729G78 Artificiell intelligens
Maskininlärning – Linjär regression

Robin Keskisärkkä

HCS/IDA

LINKÖPINGS UNIVERSITET

Val av modell

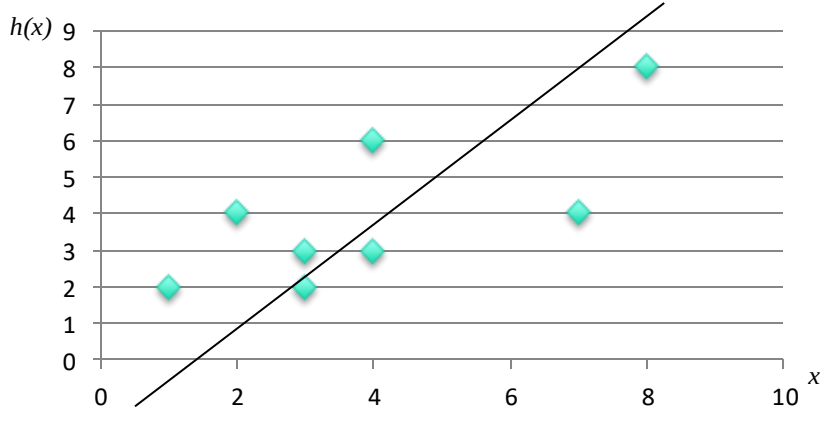
- Välj den modell som är bäst för osedd data
  - Förutsätter att testdata kommer från samma fördelning som träningsdata
  - Bästa modellen har minst andel fel (*error rate)*
- Dela data i tre delar
  - Träningsdata
  - Valideringsdata (*validation set).* Används för att justera och optimera modellen under träning.
  - Testdata

Effektiv användning av träningsdata

- Minimera risk för överträning (overfitting) och underträning (underfitting)
- *K-fold cross-validation*
  - Dela tränings- och valideringsdata i K delar
  - En del används som valideringsdata medan resterande används som träningsdata.
  - Upprepa K-gånger där all data används för validering en gång
- Mindre bias genom att använda flera valideringsuppsättningar

Linjär regression

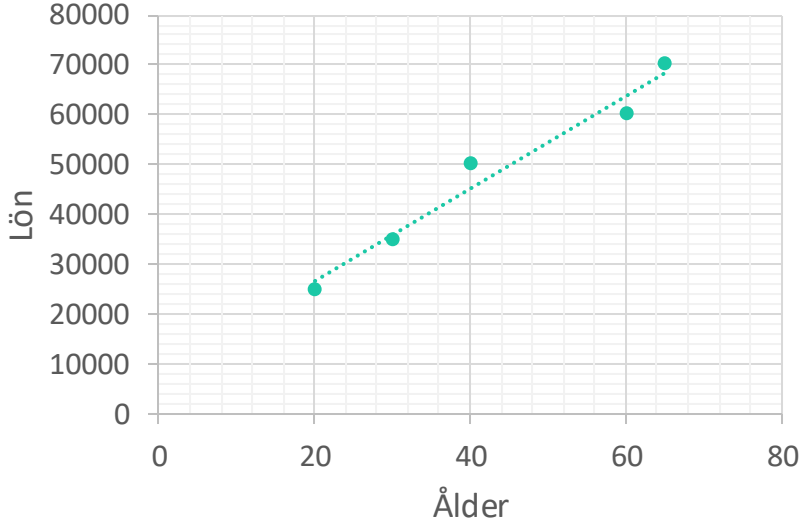
- Hitta den linje som bäst beskriver en datamängd
  - Modellantagande: sambandet kan beskrivas med en rät linje
- $h(x) = w_1 x + w_0$ där w_1 anger lutningen och w_0 förskjutningen från origo
- Jämför avvikelsen mot linjen i varje punkt och minimera felet



OBS: Räta linjens funktion: $y = kx + m$

Exempel

Ålder	Lön
20	25000
30	35000
40	50000
60	60000
65	70000



Minimera förlust

- Vill inte bara minska felet
  - Ex. värre att klassificera icke-spam som spam än tvärtom
- Vill maximera nyttan, minimera förlusten (*loss*)
  - $L(x, t, h(x))$ ger differensen i ”nytta” mellan att välja korrekt svar t och det predicerade svaret $h(x)$ utifrån indata x
  - Förenklas ofta till $L(t, h(x))$
    Ex. $L(spam, icke\text{–}spam) = 1$, $L(icke\text{–}spam, spam) = 10$
- Minimera *loss* över alla par av korrekt värde/ predicerat värde

Felfunktioner (*loss functions*)

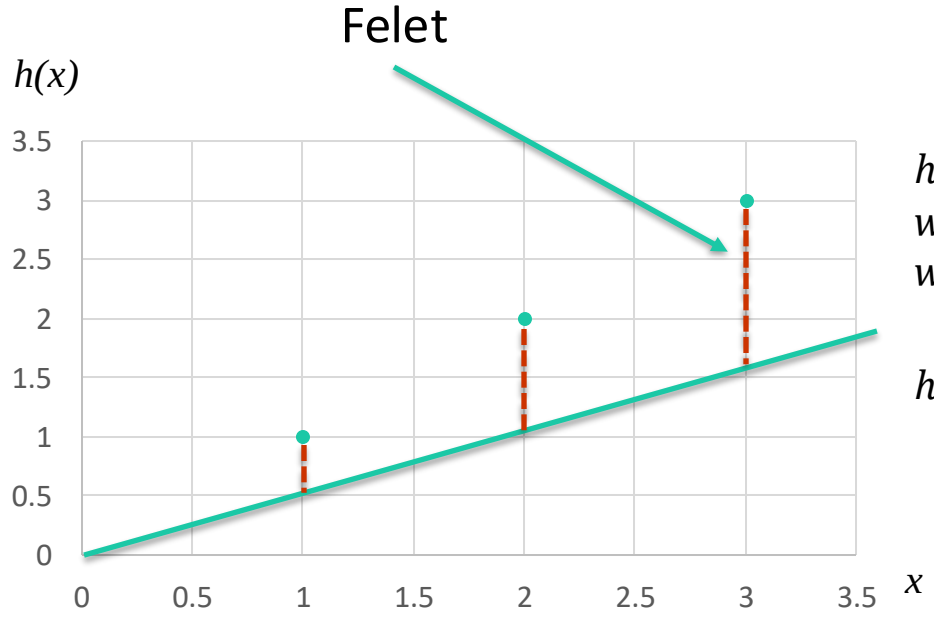
- Finns flera sätt att räkna ut felet mellan det korrekta värdet t och värdet $h(x)$
  - Absolutvärdet, L_1: $\sum_{i=1}^{N} |ti - h(x_i)|$
  - Kvadratfelet, L_2: $\sum_{i=1}^{N} (ti - h(xi))^2$ ← Vanligast. Stora fel betraktas som värre än små och negativa fel kan inte ta ut positiva.
  - Värsta fallet: $\max_i |t_i - h(x_i)|$

Förlusten (*loss*)

$t_1 = 1$
$h(x_1) = 0{,}5$

$t_2 = 2$
$h(x_2) = 1$

$t_3 = 3$
$h(x_3) = 1{,}5$



$h(x) = w_1 x + w_0$
$w_0 = 0$
$w_1 = 0{,}5$

$h(x) = 0{,}5x$

Kvadratfelet, L_2: $(1 - 0{,}5)^2 + (2 - 1)^2 + (3 - 1{,}5)^2 = 0{,}25 + 1 + 2{,}25 = 3{,}5$

Felfunktionen för L_2

Kvadratfelet blir större ju mer träningsdata som används, medelkvadratfel (MSE)

$$J(\mathrm{w}_0, \mathrm{w}_1) = \frac{1}{N}\sum_{i=1}^{N}(t_i - h(x_i))^2$$

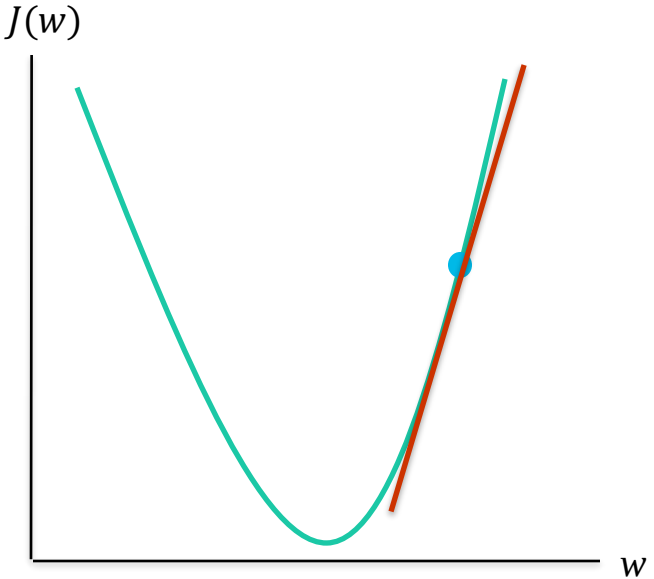
där t_i är korrekt värde och $h(x_i) = w_1 x_i + w_0$

Felfunktionen för L_2

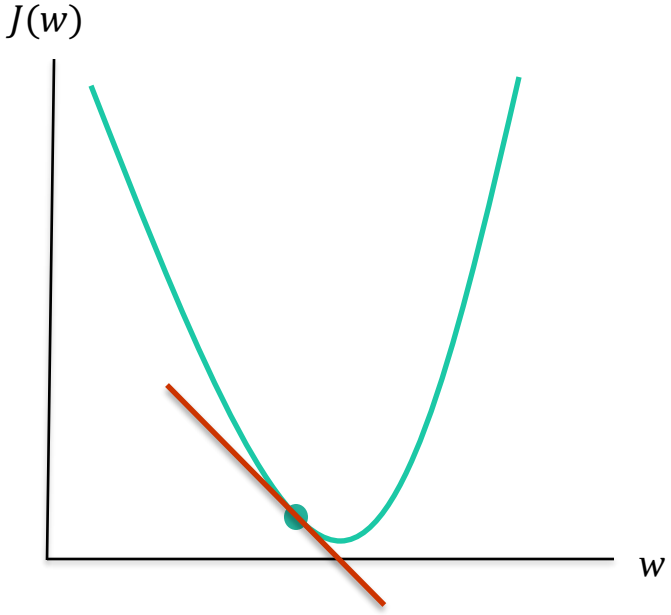
$$J(\mathrm{w}_0, \mathrm{w}_1) = \frac{1}{N}\sum_{i=1}^{N}(t_i - h(x_i))^2$$

- Felfunktionen $J(\mathrm{w}_0, \mathrm{w}_1)$ ger oss ett enkelt sätt att teckna felet för olika värden på parametrarna w_0 och w_1
- Målet är att justera parametrarna så att felet blir så litet som möjligt
- En vanlig metod är gradientsökning

Gradientsökning, intuition



$w = w - \text{stort värde}$



$w = w + \text{litet värde}$

Kurvan beskriver felfunktionen. Lutningen säger hur vi borde justera värdet på w

Gradientsökning

Steg 0: Börja med ett godtyckligt värde för w

Steg 1: Räkna ut felfunktionens lutning (derivatan) i den punkten

Steg 2: Gå i lutningens motsatta riktning:

- Om tangenten har positiv lutning, minska värdet på w
- Om tangenten har negativ lutning, höj värdet på w
- Detalj: Storleken på justeringen är lutningen multiplicerat med en inlärningsparameter

Upprepa steg 1 och 2 tills felet blir tillräckligt litet

Derivata, recap

- Derivatan mäter hur en funktion förändras med avseende på en variabel
- För en funktion $f(x)$ definieras derivatan som:

$$f'(x) = \lim_{h \to 0} \left(\frac{f(x+h) - f(x)}{h} \right)$$

- Om $y = f(x)$ kan vi också skriva $\frac{dy}{dx}$ eller $\frac{df(x)}{dx}$

”Hur förändras $f(x)$ med avseende på x i en viss punkt”

Partiell derivata

- Idén om derivata kan även tillämpas om man har flera variabler genom att man håller vissa variabler konstanta
- Låt $f(x, y) = x^2 + y^3$
  - Partiella derivatan med avseende på x:
    $\frac{\partial f(x,y)}{\partial x} = 2x$
  - Partiella derivatan med avseende på y:
    $\frac{\partial f(x,y)}{\partial y} = 3y$
- Det innebär att vi kan justera lutningen för flera variabler separat!

Gradienten

Justera w_0 och w_1 i gradientens riktning

$$J(\mathrm{w}_0, \mathrm{w}_1) = \frac{1}{N}\sum_{i=1}^{N}(t_i - h(x_i))^2 \qquad \text{där } h(x_i) = w_1 x_i + w_0$$

$$\frac{\partial J(w_0, w_1)}{\partial w_0} = \frac{1}{N}\sum_{i=1}^{N} -2(t_i - h(x_i))$$

$$\frac{\partial J(w_0, w_1)}{\partial w_1} = \frac{1}{N}\sum_{i=1}^{N} -2x_i(t_i - h(x_i))$$

Parameteruppdatering

Inlärningsparametern η (eta) avgör hur snabbt vi uppdaterar w.

$$w_0 = w_0 + \eta \frac{1}{N} \sum_{i=1}^{N} (t_i - h(x_i))$$

$$w_1 = w_1 + \eta \frac{1}{N} \sum_{i=1}^{N} x_i (t_i - h(x_i))$$

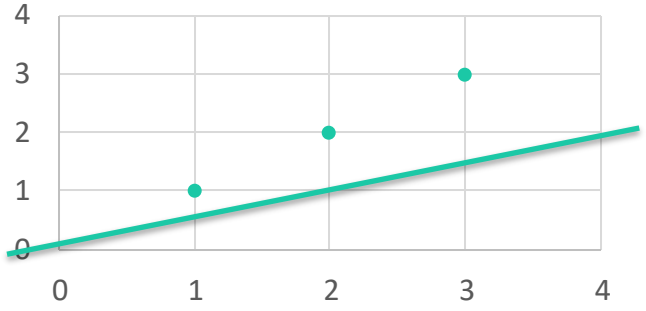
OBS: Konstanten 2 i formeln på föregående bild fångas här i η eftersom vi kan välja hur vi vill justera vikterna genom inlärningsparametern.

Exempel

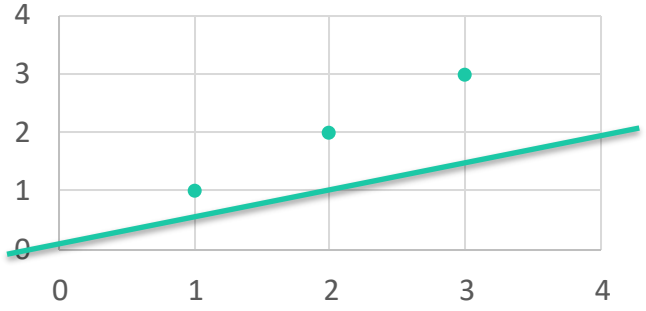
$\eta = 0{,}1, w_0 = 0$
dvs $h(x) = w_1 x$

Initialt $w_1 = 0{,}5$

$$w_1 = w_1 + \eta \frac{1}{N} \sum_{i=1}^{N} x_i (t_i - h(x_i))$$



Exempel



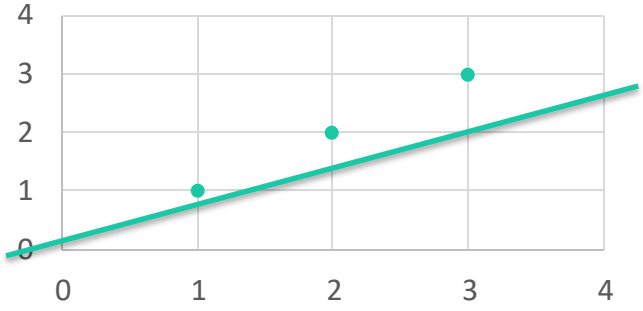
$\eta = 0{,}1, w_0 = 0$
dvs $h(x) = w_1 x$

Initialt $w_1 = 0{,}5$

$$w_1 = w_1 + \eta \frac{1}{N} \sum_{i=1}^{N} x_i (t_i - h(x_i))$$

$$w_1 = 0{,}5 + 0{,}1 \frac{1}{3} \big(1(1 - 0{,}5) + 2(2 - 1) + 3(3 - 1{,}5)\big) = 0{,}733$$

Exempel



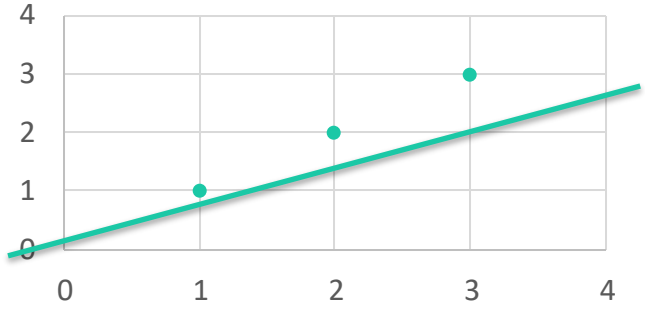
$\eta = 0{,}1, w_0 = 0$
dvs $h(x) = w_1 x$

Initialt $w_1 = 0{,}5$

$$w_1 = w_1 + \eta \frac{1}{N} \sum_{i=1}^{N} x_i (t_i - h(x_i))$$

$$w_1 = 0{,}5 + 0{,}1 \frac{1}{3} \big(1(1 - 0{,}5) + 2(2 - 1) + 3(3 - 1{,}5)\big) = 0{,}733$$

Exempel



$\eta = 0{,}1, w_0 = 0$
dvs $h(x) = w_1 x$

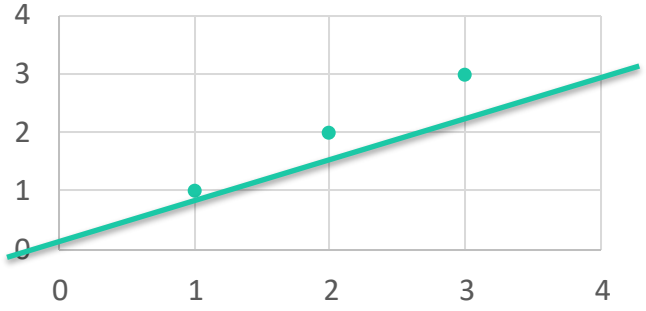
Initialt $w_1 = 0{,}5$

$$w_1 = w_1 + \eta \frac{1}{N} \sum_{i=1}^{N} x_i (t_i - h(x_i))$$

$$w_1 = 0{,}5 + 0{,}1 \frac{1}{3} \big(1(1 - 0{,}5) + 2(2 - 1) + 3(3 - 1{,}5)\big) = 0{,}733$$

$$\mathrm{w}_1 = 0{,}733 + 0{,}1 \frac{1}{3} \big(1(1 - 0{,}733) + 2(2 - 1{,}466) + 3(3 - 2{,}199)\big) = 0{,}857$$

Exempel



$\eta = 0{,}1, w_0 = 0$
dvs $h(x) = w_1 x$

Initialt $w_1 = 0{,}5$

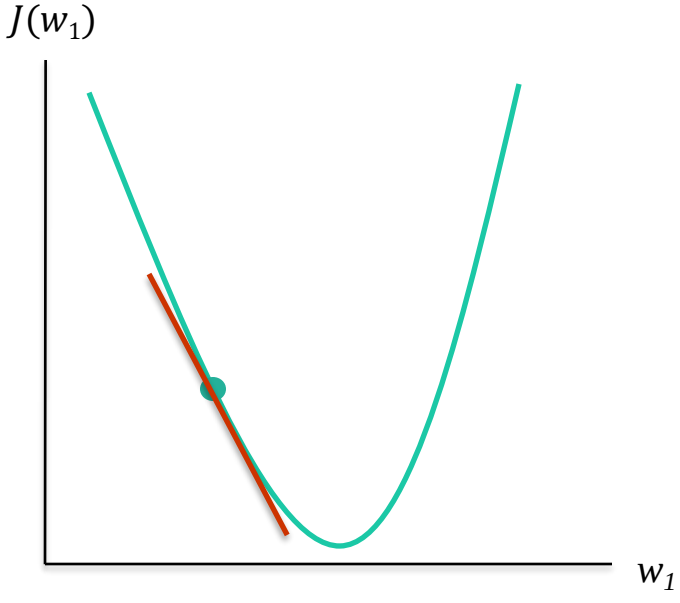
$$w_1 = w_1 + \eta \frac{1}{N} \sum_{i=1}^{N} x_i (t_i - h(x_i))$$

$$w_1 = 0{,}5 + 0{,}1 \frac{1}{3} \big(1(1 - 0{,}5) + 2(2 - 1) + 3(3 - 1{,}5)\big) = 0{,}733$$

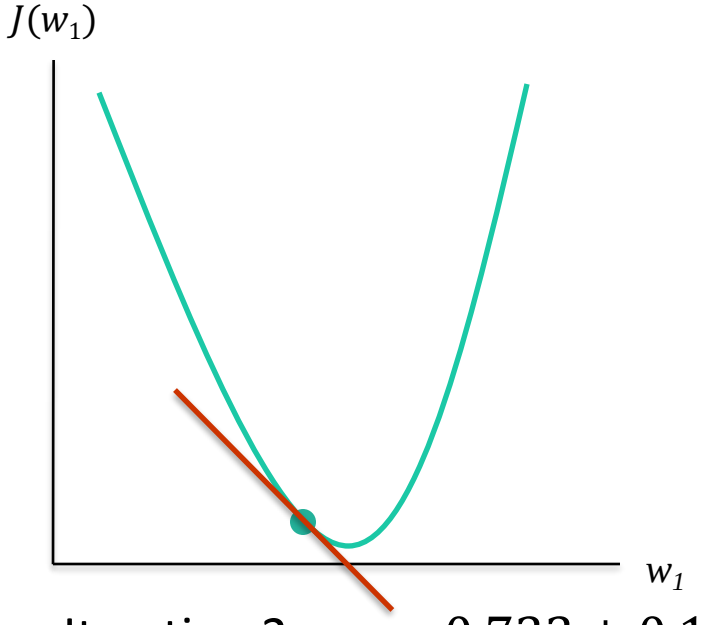
$$\mathrm{w}_1 = 0{,}733 + 0{,}1 \frac{1}{3} \big(1(1 - 0{,}733) + 2(2 - 1{,}466) + 3(3 - 2{,}199)\big) = 0{,}857$$

…

Gradientsökning



Iteration 1: $w_1 = 0{,}5 + 0{,}233$



Iteration 2: $w_1 = 0{,}733 + 0{,}124$

Regularisering

- Används för att undvika overfitting
- Minskar generaliseringsfelet utan att träningsfelet blir mindre
- Exempel: L_2-regularisering
  Lägg till en extra term till felfunktionen som blir större ju mer komplex funktionen h är

$$\sum_{i=1}^{N}(t_i - h(x_i))^2 + \lambda Complexity(h)$$

λ är en så kallad *hyperparameter* som tränas till att balansera kostnaden.

En hyperparameter är helt enkelt en parameter som styr hur maskininlärningsmodellen tränas. η (inlärningsparametern) räknas också som en hyperparameter.

Multipel linjär regression

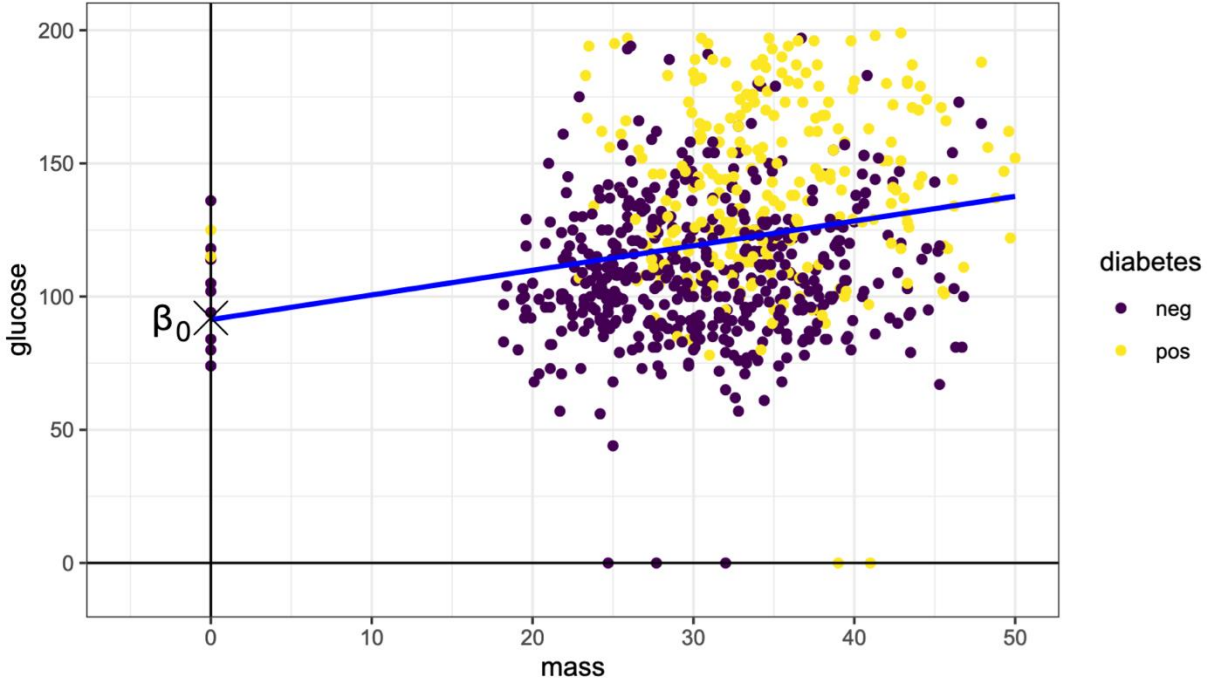
- Linjär regression: $h(x) = w_1 x + w_0$
- Multipel linjär regression: $h(x) = w_1 x_1 + w_1 x_1 + \cdots + w_n x_n + w_0$

$$h(x_1, \ldots, x_n) = \sum_{i=0}^{n} \mathrm{w}_i x_i$$

- Gradientsökningen kan generaliseras till att stödja flera särdrag x och parametrar *w*
- Sambandet mellan indata och utdata blir nu ett *hyperplan*, dvs. yta som delar ett *n*-dimensionellt rum

Exempel: Hyperplan

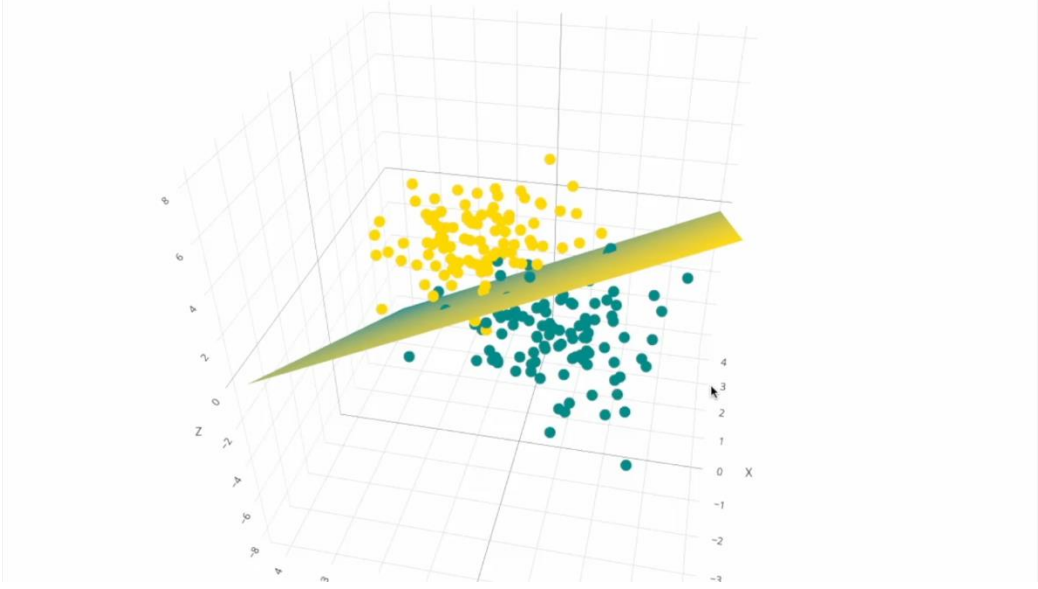
Glukosnivå som en funktion av vikt, svagt positiv lutning



Källa: https://science.nu/amne/multipel-regression/

Exempel: Hyperplan

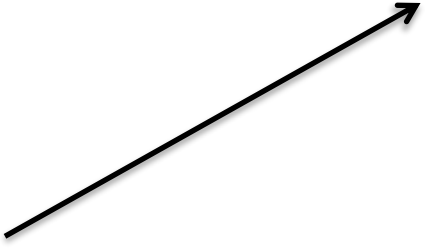
Fler särdrag, kraftfullare modell



Källa: https://science.nu/amne/multipel-regression/

Vektorer

- Särdrag och parametrar kan skrivas som vektorer, skrivs ofta med fet stil ex. $\boldsymbol{x}$ och $\boldsymbol{w}$
- Skalär
  - Ett tal, ex. längd (2m), tid (12:30), temperatur (170K)
- Vektor
  - Riktning och längd, t.ex. hastighet, kraft

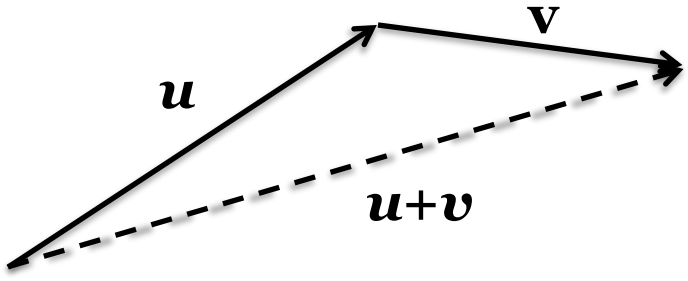


Några vektorsamband

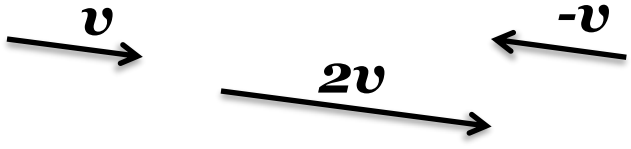
- Likhet

  - Lika lång och med samma riktning men inte nödvändigtvis samma startpunkt

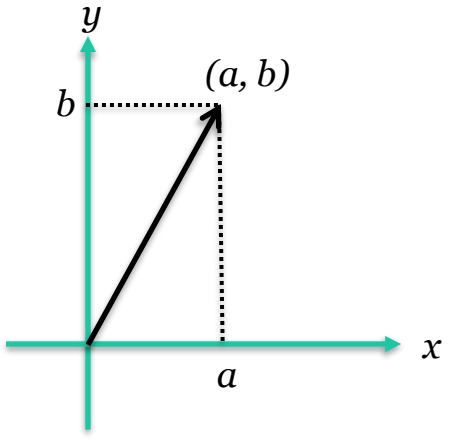
- Addition

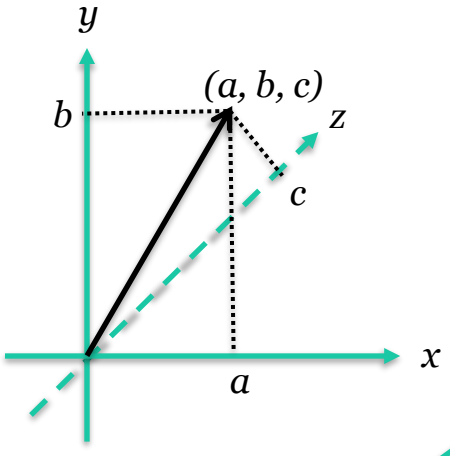


- Multiplikation



Punkter och vektorer





Både parenteser och hakparenteser

Vektorer kan identifieras som punkter: $[a, b]$ och $[a, b, c]$

Kan generaliseras till vilken dimension som helst: $[a_0, a_1, \ldots, a_2]$

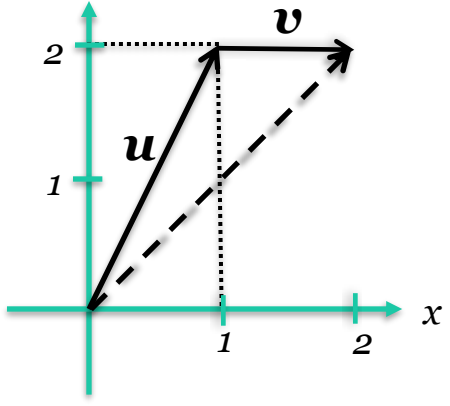
Vektoroperationer

$\boldsymbol{z} = [a_1, b_1, c_1]$
$\boldsymbol{w} = [a_2, b_2, c_2]$

Addition
$\boldsymbol{z} + \boldsymbol{w} = [a_1 + a_2, b_1 + b_2, c_1 + c_2]$

Multiplikation med skalär
$2\boldsymbol{z} = [2a_1, 2b_1, 2c_1]$

$\boldsymbol{u} = [1, 2]$
$\boldsymbol{z} = [1, 0]$
$\boldsymbol{u} + \boldsymbol{z} = [2, 2]$



Skalärprodukt

Produkten av två vektorer ger en skalär (ett tal)

$\boldsymbol{z} = [a_1, b_1, c_1]$

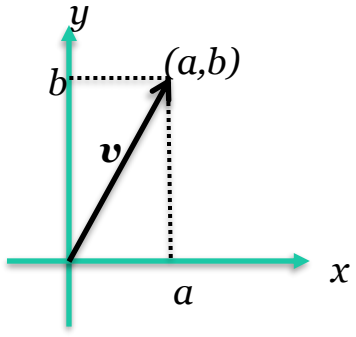
$\boldsymbol{w} = [a_2, b_2, c_2]$

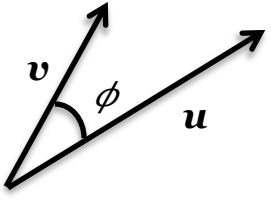
$$\boldsymbol{z} \cdot \boldsymbol{w} = a_1 a_2 + b_1 b_2 + c_1 c_2$$

Skalärprodukten innehåller information om både vektorns längd och riktning

Skalärproduktens geometriska definition

- Längden av vektorn $\boldsymbol{v} = [a, b]$ betecknas $\|\boldsymbol{v}\|$ kan beräknas med Pythagoras sats som $\|\boldsymbol{v}\| = \sqrt{a^2 + b^2}$
- Skalärprodukten
  $\boldsymbol{v} \cdot \boldsymbol{v} = \boldsymbol{aa} + \boldsymbol{bb} = \boldsymbol{a^2} + \boldsymbol{b^2}$
  $\boldsymbol{v} \cdot \boldsymbol{v} = \|\boldsymbol{v}\|^2$ dvs. $\|\boldsymbol{v}\| = \sqrt{\boldsymbol{v} \cdot \boldsymbol{v}}$
- Två vektorer $\boldsymbol{u}$ och $\boldsymbol{v}$ med vinkeln ϕ mellan sig
- Geometrisk definition av skalärprodukt
  $\boldsymbol{u} \cdot \boldsymbol{v} = \|\boldsymbol{u}\|\|\boldsymbol{v}\|\cos(\phi)$





Matriser

- Matriser är rektangulära tabeller med tal
- Smidig representation om man har mycket träningsdata. Ex.
- Exempel $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$ och $B = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}$

- A är en matris med 2 rader och 3 kolumner:
  A är en 2×3-matris
  B är en 3×2-matris
- Matriser multipliceras annorlunda

Matrismultiplikation

- Matrisen $C = AB$ fås genom att multiplicera raden i A med kolumnen i B

$$c_{ij} = \sum_{k=1}^{m} a_{ik}\, b_{kj}$$

- Antalet kolumner i A måste matcha antalet rader i B
- Exempel $\mathrm{A} = \begin{bmatrix} 1 & 3 & 4 \\ -3 & 2 & 0 \end{bmatrix}$ 2 × 3 och $\mathrm{B} = \begin{bmatrix} 2 & 1 \\ 3 & 0 \\ -1 & 1 \end{bmatrix}$ 3 × 2

$$C = \begin{bmatrix} 1 \cdot 2 + 3 \cdot 3 + 4 \cdot -1 & 1 \cdot 1 + 3 \cdot 0 + 4 \cdot 1 \\ -3 \cdot 2 + 2 \cdot 3 + 0 \cdot -1 & -3 \cdot 1 + 2 \cdot 0 + 0 \cdot 1 \end{bmatrix} = \begin{bmatrix} 7 & 5 \\ 0 & -3 \end{bmatrix}$$

Matrismultiplikation, 2

- Matriser kan transponeras
- Exempel $\mathrm{A} = \begin{bmatrix} 1 & 3 & 4 \\ -3 & 2 & 0 \end{bmatrix}$ $A^T = \begin{bmatrix} 1 & -3 \\ 3 & 2 \\ 4 & 0 \end{bmatrix}$
- $\boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}$ $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$
- $h(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} = \begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = [\mathrm{w}_1 x_1 + \mathrm{w}_2 x_2 + \mathrm{w}_3 x_3]$
- Jämför $h(x_1, x_2, x_3) = \sum_{i=1}^{3} \mathrm{w}_i x_i = \mathrm{w}_1 x_1 + \mathrm{w}_2 x_2 + \mathrm{w}_3 x_3$

Multipel linjär regression

- $h(x) = w_1 x_1 + w_1 x_1 + \cdots + w_n x_n + w_0$ kan nu formuleras i termer av vektorer/matriser
- $h(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$
- I flera dimensioner blir gradienten en vektor som innehåller alla lutningar med en lutning per dimension (dvs. en lutning per parameter som ska uppdateras $w_1, w_2, \ldots, w_n$)
- Jämför $w_i = w_i + \eta \frac{\partial J}{\partial w_i}$ där vi bara tittade på w_1
- Flera parametrar $\nabla \mathrm{J}(\boldsymbol{w}) = \left[\frac{\partial}{\partial w_0} \quad \frac{\partial}{\partial w_1} \quad \frac{\partial}{\partial w_2} \quad \ldots\right]$
- Parameteruppdatering $\boldsymbol{w} = \boldsymbol{w} + \eta \nabla \mathrm{J}(\boldsymbol{w})$