DM534 — Øvelser Uge 43

Introduktion til Datalogi, Efterår 2021

Jonas Vistrup

---

1 I

1.1

Beregn følgende:

(a) L_3-normen af $\overrightarrow{v} = (-2, 5)$. $(|-2|^3 + |5|^3)^{1/3} = 5.10447$.

(b) L_7-normen af $\overrightarrow{v} = (4.5, -3.2)$. $((4.5)^7 + (|-3.2|)^7)^{1/7} = 4.55691$.

(c) L_1-normen af $\overrightarrow{v} = (5, 9).5 + 9 = 14$.

(d) $L_{1.5}$-normen af $\overrightarrow{v} = (2, 3).(2^{1.5} + 3^{1.5})^{1/1.5} = 4.00819$.

(e) L_∞-normen af $\overrightarrow{v} = (4.5, -3.2)$. $\max(|4.5|, |-3.2|) = 4.5$

1.2

For $\overrightarrow{p} = (1, 2, 3)$ og $\overrightarrow{q} = (0, 5, -2)$, beregn følgende:

(a) Afstanden $\text{dist}_2(\overrightarrow{p}, \overrightarrow{q})$: $\left(|1-0|^2 + |2-5|^2 + |3-(-2)|^2\right)^{1/2} = 5.91608$

(b) Afstanden $\text{dist}_3(\overrightarrow{p}, \overrightarrow{q})$: $\left(|1-0|^3 + |2-5|^3 + |3-(-2)|^3\right)^{1/3} \approx 5.348...$

(c) Afstanden $\text{dist}_1(\overrightarrow{p}, \overrightarrow{q})$: $(|1-0| + |2-5| + |3-(-2)|)^{1/1} = 9$

(d) Afstanden $\text{dist}_\infty(\overrightarrow{p}, \overrightarrow{q})$: $(|1-0|^\infty + |2-5|^\infty + |3-(-2)|^\infty)^{1/\infty} = 5$

1.3

Forklar figuren midt på side 6 i Melih Kandemirs slides (som har $P = 1$ og $k = 2$). Dvs. forklar hvorfor mængden af alle punkter $\overrightarrow{q}$ i en given afstand r fra et punkt $\overrightarrow{q}$ (også kaldet kuglen om $\overrightarrow{q}$ med radius r) har en sådan facon.

SVAR: Da Manhatten norm ligger x og y afstandene sammen, så må radius r blive delt ud på x afstanden og y afstanden. Alle måder r kan uddeles på, skaber tilsammen diamant formen.

1.4

SVAR: Check Exercise 3 part 2.

1.5

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

SVAR:

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

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

1.6

Check beregningen af de to centroider i figuren på side 36 i Melih Kandemirs slides.

SVAR:

$$\text{Blue: } ((1+3+10)/3,(5+4+1)/3) = (4.66...,3.33...)$$

$$\text{Red: } ((6+7+7+7)/4,(8+9+8+7)/4) = (6.75,8)$$

1.7

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.]

SVAR: Check løsningen til Exercise 1.

1.8

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 beregner centroids efter hver update som resulter i en ændring.

De resulterer ikke altid i samme resultat.

2 II

2.1

SVAR: Check Exercise 3 part 1.

2.2

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

SVAR: Check Exercise 2.