# 2 - 12 - HVA ER EN KATT?

Nå skal vi se litt på hvordan lineæralgebra brukes i datahåndtering; så og si alt vi har lært dette året kommer til nytte. Vi skal ikke bry oss med "statistikk" som sådan, for i statistikk er man veldig opptatt av å skille mellom parametrene i sannsynlighetsmodellen som ligger til grunn og estimatorene man stapper dataene inn i (for eksempel gjennomsnitt eller empirisk standardavvik). Vi skal bry oss mest med selve lineæralgebraen.

Det er en matrise som er veldig praktisk, nemlig kolonnevektoren

$$\mathbf{1} = (1, 1, \cdots, 1)^T.$$

Det er ikke helt uvanlig å skrive $\mathbf{1}_n$ for å indikere dimensjon, men jeg gidder ikke det.

1 Regn ut $\mathbf{1}^T\mathbf{1}$, $\mathbf{1}\mathbf{1}^T$ og $\mathbf{1}^T\mathbf{x}$.

I faktiske anvendelser er det vanlig å **sentrere** datavektoren $\mathbf{x}$ ved å lage en ny datavektor der man har trukket $\overline{\mathbf{x}}$ fra alle komponentene i $\mathbf{x}$, og så jobbe med denne istedet. Med den nye vektoren $\mathbf{1}$ kan vi skrive den sentrerte variabelen som

$$\mathbf{x} - \overline{\mathbf{x}} \cdot \mathbf{1} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} - \overline{\mathbf{x}} \cdot \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} - \begin{pmatrix} \overline{\mathbf{x}} \\ \overline{\mathbf{x}} \\ \vdots \\ \overline{\mathbf{x}} \end{pmatrix} = \begin{pmatrix} x_1 - \overline{\mathbf{x}} \\ x_2 - \overline{\mathbf{x}} \\ \vdots \\ x_n - \overline{\mathbf{x}} \end{pmatrix} = \left( I - \frac{1}{n}\mathbf{1}\mathbf{1}^T \right) \mathbf{x}$$

Matrisen

$$C_n = \left( I - \frac{1}{n}\mathbf{1}\mathbf{1}^T \right)$$

kalles **sentreringsmatrisen** fordi den sentrerer $\mathbf{x}$. Husk at en **projeksjon** er en lineæroperator $P$ som tilfredsstiller $P^2 = P$.

2 Vis at $C_n$ er en projeksjon. og finn gjennomsnittet til den nye sentrerte variabelen.

Med sentrerte variable blir livet greit. I økt 2-6 utledet du uttrykket

$$\beta_1 = \frac{\mathbf{x}^T\mathbf{y} - n\bar{\mathbf{x}}\bar{\mathbf{y}}}{\mathbf{x}^T\mathbf{x} - n\left(\bar{\mathbf{x}}\right)^2} \qquad \text{og} \qquad \beta_0 = \overline{\mathbf{y}} - \beta_1\overline{\mathbf{x}}$$

for koeffisientene til det rettlinjede regresjonspolynomet.

3 Dette ser mye penere ut om $\mathbf{x}$ og $\mathbf{y}$ er sentrerte, og forklarer hvorfor noen sier at man "projiserer $\mathbf{y}$ på $\mathbf{x}$" når man kjører regresjon. Skriv opp, forklar og tolk.

Skjønte du forrige side, kan du gå med på at sentrerte variable er lurt. La oss spinne videre litt. **Multippel regresjon** handler om at forklaringsvariabelen har flere dimensjoner. Du kan for eksemepel tenke at du har målt sylindervolum $\mathbf{x}_1$ og vekt $\mathbf{x}_2$ på $n$ biler, og så er responsvariabelen $\mathbf{y}$ drivstofforbruk per kilometer. Hvis du ønsker å tilpasse et første ordens regresjonspolynom i to variable, får du det overbestemte likningssettet

$$\mathbf{y} = \beta_{10}\mathbf{x}_1 + \beta_{01}\mathbf{x}_2 + \beta_0\mathbf{1}$$

eller

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \beta_{10}\begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} + \beta_{01}\begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix} + \beta_0\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

4 Vis at dersom $\mathbf{x}_1$, $\mathbf{x}_2$ og $\mathbf{y}$ er sentrerte, er $\beta_0 = 0$.

La oss anta du har $n$ realiseringer av en $p$-dimensjonal tilfeldig variabel. Det vanlig å sette disse opp i en $n \times p$-matrise:

$$X = \left( \begin{array}{c|c|c|c} & & & \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \\ & & & \end{array} \right)$$

Tenk at du måler lengde, bredde og høyde på alle plankene i en plankestabel, og så sentrerer du hver av disse. Til slutt setter du alt opp i matrisen $X$ på en slik måte at rad $k$ er de sentrerte målene til planke $k$. Vi kaller $\frac{1}{n-1}X^TX$ den **empiriske kovariansmatrisen**, og den er en estimator for den antatte bakenforliggende kovariansmatrisen.

5 Vis at elementene i $\frac{1}{n-1}X^TX$ er den empiriske kovariansen mellom kolonnene i $X$.

Vi sier at en matrise $A$ er **positivt definitt** dersom

$$\mathbf{x}^TA\mathbf{x} > 0$$

for alle $\mathbf{x}$, og **positivt semidefinitt** dersom

$$\mathbf{x}^TA\mathbf{x} \geq 0$$

for alle $\mathbf{x}$.

6 Vis at den empiriske kovariansmatrisen er positivt semidefinitt.

I flervariabel statistikk finnes det noe som kalles **prinsipalkomponentanalyse**.[1] Dette er en teknikk for å lete opp akser i datamengden der samvariasjonen er sterk. Teknikken baserer seg på å projisere $X$ inn på en retning $\mathbf{v}$ slik at den empiriske variansen til de resulterende punktene blir maksimert. Dette får man til ved å lete opp egenvektorene til de største egenverdiene til kovariansmatrisen.

7 I 2-6 er det flere tovariable datamengder. Plott kolonnene mot hverandre i python, regn ut egenvektorene, og sjekk samvariasjon.

8 Er det samvariasjon mellom høyde og bredde i kaibordene fra sommeren 2023?
`https://folk.ntnu.no/mortano/python/kai/kai.csv`

9 Forklar hvorfor egenvektorene til $X^TX$ forteller om samvariasjonen mellom kolonnene i $X$. (Hint: Finn ut hva en "Rayleigh quotient" er.[2])

Noen ganger er det regnet som ugunstig dersom kolonnene i datamatrisen har veldig forskjellig lengde. Dette fikser vi ved å dele ut standardavviket i hver kolonne, akkurat som når man normaliserer vektorer ellers i lineæralgebra.

Se eller her:
`https://tma4245.math.ntnu.no/forventing-og-varians/korrelasjon/`
og her:
`https://tma4245.math.ntnu.no/forventing-og-varians/kovarians/`

[1] `https://en.wikipedia.org/wiki/Principal_component_analysis`
[2] `https://en.wikipedia.org/wiki/Rayleigh_quotient`

Nå skal vi se at det er faktisk ikke nødvendig å gange datamatrisen med seg selv for å finne kovariansmatrisen. Det finnes noe enda mer sofistiskert som heter SVD. La oss begynne med litt geometrisk fikling.

10 Skriv en rutine som tar inn $A \in \mathbb{R}^{2\times 2}$ og plotter $A\mathbf{x}$ for alle $\mathbf{x}$ på enhetssirkelen.

Fikk du til dette, vil du oppdage at koden din alltid produserer ellipser, ihvertfall så lenge $A$ har linært uavhengige kolonner. La oss finne ut mer om disse ellipsoidene. Trikset er å studere $A^TA$ og $AA^T$. Vi lar $n > p$, slik at $A$ er høyt og tynn og $A^T$ er lang og flat. Dette er ikke noen essensiell restriksjon, men vi gjør den fordi det hjelper litt på visualiseringen.

11 Både $A^TA$ og $AA^T$ er ortogonalt diagonaliserbare. Hvorfor?

Siden $A^TA$ er ortogonalt diagonaliserbar, finnes det et ortonormalt sett med egenvektorer. La oss sette disse opp i en ortogonal matrise $V$. Det samme gjelder for $AA^T$; la oss sette disse egenvektoren opp i en ortonormal matrise $U$.

12 Hva er dimensjonene til $A^TA$, $AA^T$, $V$ og $U$?
Hvor mange av egenverdiene til $AA^T$ kan være forskjellige fra null?

Her er enda en observasjon.

13 Vis at dersom $\mathbf{v}$ er en egenvektor til $A^TA$ med egenverdi $\lambda$, er $A\mathbf{v}$ en egenvektor til $AA^T$. Hva er egenverdien? Kan du si noe om fortegnet?

Hvis du skjønte forrige oppgave, skjønner du forhåpentligvis at $A^TA$ og $AA^T$ har de samme egenverdiene så lenge du ser bort fra multiplisiteten til de egenverdiene som er null. La oss nå anta at $\mathbf{v}$ er en egenvektor til $A$ med lengde 1.

14 Hva er lengden til $A\mathbf{v}$?

Lengden til $A\mathbf{v}$ kalles **singulærverdi**, og vi bruker bokstaven $\sigma_k$. Kolonnene i $U$ og $V$ kalles henholdsvis de **venstre og høyre singulærvektorene**. Hvis du tenker nøye på alt du har gjort til nå, vil du se at det går an å skrive

$$A = U\Sigma V^T$$

der $\sigma$-ene sitter i den rektangulære diagonalmatrisen $\Sigma$. Dette kalles **SVD-faktoriseringen**[3] til $A$. Denne finnes alltid, og dersom du sorterer singulærverdiene i synkende rekkefølge ($\sigma_1$ er størst) og velger alle positive, er $\Sigma$ entydig. De unitære matrisene $U$ og $V$ vil ikke nødvendigvis være entydige, men de er det i noen tilfeller.

15 Hvordan ser $\Sigma$ ut? Finn svd-faktoriseringen til

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}$$

Singulærverdidekomposisjonen er en av de viktigste matrisefaktoriseringene. Har du den har du mye, og pythonkommandoen heter numpy.linalg.svd. Nå skal vi se litt på hva svd kan brukes til. Det første er kompresjon. Vi kan skrive $A$ slik:

$$A = \sigma_1\mathbf{u}_1\mathbf{v}_1^* + \sigma_2\mathbf{u}_2\mathbf{v}_2^* + \sigma_3\mathbf{u}_3\mathbf{v}_3^* + \cdots + \sigma_r\mathbf{u}_r\mathbf{v}_r^*$$

der $r$ er antallet singulærverdier som ikke er null. Vi ser nå at dersom det er bare er to av singulærverdiene som er over for eksempel maskinpresisjon, er all informasjon i matrisen inneholdt i uttrykket

$$A = \mathbf{u}_1\sigma_1\mathbf{v}_1^* + \mathbf{u}_2\sigma_2\mathbf{v}_2^*.$$

16 La oss lage et bilde av en katt. Lag deg en matrise $A$ i python med tilfeldige tall mellom -1 og 1. Dette er bakgrunnen. Sett så ett element til 10 eller 100 eller noe sånt. Dette er katten. Kjør svd på $A$ og skriv ut

$$A - \mathbf{u}_1\sigma_1\mathbf{v}_1^* \qquad \text{og} \qquad A - (\mathbf{u}_1\sigma_1\mathbf{v}_1^* + \mathbf{u}_2\sigma_2\mathbf{v}_2^*)\,.$$

[3] https://en.wikipedia.org/wiki/Singular_value_decomposition