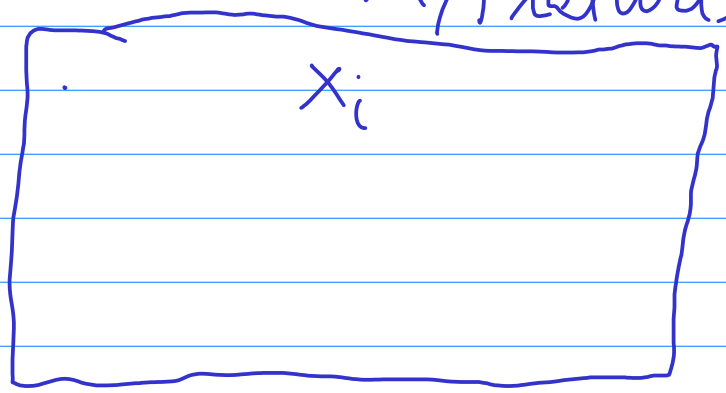
Curs 3 - Regresie logistică
Clasificare

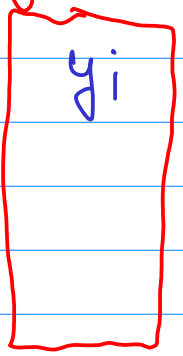
Scopul unui alg de Machine Learning (învățare pornind de la date existente)

Supervised Learning

Caracteristici/Features



Target/output



Obiectivul: Găsim f_θ care verifică (θ → parametrii modelului)

$$f_\theta(x_i) \approx y_i$$

Exemplu $x_i = (s_i, t_i)$ $\quad y_i \in \mathbb{R}$

$$f_\theta(x_i) = b + w_s s_i + w_t t_i$$

$\theta = (b, w_s, w_t)$ parametrii modelului

$$f_\theta(x_i) = b + (w_s, w_t) \cdot (s_i, t_i)$$

produs scalar

Regresie liniară

→ Pentru a găsi approx prin regresie liniară minimizăm

$$RSS(\theta) = \sum (\underbrace{b + w_s \cdot s_i + w_t t_i}_{\text{model}} - y_i)^2$$

$\theta \leftrightarrow (b, w_s, w_t)$ — get_data

→ Există formule matematice pt găsirea parametrilor optimali

→ Există pachete soft ce fac asta

sklearn

```
model = LinearRegression().fit(X_train, y_train)
prediction = model.predict(X_new)
```

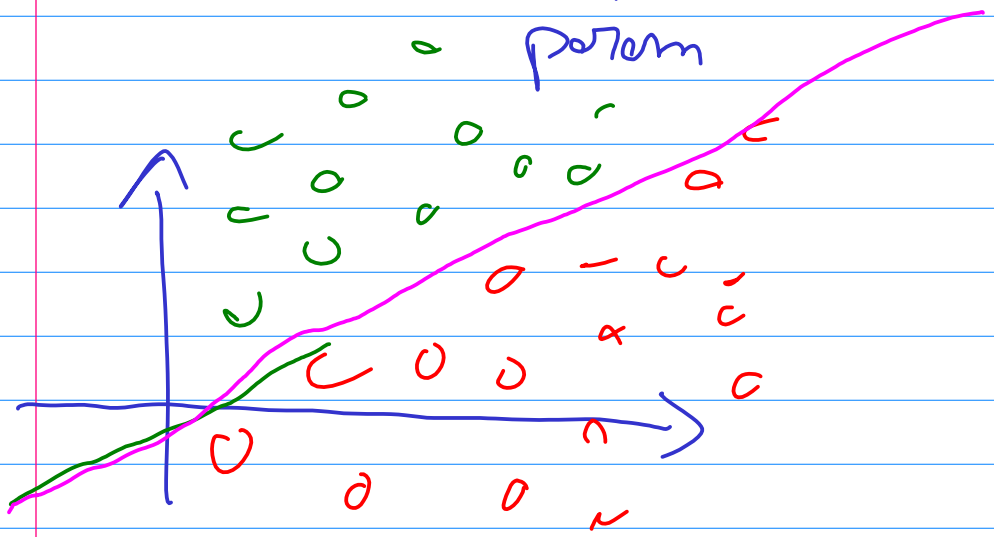
fit → găsește coeficienții

predict → fac predicții.

Regresie liniară : $x_i \longrightarrow$ valoare $\in \mathbb{R}$

Regresia logistică :

$z = b + x \cdot w \longrightarrow$ o clasă

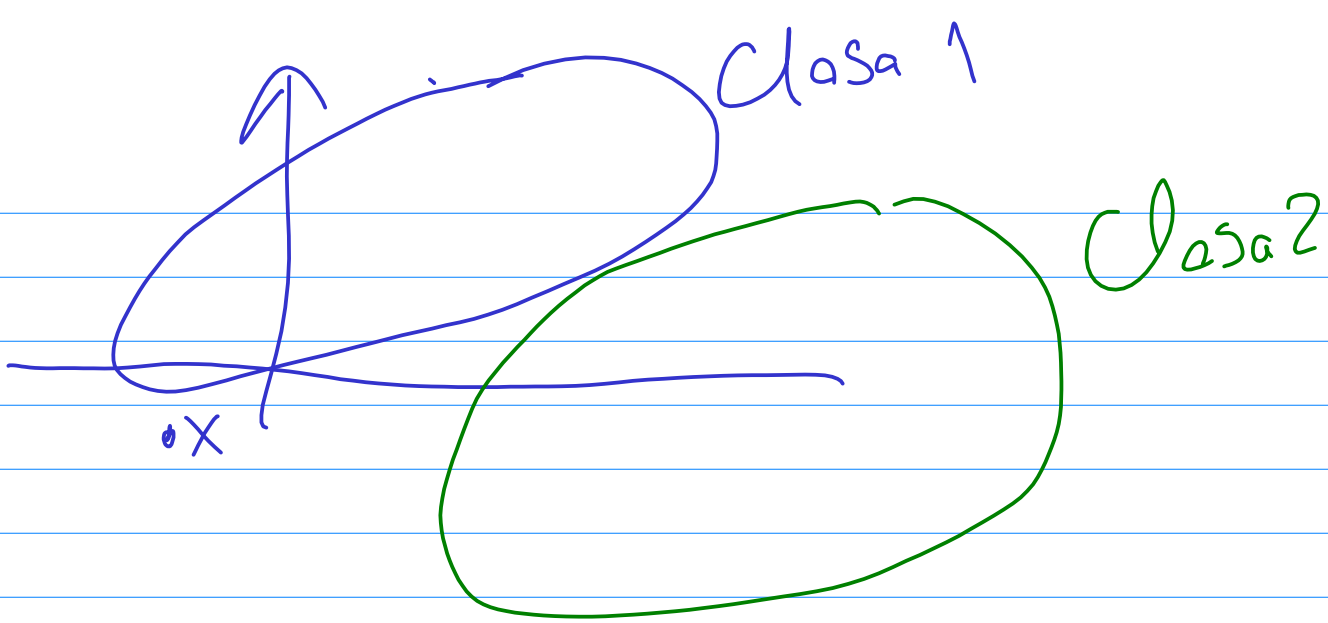


$x \longmapsto z = b + x \cdot w \longrightarrow \sigma(z) = \pi(x)$

$\sigma(z) = \frac{1}{1+e^{-z}}$ (sigmoid)



$\mathbb{R} \xrightarrow{\sigma} [0,1]$ (probabilități)



Eveniment: $x \in$ Clasa 1

$x \rightarrow$ calculăm probabilitatea dată de $\pi(b + w \cdot x)$

$$\begin{cases} p \geq \frac{1}{2} & \text{Clasa 1} \\ p \leq \frac{1}{2} & \text{Clasa 2} \end{cases}$$

Clasificator: pornim de la un input și decidem dacă aparține unei clase sau nu.

Cum formulăm problema ca o optimizare

$$J(\beta) = -\frac{1}{n} \sum y_i \cdot \log(\pi_i) + (1-y_i)\log(1-\pi_i)$$

β - parametri

$$\pi_i = \sigma(b + x_i \cdot w)$$

minimizând $J(\beta) \Rightarrow \pi_i \approx y_i \in \{0,1\}$

clasele 0, 1 pt x_i

$y_i = 0 \Rightarrow -\frac{1}{n}(1-y_i)\log(1-\pi_i)$

f_i

Dacă $\pi_i \approx 0 \Rightarrow f_i \approx 0$

Dacă $\pi_i \approx 1 \Rightarrow f_i \approx -\frac{1}{n}\log(0) = +\infty$

Deci o rețea probabilistică este penalizată puternic de funcția J de mai sus

Am găsit parametrii unui model de predicție și vrem să îl testăm.

- când face modelul preziceri corecte
- când se înșală
- când face modelul preziceri utile

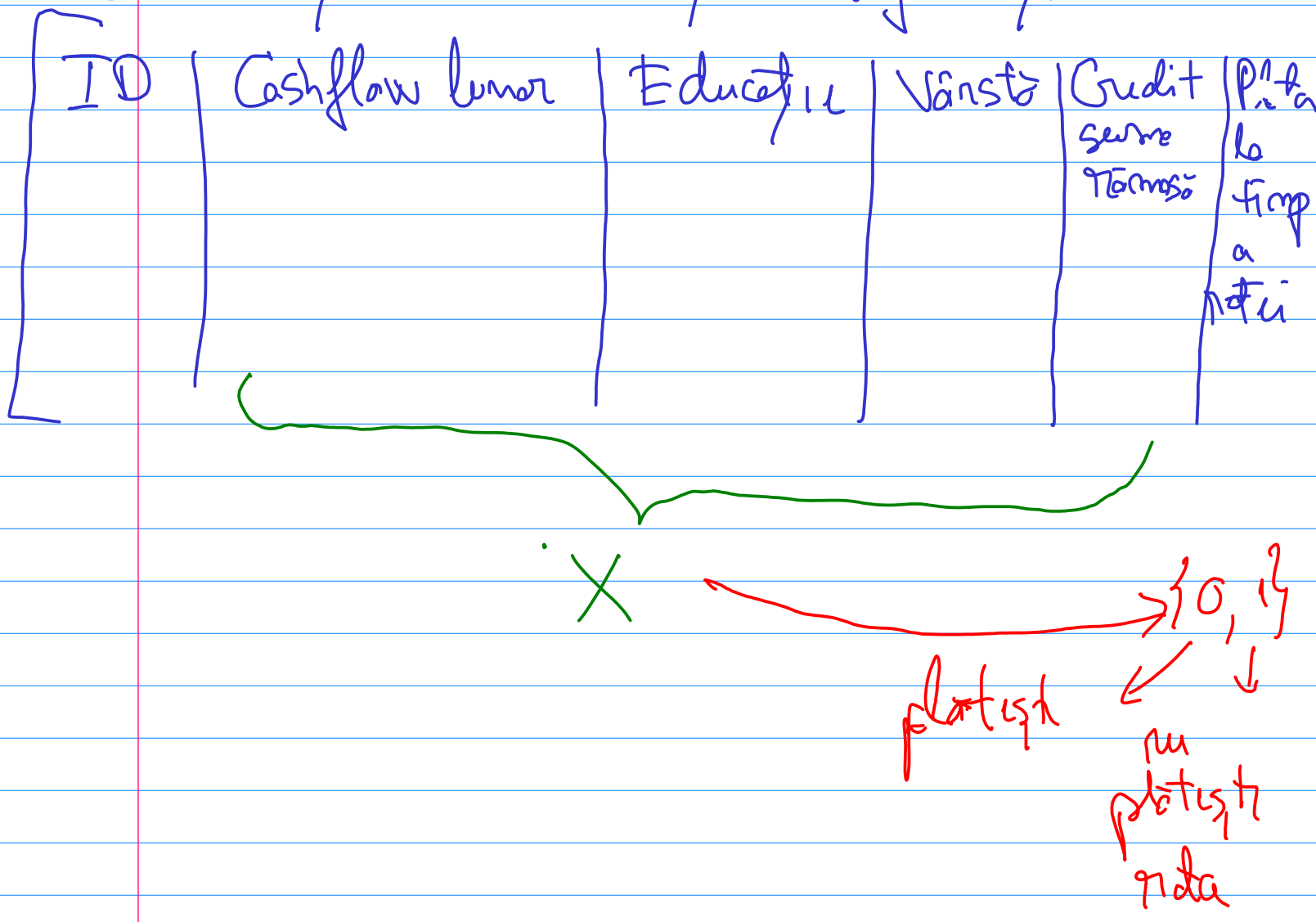
Exemplu: O bancă are o bază de date cu clienții care conține informații

ID	Cashflow lunar	Educație	Vârstă	Credit sumă rămasă	Prima la timp a ratei

$X \longrightarrow \{0, 1\}$

plătește

nu plătește rata



In baza de date :

9000 exemple : plata OK
1000 exemple : plata neefectuata

$Model_1(X) = 0$ (plata se face!)

Accuracy : $Accuracy = \frac{\text{nr previsori corecte}}{\text{total}}$

$$= \frac{9000}{10000} = 90\%$$

True Neg 9000 | 8500	False Pos 0 | 500
False Neg 1000 | 500	True Pos 0 | 500

Model 2

Precision : $\frac{TP}{TP+FP} = \frac{0}{0}$ nedefinit

Recall : $\frac{TP}{TP+FN} = \frac{0}{1000} = 0$

Model 2 : Accuracy : $\frac{TN+TP}{\text{total}} = \frac{9000}{10000} = 90\%$

Precision: $\frac{TP}{TP+FP} = \frac{500}{1000} = 50\%$

Câte "alarme false" avem.

detectează doar jumătate din problema

Recall: $\frac{TP}{TP+FN} = \frac{500}{1000} = 50\%$

Dintre clienții problematici câți am detectat.

Cost False Positive (alarmă falsă)

Cost False Negative (problemă nedetectată)

$$\begin{cases} 1 & \text{prob} \geq t \\ 0 & \text{prob} \leq t \end{cases}$$

poate fi alis

C_1 = Costul alarmei false: $C_{FP} \cdot (1-t)$

C_2 = Costul unei prob nesemnalate $C_{FN} \cdot t$

$t \nearrow \quad C_1 \searrow \quad C_2 \nearrow$

$C_1 = C_2 \qquad C_{FP} = t \cdot (C_{FP} + C_{FN})$

$$\downarrow$$

$$t = \frac{C_{FP}}{C_{FP} + C_{FN}}$$

Costul unui client care nu plătește = 1000

Costul unui alarm false = 100

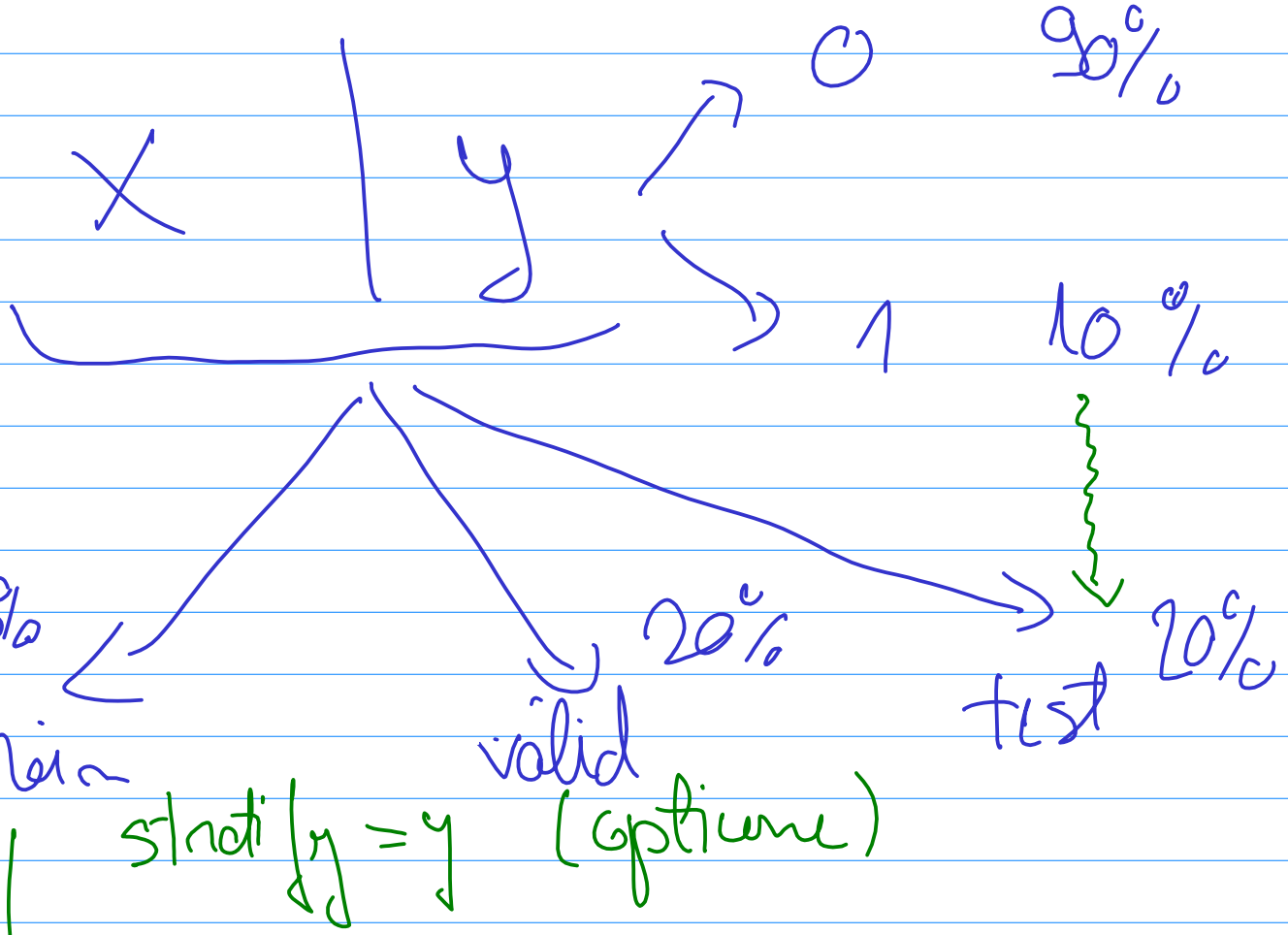
$$\frac{100}{1100} = 0{,}09$$

Partiția datelor:

Test 60%	Validare 20%	Test 20%
→ găsim parametrii modelului → fit	→ testăm precizia pt diferite modele → într-un fel contribuie la antrenament	→ doar la sfârșit

Splitarea datelor:

train_test_split
permite respectarea proporției la încărcare.



0 → 90%	0 → 90%	0 → 90%
1 → 10%	1 → 10%	1 → 10%

payload → w_p

quun → w_q

ccchu → w_c

intercept b

$$z = b + w_p \cdot p + w_q \cdot q + w_c \cdot c$$

doğum ağırlığı.