Curs 2 Regresie liniară

Continuare + Aplicații

Avem niște date:

(x_i, y_i)

x_i → parametri

y_i → caracteristici care am vrea să le putem prezice din

x_i → inputuri x_i

$x_i \xrightarrow[\text{Program informatic}]{\text{Funcție}} F(x_i) \approx y_i$

Regresia liniară:

$x_i = (X_1, X_2, \dots X_k)$ vector

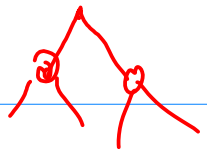
$$y_i \approx b + a_1 X_1 + a_2 X_2 + \dots + a_k X_k$$

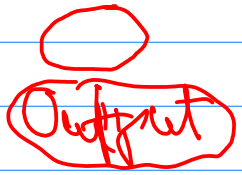
parametrii modelului

Parametrii nu sunt cunoscuți ci se găsesc pornind de la datele care le avem.

Alte tipuri de Modele de Learning

- arbori de decizie
- rețea neuronală





Procesul de învățare: permite să găsim parametrii modelului pornind de la date.

Obiectiv: efectuarea de predicții corecte, utile

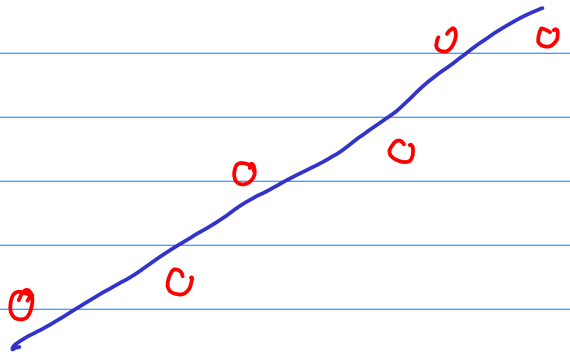
Atenție: nici un model ML nu e exact.

Regresia liniară: cel mai simplu model → utilizează doar o funcție liniară

→ rezultate modeste

Cazul 1D - 2D

(x_i, y_i)
$\in \mathbb{R} \quad \mathbb{R}$



$\hat{y}_i = b + w x_i$

bias, weight

$r_i = y_i - \hat{y}_i$

diferența între predicții și date

Pt a găsi coeficienții w_i mici pt care RSS → residual sum of squares

$$RSS(b, w) = \sum (y_i - b - w x_i)^2$$

date, predicția

Obiectiv: RSS minim.

Formula directă: $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$ (media)

$$w = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad b = \bar{y} - w\bar{x}$$

$\sum (x_i - \bar{x})(y_i - \bar{y})$ → produs scalar (np.dot)

$\langle x - \bar{x}, y - \bar{y} \rangle$

$\sum (x_i - \bar{x})(x_i - \bar{x}) = \langle x - \bar{x}, x - \bar{x} \rangle$

Măsura calității predicțiilor

MAE → mean absolute error

$\frac{1}{m} \sum |y_i - \hat{y}_i|$ ($\hat{y}_i$ = predicția)

x mic ⇒ $x^2 < |x|$

$x = 0.1$

$x^2 = 0.01 < |x| = 0.1$

RMSE → residual mean squared error

$\sqrt{\frac{1}{m} \sum (y_i - \hat{y}_i)^2}$ ⇝ ordinul 1

același unități ca și y_i

Mai multe caracteristici

2 caracteristici:

- mărimea sarcinii pt server
- lungimea cozii sarcinilor.

Estimarea timpului de execuție -

$$\hat{y} = b + w_s s + w_q q$$

1 output — 2 input-uri

Dimensiune N1

$$\hat{y} = 1 \cdot b + w_1 x_1 + w_2 x_2 + \dots + w_p \cdot x_p$$

$$= \langle (1, x_1, x_2 \dots, x_p), (b, w_1, \dots, w_p) \rangle$$

input — parametri

Pt mai multi input-uri

$$\begin{pmatrix} 1, x_{11}, x_{12} & x_{1p} \\ 1, x_{21}, x_{22} \dots & x_{2p} \\ \vdots & \\ X & \end{pmatrix} \begin{pmatrix} b \\ w_1 \\ \vdots \\ w_p \\ \beta \end{pmatrix} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{pmatrix}$$

Predicțiile $\approx y_i$ (datele)

$$\underbrace{\begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix}} \approx \underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{y \to \text{vect datelor}}$$

$$X \cdot \beta \approx y$$

$$\underbrace{X \cdot \beta - y}_{\dim n} \approx 0 \qquad \left\| \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \right\| = \sqrt{x_1^2 + \dots + x_n^2}$$

$J(\beta) = \| X \cdot \beta - y \|$ mică, minimală.

$\beta = (b, w_1, \dots, w_p)$ param. modelului

Pt a găsi β rezolvăm

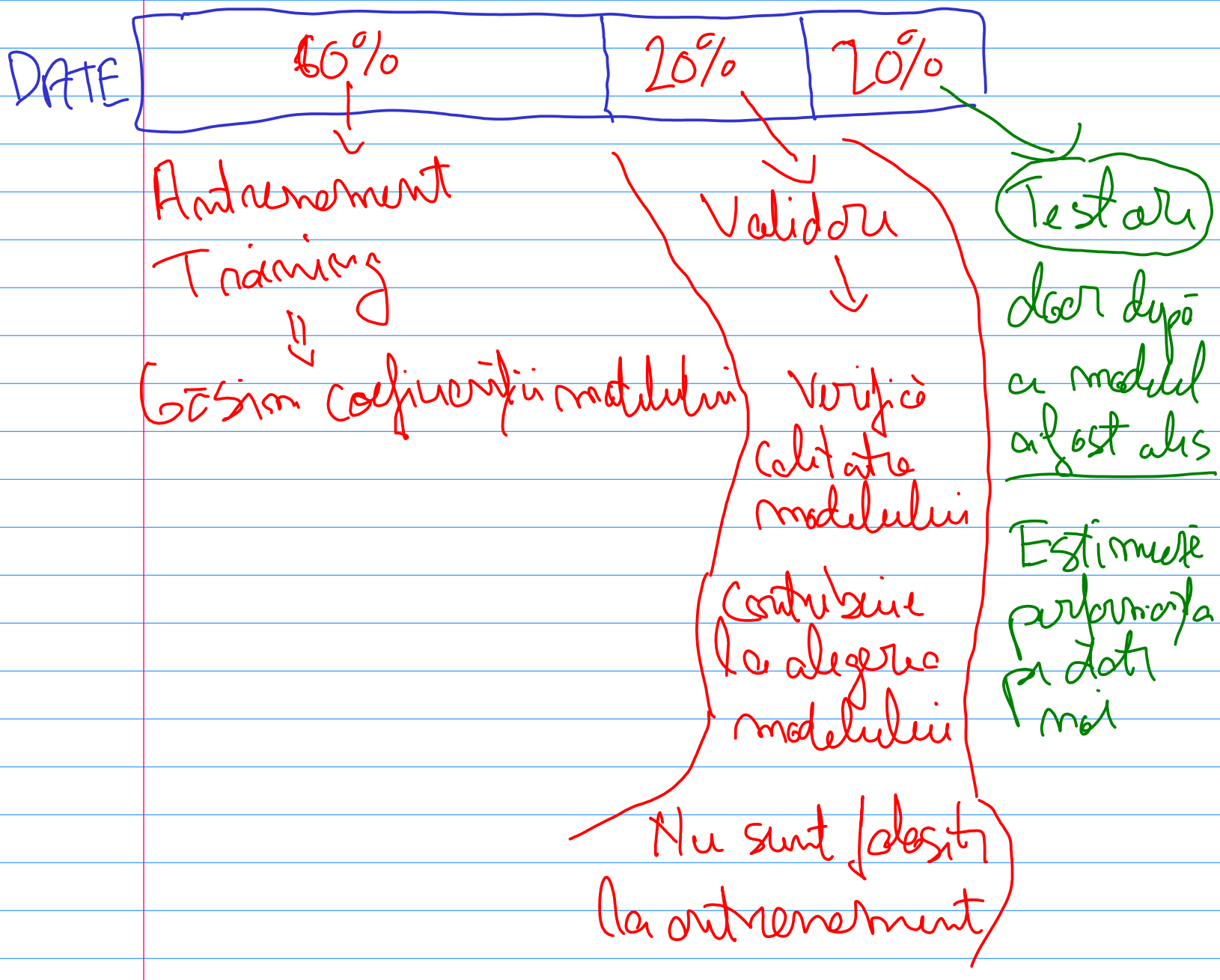
$$X^T X \beta = X^T y \to \text{se află } \beta$$

E important să re-scalăm cantitățile.

timpul în s	d, distanța în km
t	d
1	0,01
10	0,05

	y
	0,5
	0,1

Procesarea datelor


DATE
60%
20%
20%
Antrenament
Training
Testare
Validare
Învățare
Învățare
Învăț

→ Underfitting :- am putea "și mai bine"
- modelul nu este optimal
- antrenamentul ar putea continua

→ Overfitting :- modelul e prea bun pe datele de training
- nu generalizează corect pe datele de test.

U(0,5;10) → [un număr random în [0,5;b]
Toate numerele au aceeași probabilitate.

Poisson

Notebook A

payload queue length

1. latency $\approx b + w_s S + w_q \cdot q$

a) latency $\approx b + w_s \cdot S$ $w_q = 0$

Pt regresie: minimizăm $J(\beta)$

→ formulă

→ Linear Regression

→ Aplicăm un alg. optimizare

$$\beta_{m+1} = \beta_m - \eta \nabla J(\beta_m)$$

repeat until $\beta_{m+1} \approx \beta_m$