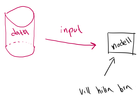
Importance of data quality



Vill hitta bra parametrar som beskriver ett visst samband

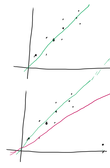
Vad kännetecknar att datan är dålig på något sätt?

Vad betyder dålig data?

- Många nulls
- Ingen sambands/irrelevant data
- Ofullständig data
- Smutsig data
- felaktig data
- OSV

Exempel på hur ett enda pytteliten felaktigt värde i din data kan försämra prestandan av din tränade modell

Anta att vi har en feature x_1 och en target y_1



$$\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$$

Som vi ser här kan ett enda fel skapa en usel modell. Viktigt att åtgärda dessa fel, på något sätt

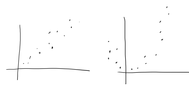
---

Ok, men vilken data vill man samla och träna sin modell på?

Svar: Relevant data!

Du vill ha features som direkt har en påverkan på din önskade target

Det är viktigt att ha features som är relevanta. Ingenting stoppar dig från att även använda irrelevanta features, men detta kommer ha en negativ påverkan på modellens prestanda eftersom att dessa features kommer störa modellen



Features

x_1, x_2, x_3, x_4, x_5 $\quad y$

Du väljer helt och hållet själv vilka features du vill använda för att träna en modell på.

Hur du väljer dina features ska inledningsvis guidas av din intuition och domänexpertis. Du vill endast ha relevanta features inblandade.

Du kan också skapa din egna features!

$$x_6 = x_1 \cdot x_2$$

Ok, jag vill träna en linjär modell med x1, x2 och x6 som features. Hur ser det ut?

$$y = w_3 \cdot x_6 + w_2 \cdot x_2 + w_1 \cdot x_1 + w_0$$

Du kan skapa dina egna features precis hur du vill. Observera, det här är ingen lekstuga utan skapandet av dina features ska ske enlegt någon logik/domän expertis

$$x_7 = x_3 x_4 \sqrt{x_5}$$

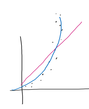
$$x_8 = x_4 + 2$$

Denna process att hitta rätt och bra features kallas FEATURE ENGINEERING

exempel

x_1, y

$y = w_1 \cdot x_1 + w_0$



$$x_2 = x_1 x_1 = (x_1)^2$$

$$y = w_2 x_2 + w_1 \cdot x_1 + w_0 = w_2 (x_1)^2 + w_1 \cdot x_1 + w_0$$

Nu har vi sett hur förödande och jobbig dålig data är.

För att träna bra modeller behöver vi bra data. I många fall behöver vi också mycket (kvantitet), bra data. Men i en hel del fall klarar vi oss undan med en mindre mängd (kvantitet) data, så länge den datan är riktigt bra

kvalitet > kvantitet