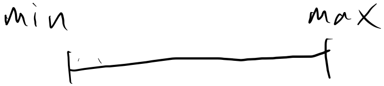
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Vad är en "skala".



Ett sätt: min-max skalering.

Känsligt för extremvärden/outliers.

Ett annat sätt: standardisering

$$\frac{X - \mu}{\sigma}$$

Varför?

Reguljärisering:

ℓ_2-norm: $C(\beta) = RSS + \sum_{i=1}^{d} \beta_i^2$ Ridge-regression

Måste ha β_i på samma skala.

Om '0' skall ha betydelsen "ingenting" så behövs ofta skalering.

Y skaleras oftast <u>inte</u>.

Minsta-kvadrat metoden är ett specialfall av maximum-likelihood metoden.

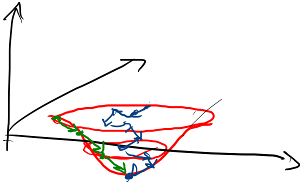
$$\left(\max C(\beta) \approx \prod p \prod 1-p \qquad (\text{en klass}) \right)$$

Då fungerar inte längre formeln för b som vi hade för linjär regression.

Vi måste hitta ett annat sätt att optimera över kostnadsfunktionen!

Gradient descent är lösningen!

steglängd viktigt!



om ytan är "knagglig" så hoppar vi omkring massa (instabil)

SGD - stochastic gradient descent

Välj en slumpmässig punkt! Nu konvergerar inte metoden! Vi måste ha ett stoppvillkor.

Den fastnar dock inte i lokala minimum!

mini-batch Gradient Descent

Ta en slumpmässig dolmängd och räkna gradient.

Iterativ metod!

- stoppvillkor
- varje iteration kallas en epok.

OLS: $O(np^2)$, n - stickprovstorlek

p - antalet dim.

SGD: $O(n)$

$$x_{n+1} = r(1 - x_n)$$

$$x_0 = 0.5$$

Vi behöver justera några parametrar för att SGD skall bli bra.

- Steglängden bör förändras! Adaptive Grad. Descent
- Moment! ADAM