Øvingsforelesning til 12 - Interpolasjon, regresjon og markovkjeder

Interpolasjon og regresjon

1. Gitt et ligningssystem $A\mathbf{x} = \mathbf{b}$ som ikke har en løsning. Jeg kan bruke minste kvadratersmetode for å

- minimere lengden til vektoren $\mathbf{x}$
- finne den beste approksimasjon til en løsning for systemet
- jeg vet ikke fordi jeg ikke har forberedt meg.

Løsning: Vi kan bruke metoden for å finne den beste approksimasjon til en løsning for systemet. Vi bruker den for å finne $\hat{\mathbf{x}}$ slik at $A\hat{\mathbf{x}}$ har minst avstand til $\mathbf{b}$ (og vi bruker summen av kvadratene til koordinatene til å beregne avstander).

2. (Vår 2020, oppgave 31) Finn linjen som går gjennom eller passer best til punktene i $\mathbb{R}^2$

$$\begin{bmatrix}-1\\-1\end{bmatrix}, \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1\\2\end{bmatrix}.$$

Løsning: Vi leter etter en linje $y = cx+d$ som passer best til punktene. Husk at dersom det fantes en linje gjennom alle punktene så kunne vi finne $c, d \in \mathbb{R}$ slik at

$$\begin{bmatrix}-1 & 1\\ 0 & 1\\ 1 & 1\end{bmatrix}\begin{bmatrix}c\\d\end{bmatrix} = \begin{bmatrix}-1\\0\\2\end{bmatrix}.$$

Dette ligningssystemet har ikke noen løsninger fordi når vi Gausseliminerer får vi

$$\left[\begin{array}{cc|c} -1 & 1 & -1\\ 0 & 1 & 0\\ 1 & 1 & 2\end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 0 & 1\\ 0 & 1 & 0\\ 0 & 0 & 1\end{array}\right].$$

Punktene ligger altså ikke på en linje, og vi kan bare finne en linje som minimerer avstanden til alle punktene.

Det gjør vi ved å bruke minste kvadraters metode på

$$A = \begin{bmatrix}-1 & 1\\ 0 & 1\\ 1 & 1\end{bmatrix} \text{ og } \mathbf{b} = \begin{bmatrix}-1\\0\\2\end{bmatrix}.$$

Vi beregner

$$A^TA = \begin{bmatrix}-1 & 0 & 1\\ 1 & 1 & 1\end{bmatrix}\begin{bmatrix}-1 & 1\\ 0 & 1\\ 1 & 1\end{bmatrix} = \begin{bmatrix}2 & 0\\ 0 & 3\end{bmatrix}$$

og

$$A^T\mathbf{b} = \begin{bmatrix}-1 & 0 & 1\\ 1 & 1 & 1\end{bmatrix}\begin{bmatrix}-1\\0\\2\end{bmatrix} = \begin{bmatrix}3\\1\end{bmatrix}.$$

Nå må vi løse ligningssystemet med totalmatrisen

$$\left[\begin{array}{cc|c} 2 & 0 & 3\\ 0 & 3 & 1\end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 0 & 3/2\\ 0 & 1 & 1/3\end{array}\right].$$

Linjen er altså $y = \frac{3}{2}x + \frac{1}{3}$.

3. La A være en 4×2-matrise. Hvilken av matrisene kan være lik $A^T \cdot A$?

$$L = \begin{bmatrix}14 & 6\\ 6 & 4\end{bmatrix},\ M = \begin{bmatrix}14 & 6\\ 4 & 4\end{bmatrix},\ N = \begin{bmatrix}14 & 0\\ 6 & 14\end{bmatrix}$$

Løsning: Bare L kan være A^TA fordi A^TA er en symmetrisk matrise, dvs elementene over og under diagonalen er like.

4. Vi skal finne polynomer som går gjennom eller passer best til punktene i $\mathbb{R}^2$

$$\begin{bmatrix}0\\-1\end{bmatrix}, \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}2\\5\end{bmatrix}, \begin{bmatrix}3\\20\end{bmatrix}.$$

a) Finnes det en linje på formen $y = ax + b$ som går gjennom alle punktene? Hvis ikke, bruk minste kvadraters metode til å finne linjen som passer best til punktene.

Løsning: Hvis en slik linje $y = ax + b$ eksisterer, må den oppfyller

$$\begin{aligned} -1 &= a\cdot 0 + b\\ 0 &= a\cdot 1 + b\\ 5 &= a\cdot 2 + b\\ 20 &= a\cdot 3 + b. \end{aligned}$$

Dette er et system med fire ligninger, men kun to variabler. Fra den første ligningen får vi $b = -1$ og så fra den andre $a = 1$. Men $5 \neq 1 - 2$ og systemet har ingen løsninger a, b. Dermed kan vi bare prøve å finne den beste approksimasjonen ved hjelp av minstre kvadraters metoden.

Vi ser altså på systemet $A\mathbf{x} = \mathbf{b}$ med

$$A = \begin{bmatrix}0 & 1\\ 1 & 1\\ 2 & 1\\ 3 & 1\end{bmatrix},\ \mathbf{x} = \begin{bmatrix}a\\b\end{bmatrix},\ \mathbf{b} = \begin{bmatrix}-1\\0\\5\\20\end{bmatrix}.$$

Vi beregner

$$A^TA = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix} = \begin{bmatrix} 14 & 6 \\ 6 & 4 \end{bmatrix}$$

og

$$A^T\mathbf{b} = \begin{bmatrix} 70 \\ 24 \end{bmatrix}.$$

Nå må vi løse systemet $A^TA\mathbf{x} = A^T\mathbf{b}$:

$$\left[\begin{array}{cc|c} 14 & 6 & 70 \\ 6 & 4 & 24 \end{array}\right] \sim \left[\begin{array}{cc|c} 7 & 3 & 35 \\ 3 & 2 & 12 \end{array}\right]$$
$$\sim \left[\begin{array}{cc|c} 1 & -1 & 11 \\ 3 & 2 & 12 \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 0 & 34/5 \\ 0 & 1 & -21/5 \end{array}\right].$$

Linjen som passer best til punktene er altså

$$y = \frac{34}{5}x - \frac{21}{5}.$$

b) Det finnes ikke et annengradspolynom som går gjennom alle punktene heller. Hvilket ligningssystem må vi se på for å vise dette? Hva må vi sjekke for dette systemet?

Løsning: Hvis det fantes et annengradspolynom gjennom alle punktene, kunne vi finne at polynom $q(x) = ax^2 + bx + c$ slik at

$$\begin{aligned} q(0) &= a \cdot 0 + b \cdot 0 + c = -1 \\ q(1) &= a \cdot 1^2 + b \cdot 1 + c = 0 \\ q(1) &= a \cdot 2^2 + b \cdot 2 + c = 5 \\ q(1) &= a \cdot 3^2 + b \cdot 3 + c = 20. \end{aligned}$$

Dermed må vi prøve å løse systemet $A\mathbf{x} = \mathbf{b}$ med

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \end{bmatrix}, \ \mathbf{x} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} -1 \\ 0 \\ 5 \\ 20 \end{bmatrix}.$$

For å vise at det ikke finnes et annengradspolynom som går gjennom alle punktene må vi vise at dette systemet ikke har en løsning.

c) Men det finnes et unikt tredjegradspolynom som går gjennom alle punktene. Sett opp et ligningssystem for koeffisientene til dette polynomet, og finn koeffisientene.

Løsning:

En tredjegradspolynom er på formen

$$p(x) = ax^3 + bx^2 + cx + d.$$

Vi må altså finne koeffisienter a, b, c, d slik at

$$\begin{aligned} p(0) &= a \cdot 0 + b \cdot 0 + c \cdot 0 + d = -1 \\ p(1) &= a \cdot 1^3 + b \cdot 1^2 + c \cdot 1 + d = 0 \\ p(2) &= a \cdot 2^3 + b \cdot 2^2 + c \cdot 2 + d = 5 \\ p(3) &= a \cdot 3^3 + b \cdot 3^2 + c \cdot 3 + d = 20. \end{aligned}$$

Dermed må vi løse systemet $A\mathbf{x} = \mathbf{b}$ med

$$A = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 8 & 4 & 2 & 1 \\ 27 & 9 & 3 & 1 \end{bmatrix}, \ \mathbf{x} = \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} -1 \\ 0 \\ 5 \\ 20 \end{bmatrix}.$$

Vi bruker Gausseliminasjon for å finne $\mathbf{x}$:

$$\left[\begin{array}{cccc|c} 0 & 0 & 0 & 1 & -1 \\ 1 & 1 & 1 & 1 & 0 \\ 8 & 4 & 2 & 1 & 5 \\ 27 & 9 & 3 & 1 & 20 \end{array}\right] \sim \left[\begin{array}{cccc|c} 0 & 0 & 0 & 1 & -1 \\ 1 & 1 & 1 & 0 & 1 \\ 8 & 4 & 2 & 0 & 6 \\ 27 & 9 & 3 & 0 & 21 \end{array}\right]$$
$$\sim \left[\begin{array}{cccc|c} 0 & 0 & 0 & 1 & -1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & -2 & -3 & 0 & -1 \\ 0 & -6 & -8 & 0 & -2 \end{array}\right] \sim \left[\begin{array}{cccc|c} 0 & 0 & 0 & 1 & -1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 1 \end{array}\right].$$

Vi får altså $\mathbf{x} = \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}$ og

$$p(x) = x^3 - x^2 + x - 1.$$

d) Kan du nå gjette hva x må være for at den følgende setningen er sant: Gitt m datapunkter i $\mathbb{R}^2$. Da finnes det et unikt x-tegradsplynom som går gjennom alle punktene. Hva ville du gjort for å vise at $x$en du valgte er den riktige?

Løsning: Vi må velge $x = m - 1$. For å bevise vår påstand må vi vise at systemet $A\mathbf{x} = \mathbf{b}$ med

$$A = \begin{bmatrix} x_1^{m-1} & x_1^{m-2} & \dots & x_1 & 1 \\ x_2^{m-1} & x_2^{m-2} & \dots & x_2 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_m^{m-1} & x_m^{m-2} & \dots & x_m & 1 \end{bmatrix}$$

altid har en unik løsning, der $(x_1, y_1), \dots, (x_m, y_m)$ er datapunktene. Med andre ord vi må vise at $\operatorname{Col} A = \mathbb{R}^m$ eller at A har m lineært uavhengige kolonner. Fordi A er en kvadratisk matrise, kan vi beregne $\det A$ og vise påstanden ved å sjekke at $\det A \neq 0$. Matrisen A er nesten den samme som *Vandermondematrisen*. Du kan lese mer om denne spennende og nyttige matrisen i lærebøker eller på nett. Det kan da vises at

$$\det A = \pm \prod_{1 \le i < j \le m} (x_j - x_i).$$

Determinanten $\det A$ er altså $\neq 0$ så lenge $x_i \neq x_j$ for $i \neq j$.

Markovkjeder

1. Markovkjeder er

- en viktig type stokastiske prosesser
- ikke
- jeg vet ikke fordi jeg ikke har forberedt meg.

Løsning: Kort oppsummert beskriver Markovkjeder prosesser der det finnes faste overgangssannsynligheter mellom forskjellige tilstander. Slike prosesser møter vi overalt. Derfor er Markovkjeder kjempenyttige.

2. En likevektsvektor

- legger like mye vekt på alle sine koordinater
- er det samme som nullvektoren
- beskriver tilstanden en Markovkjede konvergerer til.

Løsning: En likevektsvektor beskriver tilstanden en Markovkjede konvergerer til. Den er især en sannsynlighetsvektor, dvs alle koordinatene er ≥ 0 og summen av dem er lik 1. Det finnes en unik likevektsvektor for hver Markovkjede med en regulær stokastisk overgangsmatrise.

3. Hvilken av vektorene kan være en likevektsvektor for en Markovkjede?

$$\mathbf{u} = \begin{bmatrix} 1/3 \\ 4/3 \\ -2/3 \end{bmatrix}, \ \mathbf{v} = \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \end{bmatrix}, \ \mathbf{w} = \begin{bmatrix} 1/3 \\ 1/2 \\ 1/6 \end{bmatrix}$$

Løsning: Bare **w** oppfyller kravene: alle koordinatene er ≥ 0 og summeres til 1.

4. Hvilken av matrisene er en stokastisk matrise

$$L = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix}, \ M = \begin{bmatrix} 1/3 & 1 \\ 2/3 & 0 \end{bmatrix}, \ N = \begin{bmatrix} 1/3 & 4/3 \\ 2/3 & -1/3 \end{bmatrix}$$

Løsning: Bare M har sannsynlighetsvektorer som kolonner, dvs summen av elementene i hver kolonne er 1 og alle elementene er ≥ 0.

5. Er følgende stokastiske matriser regulære? Begrunn svaret ditt.

a) $P = \begin{bmatrix} 0.2 & 1 \\ 0.8 & 0 \end{bmatrix}$

Løsning: P er regulær fordi

$$P^2 = \begin{bmatrix} 0.84 & 0.2 \\ 0.16 & 0.8 \end{bmatrix}.$$

b) $Q = \begin{bmatrix} 1 & 0.2 \\ 0 & 0.8 \end{bmatrix}$

Løsning: Q er ikke regulær fordi

$$Q^n = \begin{bmatrix} 1 & 1 - 0.8^n \\ 0 & 0.8^n \end{bmatrix}$$

har alltid en 0 i det nederste venstre hjørnet.

6. (Høst 2020, oppgave 6) For å simulere proteinene som danner taggene til koronaviruset SARS-CoV-2 bruker forskere overgangsalgoritmer (for eksempel i prosjektet Folding@Home ved Stanford). De trenger din hjelp til å teste om algoritmen de har implementert fungerer som den skal, og har derfor forenklet utregningene slik at du kan gjøre dem for hånd.

Proteinet S kan innta tre ulike tilstander, A, B og C. Hvert mikrosekund, altså hvert 10^{-6}-sekund, kan den bytte tilstand.

- Gitt at S er i tilstand A, og man venter et mikrosekund så er det: 2/3 sannsynlighet for at S fortsatt er i tilstand A, 1/3 sannsynlighet for at S går til tilstand B, 0 sannsynlighet for at S går til tilstand C.
- Gitt at S er i tilstand B, og man venter et mikrosekund så er det: 0 sannsynlighet for at S går til A, 3/4 sannsynlighet for at S fortsatt er i tilstand B, 1/4 sannsynlighet for at S går til tilstand C.
- Gitt at S er i tilstand C, og man venter et mikrosekund så er det: 1/3 sannsynlighet for at S går til A, 1/3 sannsynlighet for at S går til tilstand B, 1/3 sannsynlighet for at S fortsatt er i tilstand C.

a) Finn den stokastiske matrisen M som representerer hvordan S endrer tilstander for hvert mikrosekund.

Løsning:

Prosessen kan beskrives ved den stokastiske matrisen

$$M = \begin{bmatrix} \frac{2}{3} & 0 & \frac{1}{3} \\ \frac{1}{3} & \frac{3}{4} & \frac{1}{3} \\ 0 & \frac{1}{4} & \frac{1}{3} \end{bmatrix}.$$

Her er første rad sjansen for tilstand A, andre rad for B og tredje rad for C, og første kolonne gjelder dersom S var i tilstand A mikrosekunden før, andre kolonne dersom det var i tilstand B mikrosekunden før og tredje kolonne dersom det var il tilstand C mikrosekunden før.

b) Gitt at S er i tilstand A, hva er sannsynligheten for at S er i tilstand B to (2) mikrosekunder senere?

Løsning:

Den første kolonnen i M gir oss sannsynlighetene for å gå videre fra A i ett mikrosekund. Andre raden i M gir oss sannsynlighetene for at S er i tilstand B etter at det har vært i tilstandene A, B eller C etter ett mikrosekund. Det er altså flere sannsynligheter vi må gange med hverandre og så legge sammen. Matriseproduktet M^2 gjør akkurat det for oss. Den samler opp alle produkter av sannsynlighetene for å komme til et tilstand til et annet etter to mikrosekunder. Vi finner altså sannsynligheten som andre element in den første kolonnen i M^2: $\frac{2}{9} + \frac{3}{12} = \frac{17}{36}$.

c) Forskerene simulerer at proteinet S får endre tilstand fritt, slik det vil, i et helt sekund. Estimer sannsynligheten for at S er i hver av de 3 tilstandene på slutten av sekundet. Hvilken tilstand er det mest sannsynlig at S vil være i?

Løsning:

Et sekund består av en million mikrosekunder. Det betyr at vi er interesserte i tilstanden til S i det lange løp. Likevektsvektoren **v** til M vil altså beskrive denne tilstanden. Derfor vil vi løse systemet med totalmatrisen

$$M - I_3 = \begin{bmatrix} \frac{2}{3} - 1 & 0 & \frac{1}{3} \\ \frac{1}{3} & \frac{3}{4} - 1 & \frac{1}{3} \\ 0 & \frac{1}{4} & \frac{1}{3} - 1 \end{bmatrix} = \begin{bmatrix} -\frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{3} & -\frac{1}{4} & \frac{1}{3} \\ 0 & \frac{1}{4} & -\frac{2}{3} \end{bmatrix}$$
$$\sim \begin{bmatrix} -\frac{1}{3} & 0 & \frac{1}{3} \\ 0 & -\frac{1}{4} & \frac{2}{3} \\ 0 & \frac{1}{4} & -\frac{2}{3} \end{bmatrix} \sim \begin{bmatrix} -\frac{1}{3} & 0 & \frac{1}{3} \\ 0 & -\frac{1}{4} & \frac{2}{3} \\ 0 & 0 & 0 \end{bmatrix}.$$

Dette viser at koordinatene v_1, v_2, v_3 til $\mathbf{v}$ oppfyller altså

$$v_2 = \frac{3}{8} v_3 \text{ og } v_1 = v_3.$$

Det betyr at $\mathbf{v}$ er på formen

$$\mathbf{v} = t \begin{bmatrix} 1 \\ \frac{8}{3} \\ 1 \end{bmatrix}.$$

I en likevektsvektor er summen av elementene lik 1. Vi får altså

$$\mathbf{v} = \frac{3}{14} \begin{bmatrix} 1 \\ \frac{8}{3} \\ 1 \end{bmatrix}.$$

Nå må vi se hvilken av koordinatene til $\mathbf{v}$ er størst og vi ser at tilstand B er mest sannsynlig.