DM534 — Øvelser Uge 43

Introduktion til Datalogi, Efterår 2021

Jonas Vistrup, med tilføjelser af Rolf Fagerberg

---

1 I

1.1

Vi ser på følgende datasæt

$$S = \{(2,3),(4,8),(6,7)\}$$

og bruger den rette linje $y = ax + b$ med $a = 1.25$ og $b = 0.5$ som model for data.

Hvad er værdien af *Loss*-funktionen $L(a,b,S)$ i denne situation?

SVAR: Modellen $y = 1.25x + 0.5$ giver følgende tre bud på y-værdier:

$$\begin{array}{rcl} y_1^* & = & 1.25 \cdot 2 + 0.5 = 3 \\ y_2^* & = & 1.25 \cdot 4 + 0.5 = 5.5 \\ y_3^* & = & 1.25 \cdot 6 + 0.5 = 8 \end{array}$$

Værdien af *Loss*-funktionen $L(a,b,S)$ bliver derfor

$$\begin{array}{rcl} ((3-3)^2 + (5.5-8)^2 + (8-7)^2)/3 & = & (0^2 + (-2.5)^2 + 1^2)/3 \\ & = & (0 + 6.25 + 1)/3 \\ & = & 7.25/3 \\ & = & 2.416\ldots\ . \end{array}$$

1.2

Vi ser stadig på datasættet

$$S = \{(2,3),(4,8),(6,7)\}$$

fra forrige opgave.

(a) Hvad er a og b for den rette linje $y = ax + b$, som minimerer *Loss*-funktionen $L(a,b,S)$?

(b) Hvilken værdi af *Loss*-funktionen $L(a,b,S)$ opnås af denne linje?

SVAR:

For (a) bruger vi formlerne for a_{best} og b_{best}. Vi finder først $\overline{x}$ og $\overline{y}$:

$$\begin{array}{rcl} \overline{x} & = & (2+4+6)/3 = 12/3 = 4 \\ \overline{y} & = & (3+8+7)/3 = 18/3 = 6 \end{array}$$

Derfor har vi fra formlen for a_{best} at

$$\begin{aligned} \text{tæller i formlen} &= (3-6)(2-4)+(8-6)(4-4)+(7-6)(6-4) \\ &= (-3)(-2)+2\cdot 0+1\cdot 2 \\ &= 6+0+2 \\ &= 8 \end{aligned}$$

og

$$\begin{aligned} \text{nævner i formlen} &= (2-4)^2+(4-4)^2+(6-4)^2 \\ &= (-2)^2+0^2+2^2 \\ &= 4+0+4 \\ &= 8. \end{aligned}$$

Altså er $a_{best} = 8/8 = 1$. Dermed fås $b_{best} = \overline{y} - a_{best} \cdot \overline{x} = 6 - 1 \cdot 4 = 2$. Bedste model mht. denne *Loss*-funktion er dermed

$$y = x + 2.$$

For (b): Modellen $y = x + 2$ giver følgende tre bud på y-værdier:

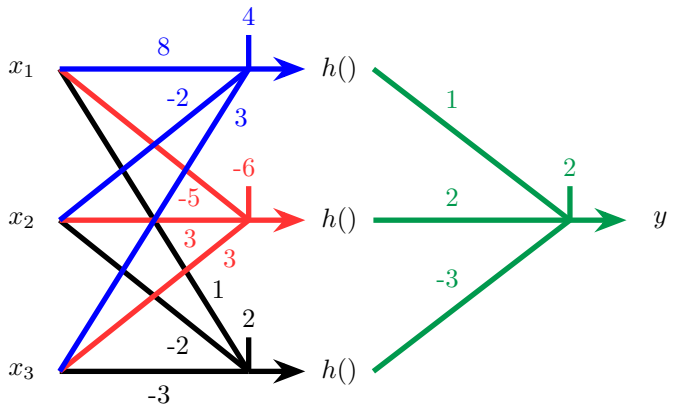
$$\begin{aligned} y_1^* &= 2+2=4 \\ y_2^* &= 4+2=6 \\ y_3^* &= 6+2=8 \end{aligned}$$

Værdien af *Loss*-funktionen $L(a, b, S)$ bliver derfor

$$((4-3)^2+(6-8)^2+(8-7)^2)/3 = (1^2+(-2)^2+1^2)/3 = (1+4+1)/3 = 2.$$

1.3

Figuren nedenfor viser et neuralt netværk med et input lag med værdier x_1, x_2 og x_3, et skjult lag af tre perceptroner (blå, rød og sort), samt et output lag med en perceptron (grøn). Hver perceptron har tre vægte (angivet på dens tre input kanaler) og en bias værdi (angivet lodret over den). Output af perceptronerne i det skjulte lag sendes igennem *activation*-funktionen $h(x) = \max(0, x)$, inden de bliver input til den sidste perceptron.



Hvis input-værdierne er $(x_1, x_2, x_3) = (2, 5, -3)$, hvad er da output-værdien y?

SVAR:

For øverste perceptron (rød) i det skjulte lag bliver output

$$2 \cdot 8 + 5(-2) + (-3)3 + 4 = 16 - 10 - 9 + 4 = 1$$

For næste perceptron (blå) i det skjulte lag bliver output

$$2(-5) + 5 \cdot 3 + (-3)3 + (-6) = (-10) + 15 - 9 - 6 = -10$$

For sidste perceptron (sort) i det skjulte lag bliver output

$$2 \cdot 1 + 5(-2) + (-3)(-3) + 2 = 2 - 10 + 9 + 2 = 3$$

Efter en tur gennem *activation*-funktionen $h(x) = \max(0, x)$ bliver $(1, -10, 3)$ til $(1, 0, 3)$. Dermed bliver output af det sidste lag

$$y = 1 \cdot 1 + 0 \cdot 2 + 3(-3) + 2 = 1 + 0 - 9 + 2 = -6$$

1.4

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

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

SVAR:

$$((2 + 5 + 4)/3, (3 + 5 + 1)/3) = (3, 66..., 3)$$

1.5

Check beregningen af de to centroider i figuren på side 37 i Melih Kandemirs slides.

SVAR:

$$\text{Blue: } ((1 + 3 + 10)/3, (5 + 4 + 1)/3) = (4.66..., 3.33...)$$
$$\text{Red: } ((6 + 7 + 7 + 7)/4, (8 + 9 + 8 + 7)/4) = (6.75, 8)$$

1.6

SVAR: Check løsningen til **Exercise Clustering-1** nedenfor.

1.7

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

Forgy-Lloyd updater alle punkter før centorids genberegnes. MacQueen genberegner centroids efter hver update som resulterer i en ændring.

De to algoritmer giver ikke altid samme resultat.

1.8

SVAR: Check løsningen til **Exercise Clustering-2** nedenfor. [Du skal blot bruge informationen om $\text{dist}_2(p, q)$ og L_2-normen (som svarer til det normale distancemål kendt fra gymnasiet) og ikke bruge informationen om $\text{dist}_1(p, q)$ og L_1-normen, eller om $\text{dist}_\infty(p, q)$ og L_∞-normen.]

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

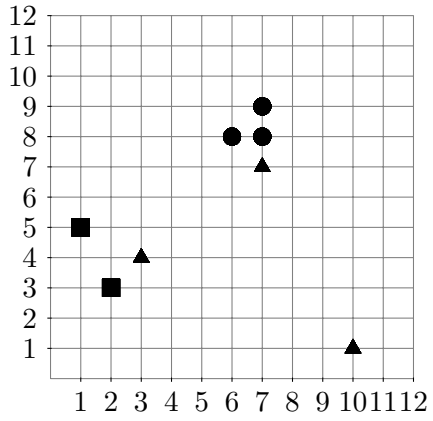
DM534: Introduction to Computer Science

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.

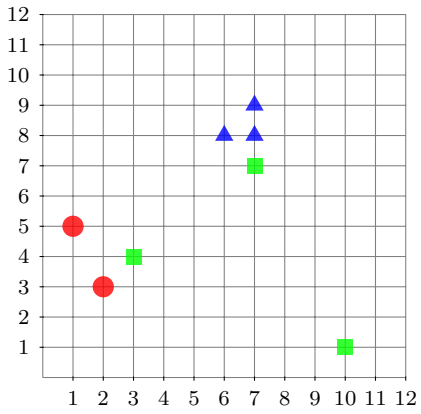
$$SSQ(x, c) = \sum_{i=1}^{d} |x_i - c_i|^2$$

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!

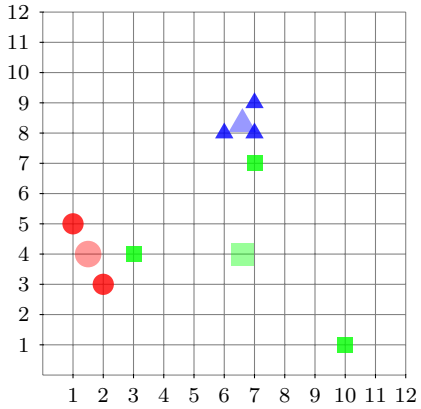
You can use the data set sketches on the next page.

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?

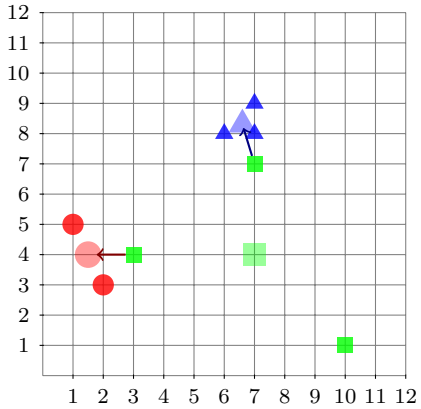


Initial clusters.

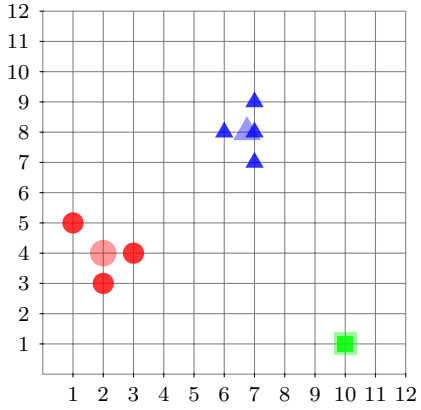


Compute centroids:

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



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.

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

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

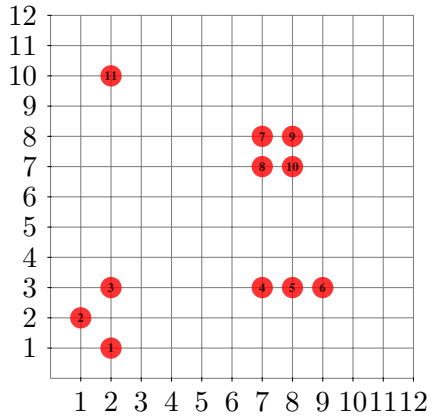
$SSQ(\mu_2, p_5) = |6.75 - 7|^2 + |8 - 7|^2 = \frac{1}{16} + 1 = 1\frac{1}{16}$
$SSQ(\mu_2, p_6) = |6.75 - 6|^2 + |8 - 8|^2 = \frac{9}{16} + 0 = \frac{9}{16}$
$SSQ(\mu_2, p_7) = |6.75 - 7|^2 + |8 - 8|^2 = \frac{1}{16} + 0 = \frac{1}{16}$
$SSQ(\mu_2, p_8) = |6.75 - 7|^2 + |8 - 9|^2 = \frac{1}{16} + 1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

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