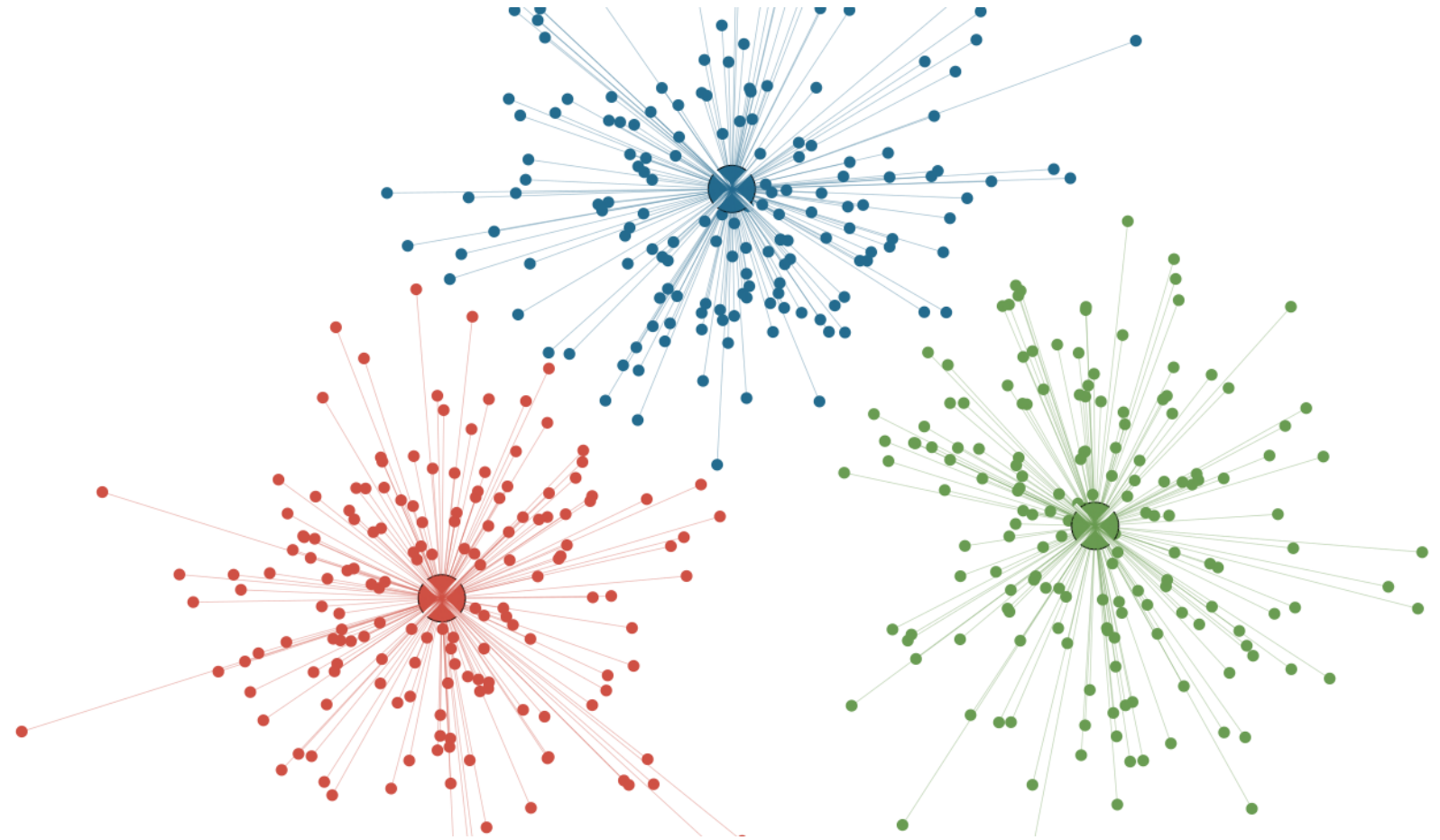




# Segmentering med k-means

Vejledende løsninger

Jesper N. Wulff, Professor

Aarhus Universitet · Institut for Økonomi

2025, 1. udgave

# Løsningsforslag

## Diskussion

**Løsning 1.** Punkter “ligner” hinanden, når afstanden mellem dem er lille. I k-means bruges typisk euklidisk afstand, som svarer til at bruge Pythagoras til at udregne afstanden mellem to punkter.

**Løsning 2.** Trin i k-means: (1) vælg  $k$  og start-centroider, (2) tildel hvert punkt til nærmeste centroide, (3) opdatér hver centroide som gennemsnit af punkterne i klyngen, (4) gentag (2)–(3) indtil konvergens (fx små centroide-flytninger eller uændrede tildelinger).

**Løsning 3.** Inerti er et mål for, hvor “tæt” punkterne ligger omkring deres nærmeste centroide. Man gør det sådan her:

1. For hvert punkt finder man afstanden til den nærmeste centroide.
2. Man kvadrerer afstandene (så store afvigelser vægter ekstra, og negative/positive afvigelser ikke kan udligne hinanden).
3. Til sidst lægger man alle de kvadrerede afstande sammen.

Skrevet som et udtryk:

$$\|x_1 - \mu_{c(1)}\|^2 + \|x_2 - \mu_{c(2)}\|^2 + \cdots + \|x_n - \mu_{c(n)}\|^2,$$

hvor  $\mu_{c(i)}$  er den centroide, som punkt  $x_i$  hører til. Inerti falder typisk fra iteration til iteration, fordi (i) tildelingen vælger den nærmeste centroide for hvert punkt (givet centroiderne), og (ii) opdateringen flytter hver centroide til gennemsnittet, som er det bedste “center” for netop de punkter (givet tildelingen).

**Løsning 4.** For stort  $k$  giver overfitting: mange små klynger, dårligere generalisering, sværere fortolkning, større følsomhed over for støj/outliers, og man risikerer klynger uden meningsfuldt segment.

## Øvelser – Basis

### Løsning 5.

$$d(A, B) = \sqrt{(5-2)^2 + (7-3)^2} = \sqrt{3^2 + 4^2} = 5.$$

**Løsning 6.** Kvadreret afstand til centret  $(300; 2,0)$ :

$$d^2((320; 2,1), (300; 2,0)) = (20)^2 + (0,1)^2 = 400 + 0,01 = 400,01$$

$$d^2((290; 1,9), (300; 2,0)) = (-10)^2 + (-0,1)^2 = 100 + 0,01 = 100,01$$

Nærmest:  $(290; 1,9)$ .

**Løsning 7.** Centroide = gennemsnit:

$$x = \frac{2+4+3}{3} = 3, \quad y = \frac{3+5+4}{3} = 4$$

Ny centroide:  $(3; 4)$ .

**Løsning 8.** Inerti betyder her: *læg de kvadrerede afstande til centret sammen*. Kvadrerede afstande til centret  $(1,5; 1,5)$ :

$$(1; 1) : (1 - 1,5)^2 + (1 - 1,5)^2 = 0,25 + 0,25 = 0,50$$

$$(2; 1) : (2 - 1,5)^2 + (1 - 1,5)^2 = 0,25 + 0,25 = 0,50$$

$$(2; 2) : (2 - 1,5)^2 + (2 - 1,5)^2 = 0,25 + 0,25 = 0,50$$

Derfor er inertien:

$$0,50 + 0,50 + 0,50 = 1,50.$$

### Løsning 9.

1. **Idéen:** Beslutningsgrænsen er de punkter, der ligger *lige langt* fra de to centre. Her ligger centrene på samme vandrette linje:  $C_1 = (0; 0)$  og  $C_2 = (4; 0)$ . Derfor er grænsen en *lodret* linje midt imellem dem.

Midtpunktet mellem  $x = 0$  og  $x = 4$  er  $x = 2$ . Altså er beslutningsgrænsen:

$$x = 2.$$

(Kontrol:  $(x; y)$  er lige langt fra begge, når  $x^2 + y^2 = (x - 4)^2 + y^2$ .  $y^2$  forsvinder på begge sider, så  $x^2 = (x - 4)^2 \Rightarrow x = 2$ .)

2. Punktet  $P = (1; 3)$  har  $x = 1 < 2$ , altså ligger det til venstre for grænsen og er derfor tættere på  $C_1 = (0; 0)$  end på  $C_2 = (4; 0)$ . **Svar:** Klynge 1.

### Løsning 10.

1. Brug kvadrerede afstande til  $C_1^{(0)} = (0; 0)$  og  $C_2^{(0)} = (4; 2)$ :

$$P_1(0; 0) : d_1^2 = 0, d_2^2 = (0 - 4)^2 + (0 - 2)^2 = 16 + 4 = 20 \Rightarrow C_1$$

$$P_2(0; 2) : d_1^2 = (2)^2 = 4, d_2^2 = (0 - 4)^2 + (2 - 2)^2 = 16 \Rightarrow C_1$$

$$P_3(4; 0) : d_1^2 = (4)^2 = 16, d_2^2 = (4 - 4)^2 + (0 - 2)^2 = 4 \Rightarrow C_2$$

$$P_4(4; 2) : d_1^2 = 20, d_2^2 = 0 \Rightarrow C_2$$

2. Opdater centroiderne (gennemsnit i hver klynge):

$$C_1^{(1)} = \left( \frac{0+0}{2}, \frac{0+2}{2} \right) = (0; 1), \quad C_2^{(1)} = \left( \frac{4+4}{2}, \frac{0+2}{2} \right) = (4; 1).$$

3. Inerti efter opdateringen (kvadrerede afstande til de nye centroider):

$$P_1 \rightarrow (0; 1) : (0 - 0)^2 + (0 - 1)^2 = 1, \quad P_2 \rightarrow (0; 1) : (0 - 0)^2 + (2 - 1)^2 = 1$$

$$P_3 \rightarrow (4; 1) : (4 - 4)^2 + (0 - 1)^2 = 1, \quad P_4 \rightarrow (4; 1) : (4 - 4)^2 + (2 - 1)^2 = 1$$

$$\text{Inerti} = 1 + 1 + 1 + 1 = 4.$$

### Løsning 11.

- 1.

$$d(A, B) = \sqrt{(1200 - 1000)^2 + (2 - 1)^2} = \sqrt{200^2 + 1} \approx 200,00$$

$$d(A, C) = \sqrt{(4000 - 1000)^2 + (3 - 1)^2} = \sqrt{3000^2 + 4} \approx 3000,00$$

2. Forbrug (kr.) dominerer, fordi talstørrelserne i den dimension er langt større end i “antal køb”, og i euklidisk afstand slår store forskelle derfor hårdt igennem.
3. Hvis man dividerer forbrug med 1000 (tusinde kr.), bliver de to dimensioner mere sammenlignelige, så “antal køb” får relativt større betydning i afstandene.

## Øvelser – udvidet

**Løsning 12.** Start-centroider:  $C_1^{(0)} = (1; 2)$  og  $C_2^{(0)} = (9; 8)$ .

*Iteration 1:* De tre venstre punkter ligger klart tættere på  $C_1^{(0)}$ , og de tre højre punkter ligger klart tættere på  $C_2^{(0)}$ .

$$\text{Klynge 1: } (1; 1), (2; 1), (1; 2) \Rightarrow C_1^{(1)} = \left( \frac{1+2+1}{3}; \frac{1+1+2}{3} \right) = (1,33; 1,33)$$

$$\text{Klynge 2: } (8; 8), (9; 8), (8; 9) \Rightarrow C_2^{(1)} = \left( \frac{8+9+8}{3}; \frac{8+8+9}{3} \right) = (8,33; 8,33)$$

*Iteration 2:* Med disse centroider ændrer tildelingerne sig ikke, så centroiderne ændrer sig heller ikke:

$$C_1^{(2)} = C_1^{(1)}, \quad C_2^{(2)} = C_2^{(1)}.$$

**Konvergens efter 2 iterationer:** Ja (tildelinger og centroider ændrer sig ikke).

## Løsning 13.

1. Eksempel:  $C_1 = (0; 0)$ ,  $C_2 = (2; 0)$  og  $P = (1; 0)$ . Da er  $d(P, C_1) = d(P, C_2) = 1$ .
2. Mulige tiebreakers: (i) vælg klyngen med lavest indeks, (ii) vælg den klynge med flest/færrest medlemmer, (iii) vælg tilfældigt.
3. Ja, tiebreak kan påvirke klyngestørrelser og dermed opdaterede centroider, især i små datasæt eller ved symmetrier.

## Løsning 14.

1. Med center  $C = (0; 0)$ :

$$(0; 0) : 0, \quad (0; 2) : 4, \quad (2; 0) : 4 \Rightarrow \text{Inerti} = 8.$$

2. Gennemsnittet af punkterne er:

$$\bar{x} = \frac{0+0+2}{3} = \frac{2}{3}, \quad \bar{y} = \frac{0+2+0}{3} = \frac{2}{3} \Rightarrow C = \left(\frac{2}{3}; \frac{2}{3}\right).$$

Kvadrerede afstande til dette center:

$$(0;0) : \left(\frac{2}{3}\right)^2 + \left(\frac{2}{3}\right)^2 = \frac{8}{9}$$

$$(0;2) : \left(\frac{2}{3}\right)^2 + \left(\frac{4}{3}\right)^2 = \frac{20}{9}$$

$$(2;0) : \left(\frac{4}{3}\right)^2 + \left(\frac{2}{3}\right)^2 = \frac{20}{9}$$

Læg sammen:

$$\text{Inerti} = \frac{8}{9} + \frac{20}{9} + \frac{20}{9} = \frac{48}{9} = \frac{16}{3} \approx 5,33.$$

3. Inertien er mindre ved gennemsnittet end ved et vilkårligt punkt som  $(0;0)$ . Det illustrerer, hvorfor gennemsnittet er et godt center i k-means: det gør de kvadrerede afstande samlet set mindst mulige for den klynge.

### Løsning 15.

1. Uden outlier: Centroider (gennemsnit)

$$C = \left(\frac{0+0+1+1}{4}; \frac{0+1+0+1}{4}\right) = (0,5; 0,5).$$

Kvadrerede afstande:

$$(0;0) : 0,5, (0;1) : 0,5, (1;0) : 0,5, (1;1) : 0,5 \Rightarrow \text{Inerti} = 2,0.$$

2. Med outlier  $(20;20)$ : Ny centroider

$$C' = \left(\frac{0+0+1+1+20}{5}; \frac{0+1+0+1+20}{5}\right) = \left(\frac{22}{5}; \frac{22}{5}\right) = (4,4; 4,4).$$

3. Outliers kan trække centroider kraftigt og øge inertien markant. Derfor er k-means følsom over for ekstreme punkter.

## Excel

**Løsning 16-20.** Se excel-filerne `k_means_k3_lærer.xlsx` og `k_means_k2_lærer.xlsx`.