35 - INTERPOLASJON OG REGRESJON

Dersom du har skjønt alt i forrige økt, blir det litt enklere å huske og forstå hvordan lineærregresjon funker. I sannsynlighetsregning pakkes dette inn i mye komplisert notasjon, og statistikerne må gjøre det slik for å bygge opp alt systematisk. Men selve matematikken er ikke så komplisert.

La oss begynne med noe som kalles **interpolasjon**. Hvis du har $n+1$ punkter (x_i, y_i) i $\mathbb{R}^2$, der $x_i \neq x_j$ for alle $i \neq j$, vil det alltid være mulig å finne et reelt polynom

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + ... + a_1 x + a_0$$

hvis graf går gjennom alle disse punktene, altså at

$$p(x_i) = y_i$$

for alle $1 \leq i \leq n+1$. DIsse likningene utgjør et $(n+1) \times (n+1)$-likningssystem for koeffisientene a_i med totalmatrise

$$\left[\begin{array}{ccccc|c} x_1^n & x_1^{n-1} & ... & x_1 & 1 & y_1 \\ x_2^n & x_2^{n-1} & ... & x_2 & 1 & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{n+1}^n & x_{n+1}^{n-1} & ... & x_{n+1} & 1 & y_{n+1} \end{array}\right]$$

Dette systemet kalles vandermondesystemet, og vi slipper faktisk å plages med det, for vi kan heller gjøre slik: For hvert punkt x_i, definerer vi et polynom

$$l_i(x) = \prod_{\substack{k=0 \\ k \neq i}}^{n} \frac{(x - x_k)}{(x_i - x_k)}.$$

1 Sjekk at polynomet $l_i(x)$ har orden n, og at det tilfredsstiller

$$l_i(x_k) = \begin{cases} 1 & \text{for } i = k \\ 0 & \text{for } i \neq k \end{cases}.$$

Hvordan tror du vi skal få bruk for disse polynomene?

Det er lett å se at

$$p_n(x) = \sum_{i=0}^{n} y_i l_i(x).$$

tilfredsstiller $p_n(x_i) = y_i$ for alle i. Polynomene l_i kalles lagrangepolynomer, og alt dette kalles Lagranges interpolasjonsmetode. Det er også klart at vi alltid kan finne et entydig interpolasjonspolynom dersom $x_i \neq x_j$ for alle $i \neq j$. [1]

2 Finn et annengradspolyom som går gjennom punktene

$$\begin{bmatrix}1\\4\end{bmatrix}, \begin{bmatrix}2\\5\end{bmatrix} \quad \text{og} \quad \begin{bmatrix}3\\6\end{bmatrix}.$$

Polynominterpolasjon ligger til grunn for veldig mye forskjellig. En klassiker er numerisk integrasjon. Den enkleste formen for numerisk integrasjon er å beregne riemannsummer. Men med polynominterpolasjon kan vi gjøre det mye bedre. La oss si at vi ønsker å finne en approksimasjon til

$$\int_a^b f(x)\, dx.$$

Dette kan gjøres ved å dele inn $[a, b]$ i et ekvidistant gitter med $a = x_0$ og $b = x_n$, skrive

$$\int_a^b f(x)\, dx \approx \int_a^b p(x)\, dx = f(x_i) \sum_{i=0}^{n} \int_a^b l_i(x)\, dx$$

og så beregne

$$\int_a^b l_i(x)\, dx$$

for alle i. For $n = 1$ får man trapesregelen

$$\int_a^b f(x)\, dx \approx \frac{b-a}{2}\left(f(a) + f(b)\right)$$

`https://en.wikipedia.org/wiki/Trapezoidal_rule`
og for $n = 2$ får man noe som kalles Simpsons metode:

$$\int_a^b f(x)\, dx \approx \frac{b-a}{6}\left(f(a) + 4f\left(\frac{b+a}{2}\right) + f(b)\right)$$

`https://en.wikipedia.org/wiki/Simpson%27s_rule`
og så videre. Dersom gitteret ikke er ekvidistant, kan man lage andre numeriske integrasjonsrutiner, for eksempel gausskvadratur:
`https://en.wikipedia.org/wiki/Gaussian_quadrature`
eller clenshawcurtiskvadratur:
`https://en.wikipedia.org/wiki/Clenshaw-Curtis_quadrature`

3 Bruk Simpson på punktene i oppgave 2.

[1] Hvis vi antar at det finnes to forskjellige polynomer p_n og q_n av grad n som interpolerer den samme tabellen, og evaluerer differansen $p_n - q_n$ i punktene x_i, ser vi at

$$p_n(x_i) - q_n(x_i) = 0 \qquad 0 \leq i \leq n.$$

Polynomet $p - q$ har maksimal grad n, og kan maksimalt ha n nullpunkter, så derfor må $p = q$.

Dersom man prøver å interpolere et datasett med et polynom som har for lav orden, vil man få det overbestemte $(n+1)\times(m+1)$-systemet

$$\left[\begin{array}{ccccc|c} x_1^m & x_1^{n-1} & \dots & x_1 & 1 & y_1 \\ x_2^m & x_2^{n-1} & \dots & x_2 & 1 & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{n+1}^m & x_{n+1}^{n-1} & \dots & x_{n+1} & 1 & y_{n+1} \end{array}\right]$$

Dette kalles **regresjon**.

4 Prøv å finne et annengradspolyom som går gjennom punktene

$$\begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}2\\1\end{bmatrix} \quad \text{og} \quad \begin{bmatrix}3\\2\end{bmatrix}.$$

Likningssystemet i oppgaven over ble

$$\begin{aligned} c &= 1 \\ a + b + c &= 0 \\ 4a + 2b + c &= 1 \\ 9a + 3b + c &= 2 \end{aligned}$$

Hvordan fikse dette? Nå trenger vi noe som kalles minste kvadraters metode. Dette er en teknikk for å finne tilnærmede løsninger til lineære systemer med flere likninger enn ukjente. La oss si at A er en $m \times n$-matrise, at $\mathbf{x}$ og $\mathbf{y}$ er kolonnevektorer i $\mathbb{C}^n$, og at vi ønsker å betrakte systemet

$$A\mathbf{x} = \mathbf{y}$$

for $m > n$. Dette systemet vil ikke ha noen løsning med mindre $\mathbf{y}$ tilfeldigvis ligger i kolonnerommet til A, så vi ønsker istedet å finne den $\mathbf{x}$ som minimerer avstanden fra $A\mathbf{x}$ til $\mathbf{y}$. Det gjør vi ved å kreve

$$A^*(A\mathbf{x} - \mathbf{b}) = \mathbf{0}$$

altså at

$$A^*A\mathbf{x} = A^*\mathbf{y}.$$

5 Hvorfor det?

Dette er et $n \times n$-system som kalles **normallikningene**, og systemet har entydig løsning om kolonnene til A er lineært uavhengige: Løsningen av systemet gir den $\mathbf{x}$ som minimerer avstanden fra $A\mathbf{x}$ til $\mathbf{y}$.

6 Bli ferdig med oppgave 4.

Når statistikere snakker om regresjonslinje, mener de alt dette og at man bruker et førsteordens polynom $ax + b$ og at man bruker minste kvadraters metode for å finne koeffisientene a og b. Økonomer kaller det trendlinje.

7 Finn regresjonslinjen til punktene

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad \text{og} \quad \begin{bmatrix} 3 \\ 2 \end{bmatrix}.$$

8 Skriv en rutine som tar inn et datasett og produserer trendlinjen.

UKENS NØTTER

t (s)	h (cm)
0	48
10	38
20	28
30	20
40	14
50	8

Jeg var på en hytte i Kragerø i helgen og badet i en sirkulær badestamp med et sirkulært høl i bånn. Når jeg skulle tømme den, målte jeg vannstanden etterhvert som stampen tømtes, og fikk tabellen i margen.

I følge boken skal vannstanden følge Torricellis lov, som sier at $\dot{h} = a\sqrt{h}$.

9 Estimer a.

Det går også an å modellere tømming av tank med modellen $\dot{h} = ah + b$.

10 Estimer a. Hvilken modell passer best?

Ifølge boken skal egentlig

$$\dot{h} = -\sqrt{\frac{2g}{b^2 - 1}h}$$

der g er tyngdeakselerasjonen og b er forholdet mellom arealet til badestampen og arealet til hølet: `https://en.wikipedia.org/wiki/Torricelli%27s_law`

11 Stampen hadde radius på ca en meter, men jeg glemte å måle radien på hølet. Hva var den?

Figur 1: Elsys Torjus koser seg i badestampen

Figur 2: Elsys Torjus hopper i havet