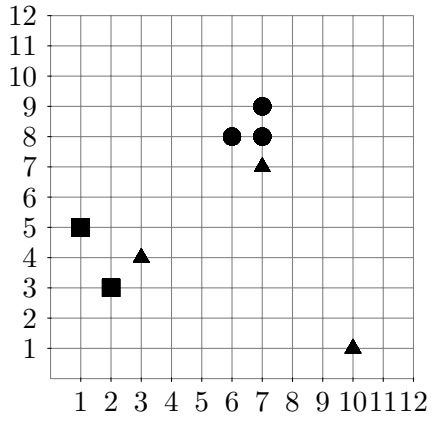
DM534: Introduction to Computer Science

Autumn Term 2019

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.

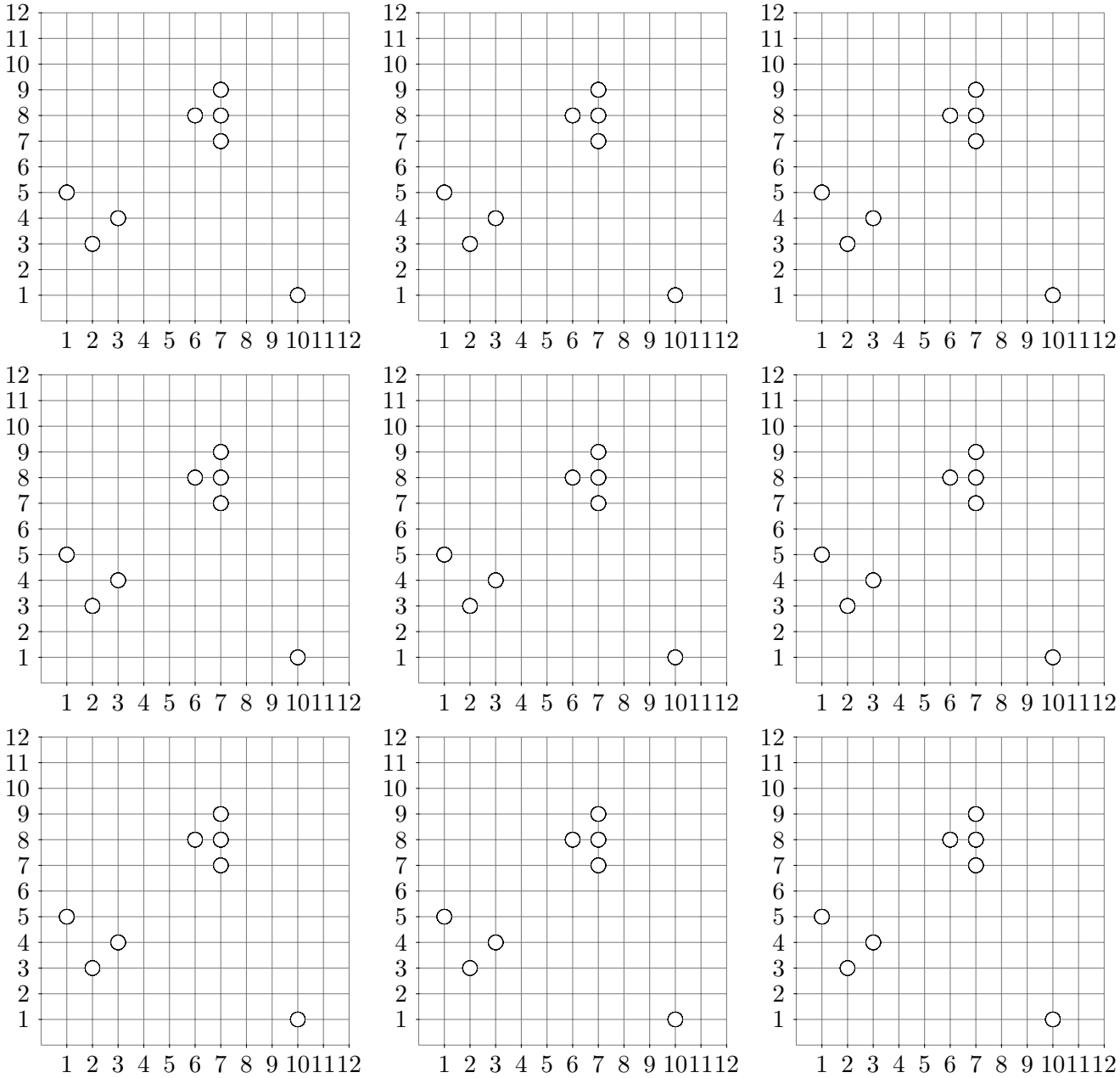
$$SSQ(x, c) = \sum_{i=1}^{d} |x_i - c_i|^2$$

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.

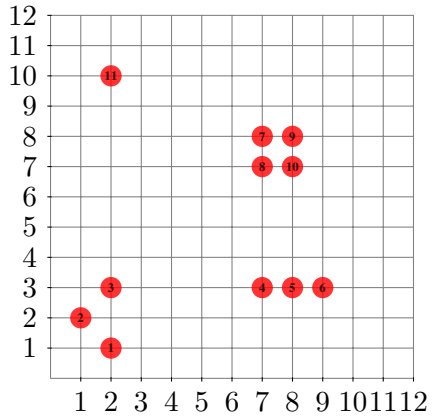
Give the final quality of the clustering (TD^2). How does it compare with the solutions for $k = 2$ discussed in the lecture? Can we conclude on $k = 3$ or $k = 2$ being the better parameter choice on this data set?

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?



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											
2											
3											
4											
5											
6											
7											
8											
9											
10											
11											

L_1 norm:

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

L_∞ norm:

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

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:

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

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

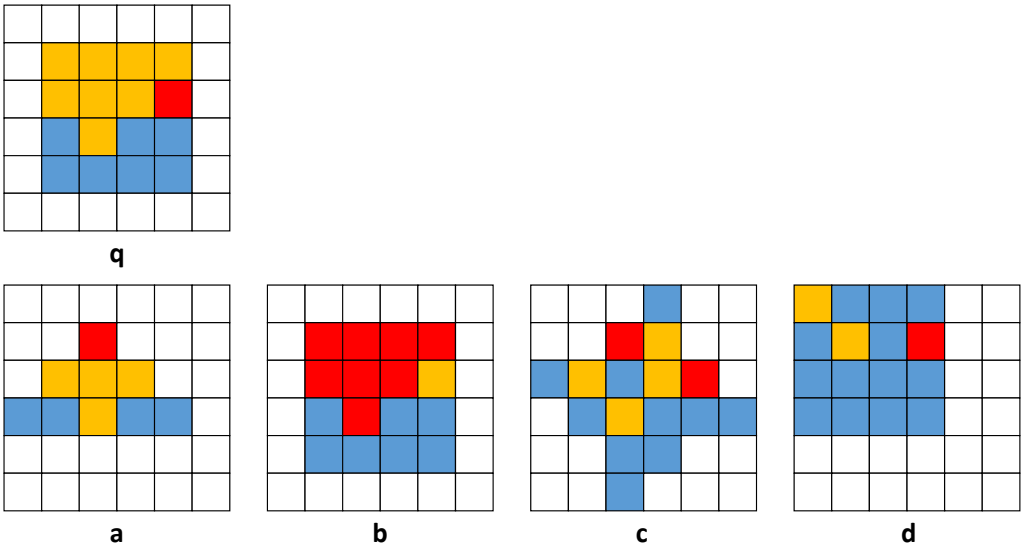


Figure 1: 6×6 pixel pictures

(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.