Seksjon 6.6: Maskinlæring og lineære modeller

Anta at vi er gitt punkter $(x_1, y_1), \dots, (x_n, y_n)$, og vil finne linjen $\underbrace{y = \beta_0 + \beta_1 x}_{\text{lineær modell}}$ som "best tilpasser" disse punktene

- $\{y_j\}$ kalles observerte verdier. $\vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ kalles en observasjonsvektor
- $\{\beta_0 + \beta_1 x_j\}$ kalles predikerte verdier for y_j
- Differansene $y_j - (\beta_0 + \beta_1 x_j)$ kalles residualer.

"best tilpasser": Det som gir minst mulig avvik mellom observerte og predikerte verdier:

$$\sum_j \left(y_j - (\beta_0 + \beta_1 x_j) \right)^2$$

$$= \left\| \vec{y} - \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \right\|^2$$

komponent j: $\begin{bmatrix} 1 & x_j \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \beta_0 + \beta_1 x_j$

Med andre ord: $\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ er minste kvadraters løsningen til

$$\underbrace{\begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}}_{\vec{\beta}} = \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{\vec{y}}$$

Konsepter:
- Linja $\beta_0 + \beta_1 x$ kalles en regresjonslinje, eller minste kvadraters linje
- β_0, β_1 kalles regresjonskoeffisienter, eller parametre
- $\begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$ kalles designmatrisen

Oppgave 2 Finn minste kvadraters linjen som best tilpasser punktene $(1,0)$, $(2,2)$, $(3,7)$, $(4,9)$.

Løsning: Observasjonsvektoren: $\vec{y} = \begin{bmatrix} 0 \\ 2 \\ 7 \\ 9 \end{bmatrix}$

Designmatrisen: $X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$

$$X^T X = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix} \qquad X^T \vec{y} = \begin{bmatrix} 18 \\ 61 \end{bmatrix}$$

$$[\,X^T X \quad X^T\vec{y}\,] = \begin{bmatrix} 4 & 10 & 18 \\ 10 & 30 & 61 \end{bmatrix} \sim \begin{bmatrix} 1 & 5/2 & 9/2 \\ 10 & 30 & 61 \end{bmatrix}$$

$$\sim \begin{bmatrix} 1 & 5/2 & 9/2 \\ 0 & 5 & 16 \end{bmatrix} \sim \begin{bmatrix} 1 & 5/2 & 9/2 \\ 0 & 1 & 16/5 \end{bmatrix}$$

$$\sim \begin{bmatrix} 1 & 0 & -7/2 \\ 0 & 1 & \underbrace{16/5}_{\beta} \end{bmatrix}$$

minste kvadraters linjen blir $\underline{\underline{y = -\frac{7}{2} + \frac{16}{5}x}}$

minste kvadraters feil blir $\|\vec{y} - X\vec{\beta}\| \approx 1.3416$ ■

Det er ikke sikkert observasjonene våre lar seg tilnærme bra med en rett linje.

La oss prøve i stedet å finne $\beta_0, \beta_1, \beta_2$ s.a.

$$y_j \approx \beta_0 + \beta_1 x_j + \beta_2 x_j^2 .$$

mer generelt, finn $\beta_0, \beta_1, \dots, \beta_k$ s.a.

$$y_j \approx \beta_0 + \beta_1 x_j + \dots + \beta_k x_j^k$$

Sum av kvadrater av residualer:

$$\sum_j \left(y_j - \left(\beta_0 + \beta_1 x_j + \dots + \beta_k x_j^k \right) \right)^2$$

$$= \sum_j \left(y_j - \begin{bmatrix} 1 & x_j & \cdots & x_j^k \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \right)^2$$

$$\left\| \vec{y} - \underbrace{\begin{bmatrix} 1 & x_1 & \cdots & x_1^k \\ 1 & x_2 & \cdots & x_2^k \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^k \end{bmatrix}}_{\text{ny designmatrise}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}}_{\text{regresjonskoeffisienter/parametre}} \right\|$$

Oppgave 2 igjen Finn minste kvadraters parabelen som ~~best~~ tilpasser punktene.

Løsning: Nå får vi

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} 0 \\ 2 \\ 7 \\ 9 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 4 & 10 & 30 \\ 10 & 30 & 100 \\ 30 & 100 & 354 \end{bmatrix} \qquad X^T \vec{y} = \begin{bmatrix} 18 \\ 61 \\ 215 \end{bmatrix}$$

I Matlab: $X^T X \setminus X^T \vec{y}$ gir resultat $\hat{\vec{x}} = \begin{bmatrix} -7/2 \\ 16/5 \\ 0 \end{bmatrix}$

minste kvadraters parabel $y = \beta_0 + \beta_1 x + \beta_2 x^2 \Rightarrow y = -\frac{7}{2} + \frac{16}{5} x + \cancel{0x^2}$

Fikk samme som for minste kvadraters linjen!

Hva med minste kvadraters tredjegradstilnærming for oppgave 2 ?

Ny designmatrise : $X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \\ 1 & 4 & 16 & 64 \end{bmatrix}$

$X^T X \setminus X^T \vec{y}$ gir nå $\vec{\beta} = \begin{bmatrix} 7 \\ -27/2 \\ 15/2 \\ -1 \end{bmatrix} \begin{array}{l} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{array}$

regresjonspolynomet : $7 - \frac{27}{2}x + \frac{15}{2}x^2 - x^3$

Minste kvadraters feil: $\|\vec{y} - X\beta\| = 0$.

Hvis man istedet vil finne en minste kvadraters tilnærming på formen $A\cos x + B\sin x$:

Vi får designmatrisen $X = \begin{bmatrix} \cos x_1 & \sin x_1 \\ \cos x_2 & \sin x_2 \\ \vdots & \vdots \\ \cos x_n & \sin x_n \end{bmatrix}$, regresjonskoeffisienter $\begin{bmatrix} A \\ B \end{bmatrix}$

Se oppgave 6.6.15 til neste uke.