Institut for Matematik og Datalogi
Syddansk Universitet, Odense

7. oktober 2022
RF

Opgaver DM573 uge 41/43

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

Repetition fra Melih Kandemirs forelæsning:

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 $\vec{v} = (4, -7, 2, 3)$ givet ved $(4^3 + 7^3 + 2^3 + 3^3)^{1/3} = 7.617\ldots$.

Man definerer desuden max-normen af $\vec{v} = (x_1, x_2, \ldots, x_k)$ ved $\max_{i=1}^{k} |x_i|$. Som eksempel er max-normen af $\vec{v} = (4, -7, 2, 3)$ givet ved $\max\{4, 7, 2, 3\} = 7$. Man kalder også max-normen for L_∞-normen (se opgave 3 for en begrundelse).

Givet en norm kan man definere afstanden mellem to punkter $\vec{p}$ og $\vec{q}$ som normen af $\vec{p} - \vec{q}$. For eksempel defineres $\text{dist}_P(\vec{p}, \vec{q})$ som L_P-normen af $\vec{p} - \vec{q}$.

I: Løses i løbet af øvelsestimerne i uge 41

1. 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-normen af $\vec{v} = (5, 9)$.
   (d) $L_{1.5}$-normen af $\vec{v} = (2, 3)$.
   (e) L_∞-normen af $\vec{v} = (4.5, -3.2)$.
2. For $\vec{p} = (1, 2, 3)$ og $\vec{q} = (0, 5, -2)$, beregn følgende:

(a) Afstanden $\text{dist}_2(\vec{p}, \vec{q})$.

(b) Afstanden $\text{dist}_3(\vec{p}, \vec{q})$.

(c) Afstanden $\text{dist}_1(\vec{p}, \vec{q})$.

(d) Afstanden $\text{dist}_\infty(\vec{p}, \vec{q})$.

3. Forklar figuren midt på side 30 i Melih Kandemirs slides. Dvs. forklar hvorfor mængden af alle punkter $\vec{q}$ i en given afstand r fra et punkt $\vec{q}$ (også kaldet "cirklen" om $\vec{q}$ med radius r) har en sådan facon, når afstand er givet ved dist_1.

   Forklar også figuren til højre på samme side.

   Forsøg også at forklare, hvorfor L_∞ er et godt navn for max-normen, 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 idé om udseendet af kugler i L_P-normer generelt, se evt. den engelske Wikipedia-side om superellipsen.]

4. Consider the five pictures given in Figure 1, each with 36 pixels.

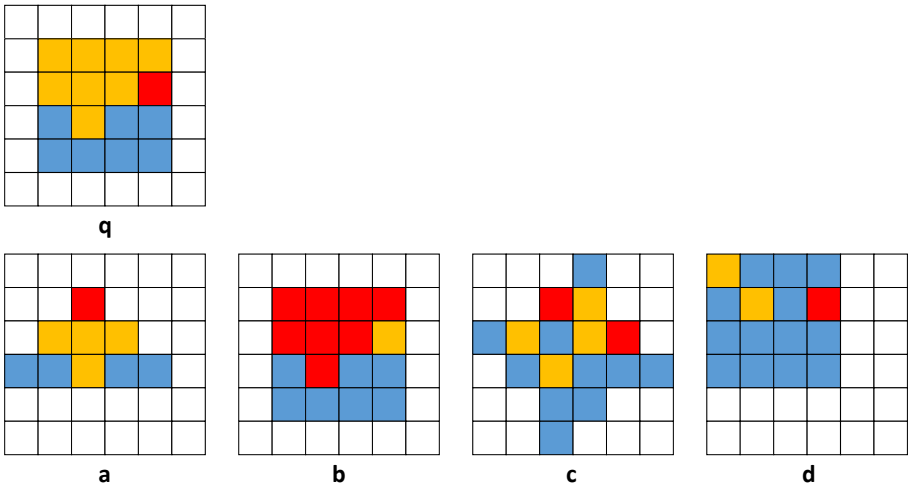


Figure 1: 6×6 pixel pictures

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

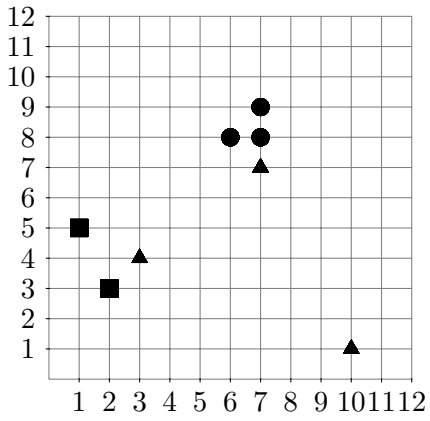
(b) For each of the pictures a to d, calculate their similarity to q using Euclidean distance (i.e., using dist_2).

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

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

6. Check beregningen af de to centroider i figuren på side 35 i Melih Kandemirs slides.

7. Consider the following data set (with 8 objects in $\mathbb{R}^2$) used in the lecture:



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

Start with computing the initial centroids, and draw the cluster assignments after each step and explain the step. As a help, you can use the data set figures last in this document.

Calculate 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?

If time permits: 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?

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

II: Løses hjemme inden øvelsestimerne i uge 43

1. For each of the following distance measures (Euclidean, Manhattan, maximum, weighted Euclidean)

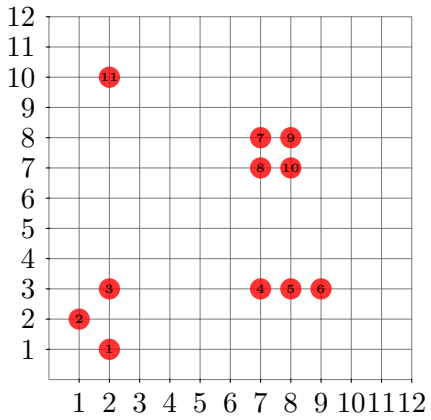
$$\begin{aligned}
\text{dist}_2(\vec{p},\vec{q}) &= \left(|p_1-q_1|^2+|p_2-q_2|^2+|p_3-q_3|^2\right)^{\frac{1}{2}} \\
\text{dist}_1(\vec{p},\vec{q}) &= |p_1-q_1|+|p_2-q_2|+|p_3-q_3| \\
\text{dist}_\infty(\vec{p},\vec{q}) &= \max(|p_1-q_1|,|p_2-q_2|,|p_3-q_3|) \\
\text{dist}_w(\vec{p},\vec{q}) &= \left(w_1|p_1-q_1|^2+w_2|p_2-q_2|^2+w_3|p_3-q_3|^2\right)^{\frac{1}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

   calculate the distance between $\vec{p}=(2,3,5)$ and $\vec{q}=(4,7,8)$. As weights w, use $\vec{w}=(1,1.5,2.5)$.

2. 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`.

   Læs først om Furthest First metoden i ovenstående slides. Løs derefter nedenstående opgave.

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



   Perform a furthest-first initialization on this data set. Do it three times, once for each of the following distance measures, where $p=(p_1,p_2)$ and $q=(q_1,q_2)$.

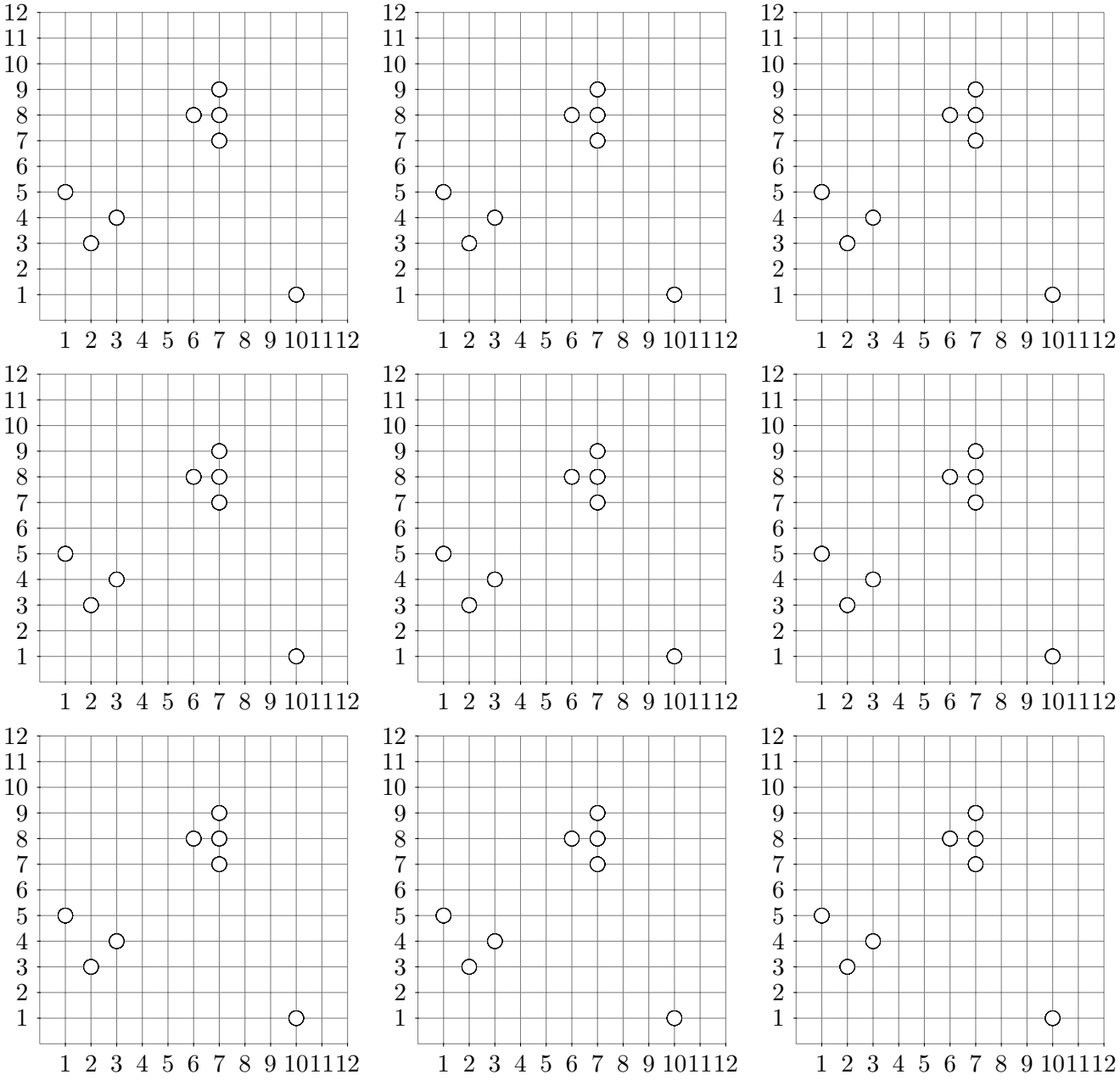
$$\begin{aligned}
\text{dist}_2(p,q) &= \left(|p_1-q_1|^2+|p_2-q_2|^2\right)^{\frac{1}{2}} \\
\text{dist}_1(p,q) &= |p_1-q_1|+|p_2-q_2| \\
\text{dist}_\infty(p,q) &= \max(|p_1-q_1|,|p_2-q_2|)
\end{aligned}$$

In more detail: Let $k = 4$, i.e., in each initialisation choose four centers. Let the point with label 3 be the first center and calculate the next three centers according to the furthest-first method. Do this three times using the three different norms. In case two or more points have the same distance, choose the point with the smallest label.

The calculations needed will fill out parts of the matrix of all pair-wise distances between all points (note: only the upper triangle is required since the distance functions are symmetric). As a help, you can use the tables last in this document to store your calculated distances.

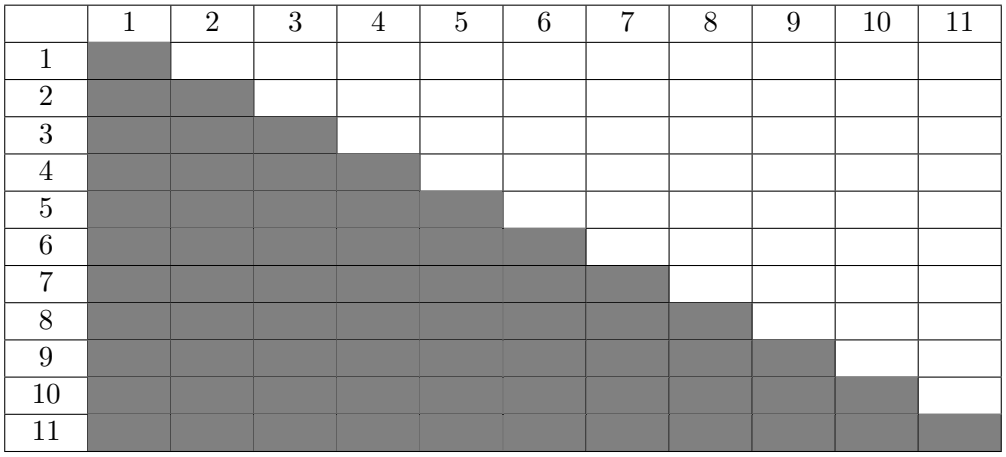
Does the sequence of points differ between the norms?

Datasæt figurer til brug for opgave I.7

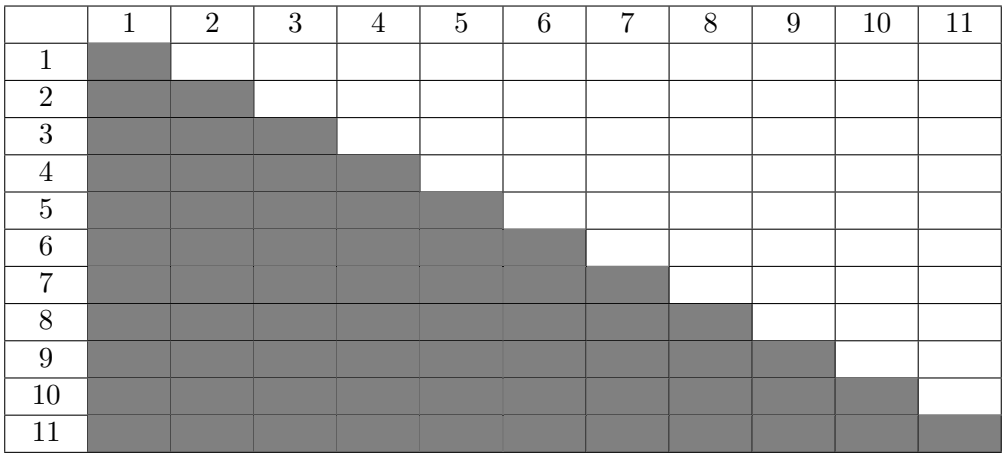


Datasæt figurer til brug for opgave II.2

L_2 norm:



L_1 norm:



L_∞ norm:

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