Analiza și procesarea datelor prin tehnici de Învățare Automată

4. Tehnici de învățare supervizată 2: Regresie

Șef Lucrări Dr. Ing. Horia Modran
Contact: horia.modran@unitbv.ro / modranhoria@gmail.com
Tel: 0770171577

Universitatea Transilvania din Brașov
FACULTATEA DE INGINERIE ELECTRICĂ ȘI ȘTIINȚA CALCULATOARELOR

2024 - 2025

Clasificare și Regresie

- Învățarea unei funcții discrete: Clasificare
  - Clasificare binară:
    - fiecare exemplu este clasificat ca adevărat (pozitiv) sau fals (negativ)
  - poate, de asemenea, clasifica în mai multe clase (3, 4, 5...)
- Învățarea unei funcții continue: Regresie
  - Regresie liniară
  - Regresie logistică

Regresie

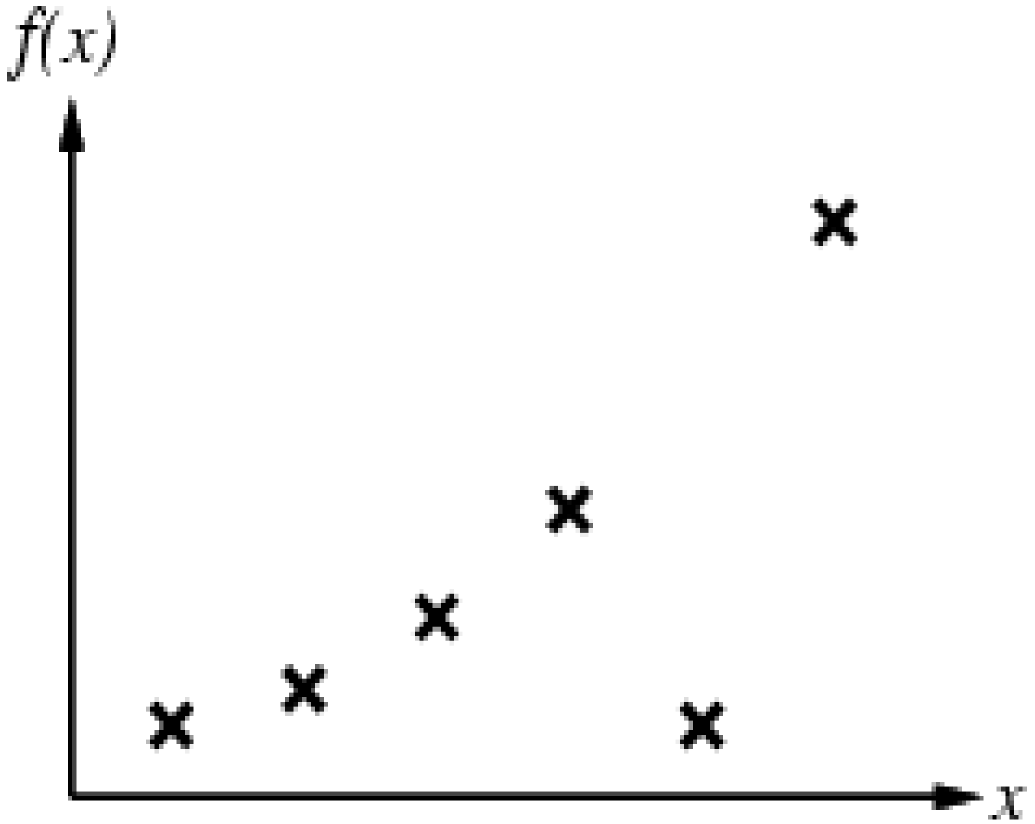
- Tehnică de învățare supervizată
- Urmărește să determine corelația între variabile
- În contextul învățării automate: prezicerea unei variabile țintă continue în funcție de una sau mai multe variabile
- Utilizată în
  - predicție/prognoză
  - modelarea seriilor de timp
  - determinarea relației cauză-efect

Corelație

- Asociere liniară între două variabile
- Arată cum să determinăm atât natura, cât și puterea relației dintre două variabile
- Corelația este între -1 și +1
- Corelația zero indică faptul că nu există nicio relație între variabile
- Coeficientul de corelație Pearson
  - cea mai cunoscută măsură a dependenței dintre două mărimi

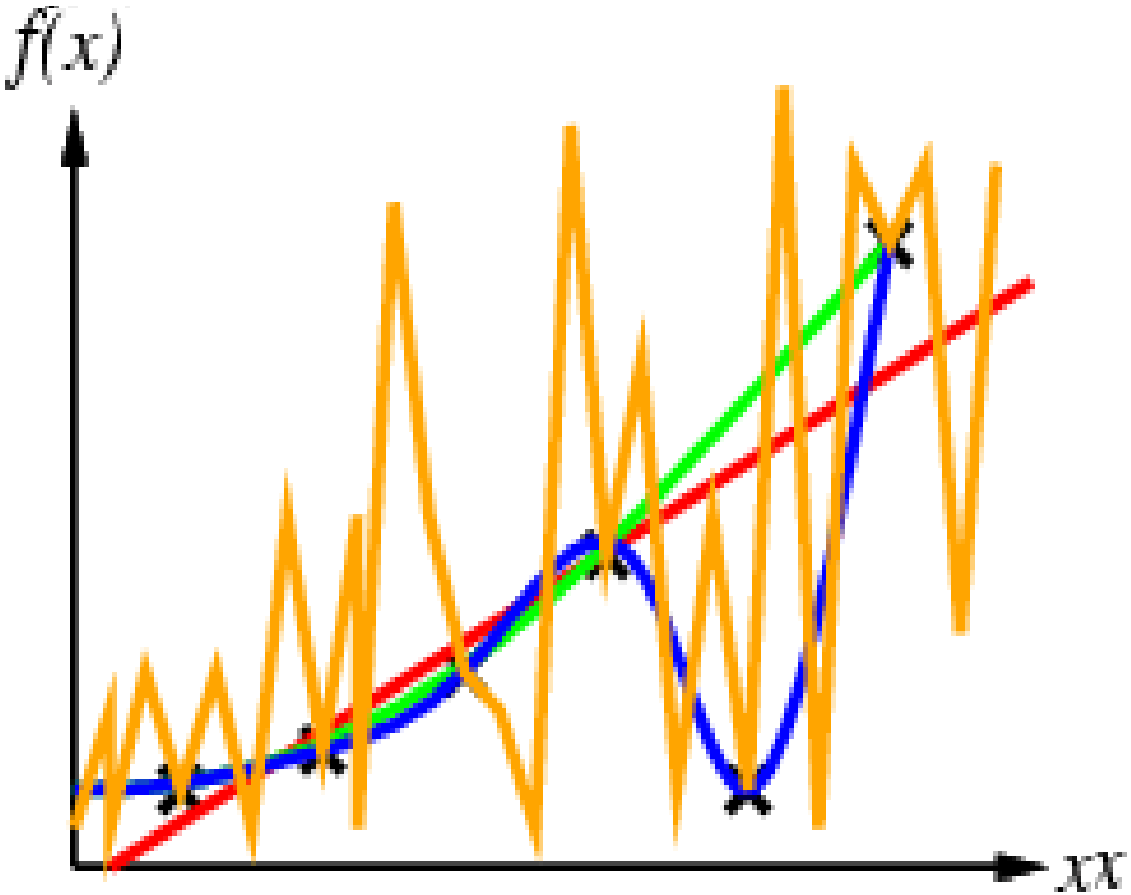
Aproximarea unei funcții

- Să se determine funcția care trece prin punctele de mai jos



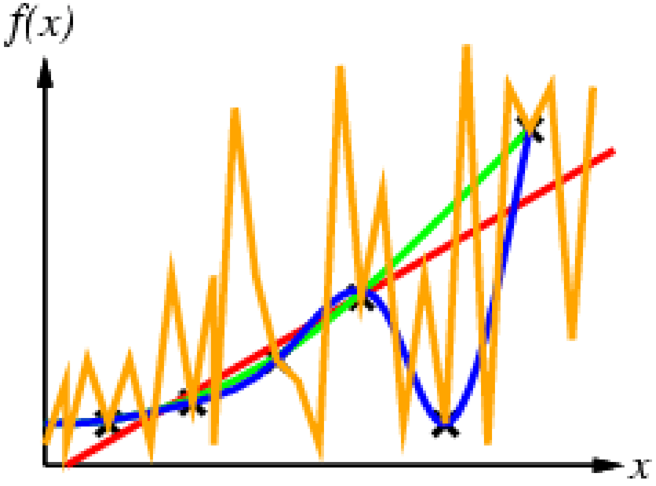
Aproximarea unei funcții

- Modalități de aboradare



Briciul lui Ocam

- Trebuie preferată cea mai simplă ipoteză consistentă cu datele
- În caz de egalitate, modelele mai simple tind să generalizeze mai bine decât cele complexe



William of Occam (1285-1347) – filozof englez

Entia non sunt multiplicanda praeter necessitatem

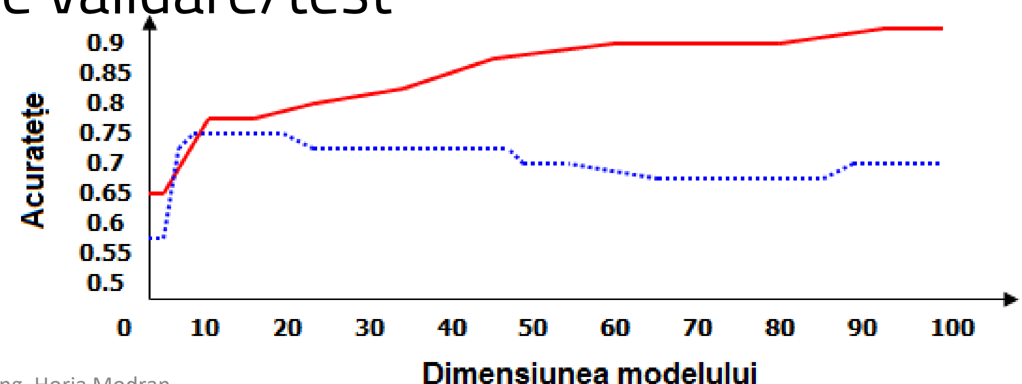
~ Cea mai simplă explicație este cea mai bună

Generalizarea

- Scopul este găsirea unui model pentru determinarea valorii clasei în funcție de valorile celorlalte atribute cu eroare cât mai mică
- Modelul trebuie să aibă capacitate de generalizare, care arată cât de bun este modelul pentru date **noi**
- De obicei există o mulțime de antrenare pentru crearea modelului și o mulțime de test pentru verificarea capacității de generalizare

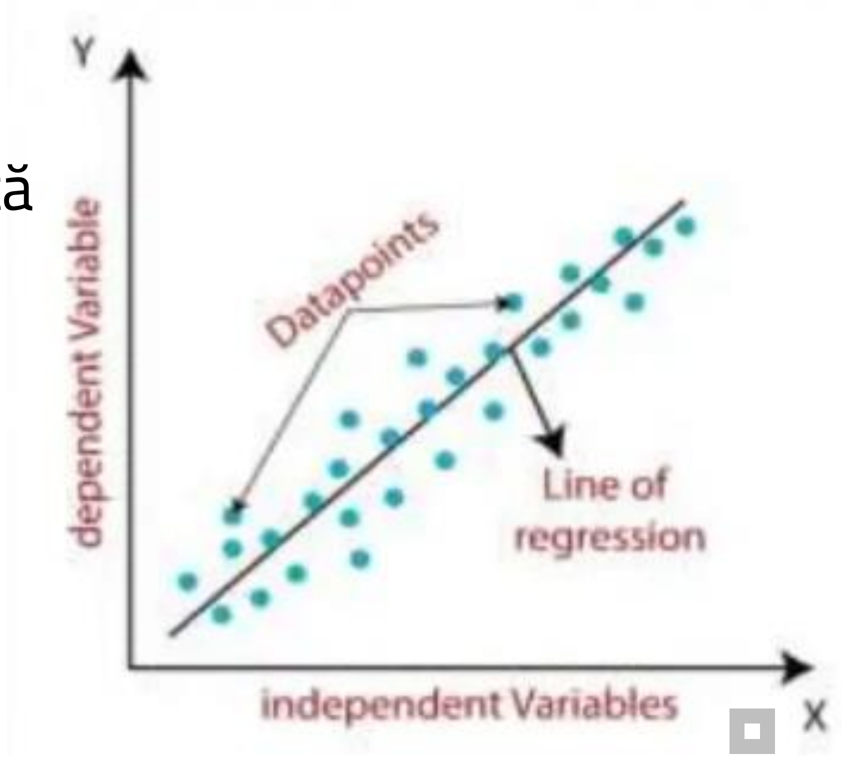
Generalitatea unui model

- Subpotrivirea (*underfitting*): modelul este prea simplu și nu poate învăța distribuția datelor
- Suprapotrivirea (*overfitting*): modelul este prea complex și poate fi influențat de zgomot și date irelevante
- Un model suprapotrivit are performanțe foarte bune pe mulțimea de antrenare, dar performanțe slabe pe mulțimea de validare/test



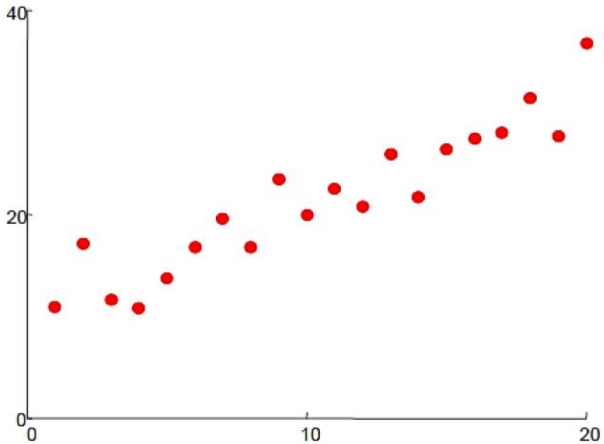
Regresie Liniară

- **Regresia Liniară** este utilizată pentru a prezice valoarea unei variabile pe baza valorilor uneia/mai multor variabile
- Este o metodă de învățare supervizată
- Scopul: determinarea relației liniare dintre o variabilă dependentă și una/mai multe variabile independente

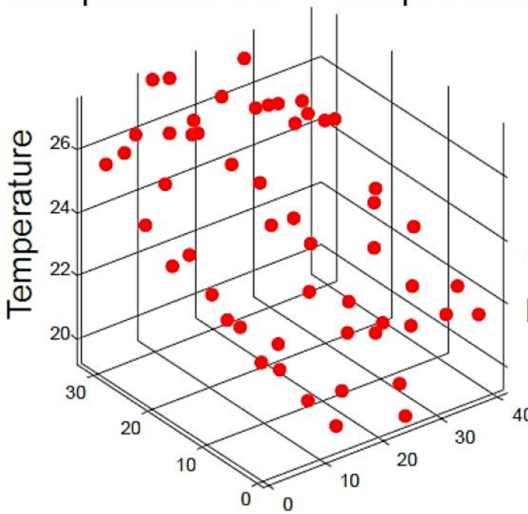


Regresie Liniară

Samples with ONE independent variable

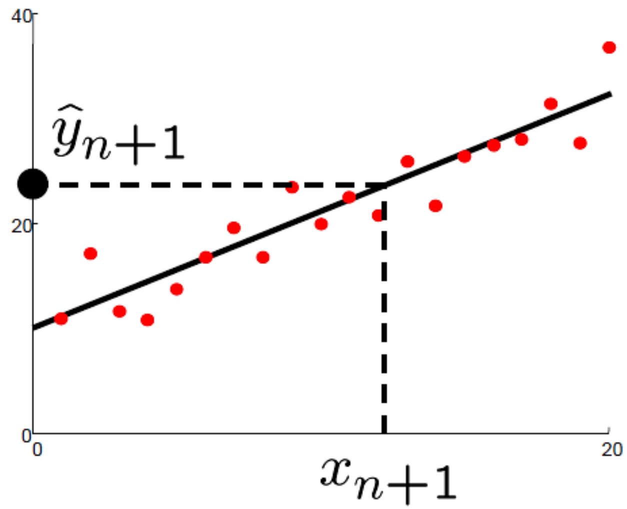


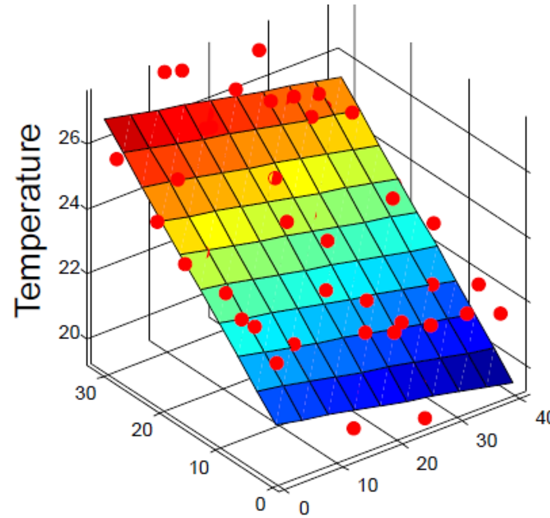
Samples with TWO independent variables



Given examples $(x_i, y_i)_{i=1 \ldots n}$

Predict y_{n+1} given a new point x_{n+1}





Regresia Liniară

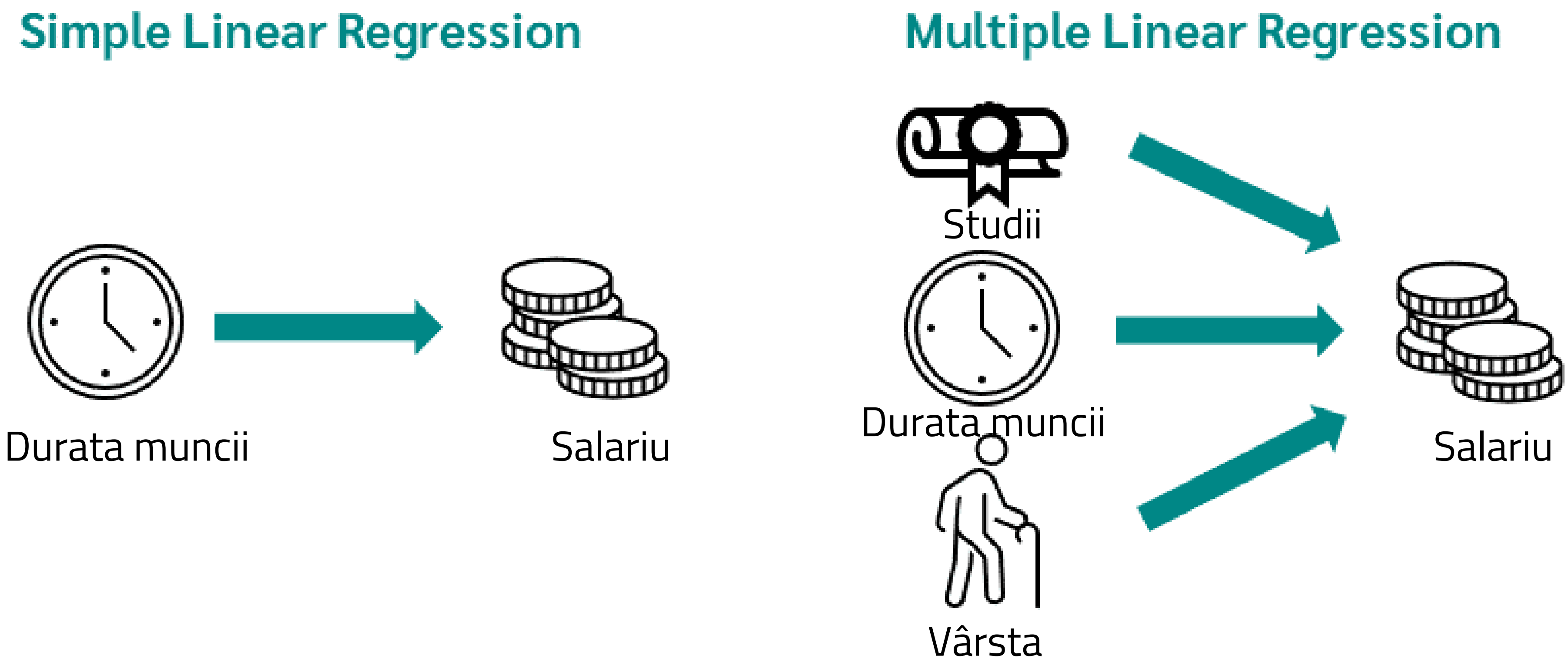
- Considerăm un model

$$\boldsymbol{y = a + bX + \varepsilon}$$

, unde $\mathbf{a}$ și $\mathbf{b}$ sunt interceptarea și panta (cunoscuți drept coeficienți sau parametri), iar $\boldsymbol{\varepsilon}$ este termenul de eroare

- Regresie liniară simplă
  - o singură variabilă independentă este utilizată
- Regresie liniară multiplă
  - două sau mai multe variabile independente sunt folosite pentru predicție

Tipuri de regresie liniară



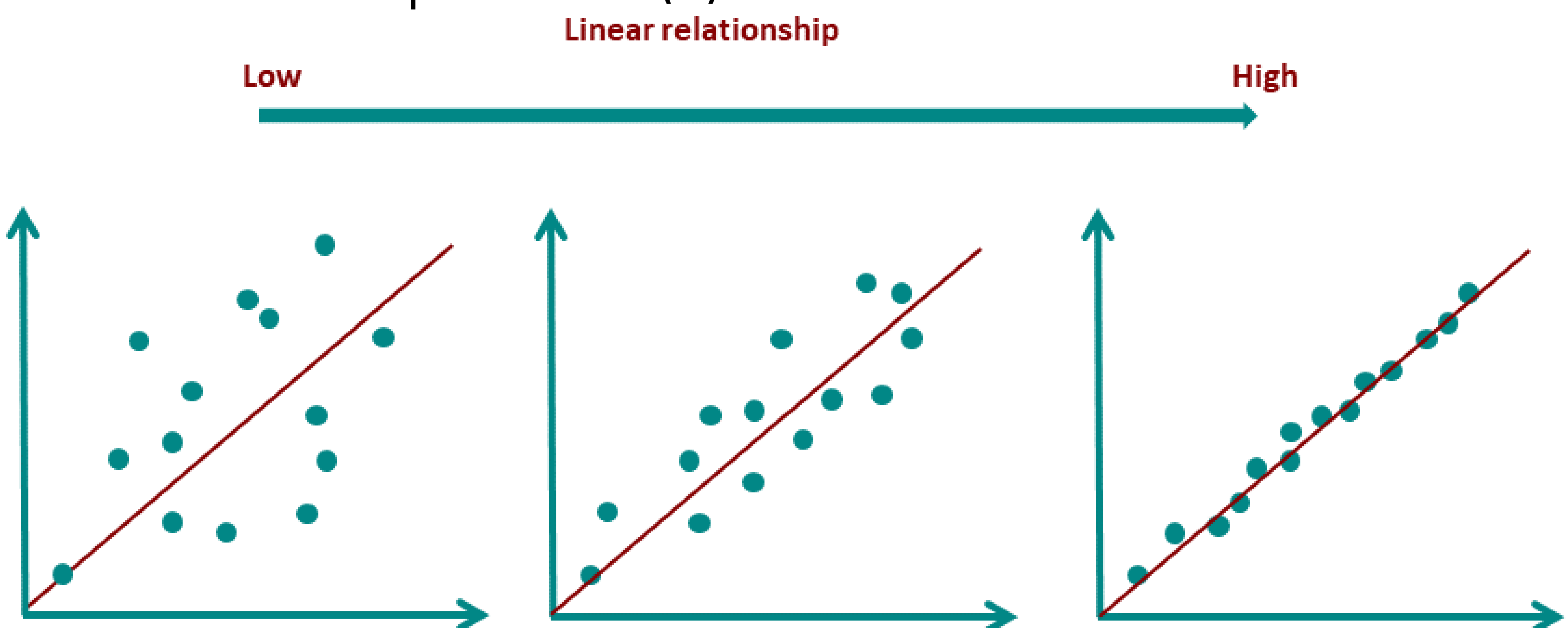
Regresia Liniară simplă

- Exemplu: **Are** **înălțimea** **influență asupra** **greutății** **unei persoane**?

Variabilă independentă

Variabilă dependentă

- Scop: prezicerea valorii variabilei dependente (y) pe baza variabilei independente (**X**)



Regresia Liniară simplă

$$y_i = a + bX_i + \varepsilon_i$$

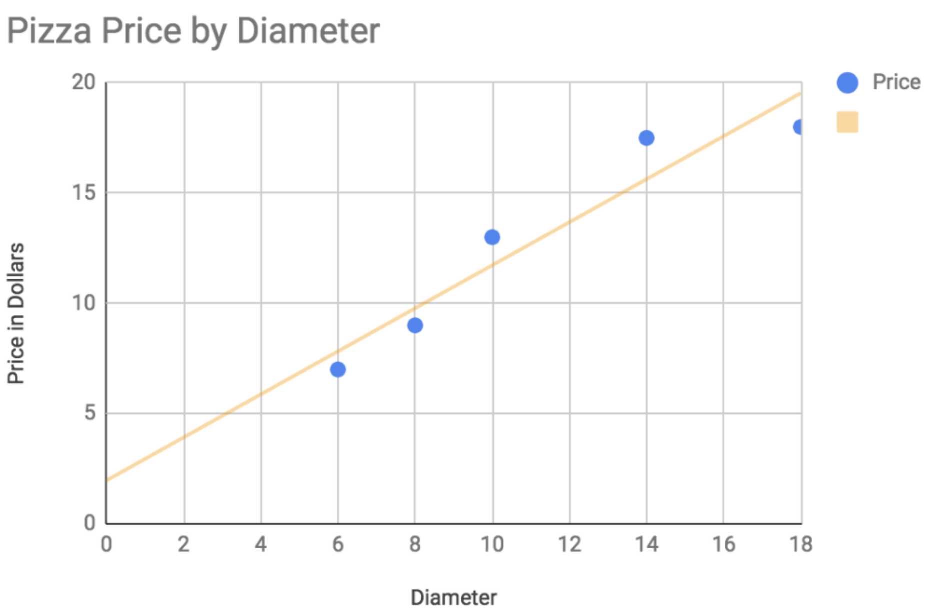
, unde y = variabila dependentă, x – variabila independentă, $\boldsymbol{a}$ – intercept, $\boldsymbol{b}$ – panta liniei, $\boldsymbol{\varepsilon}$ – termenul de eroare

- **Model Simplu**: doar un X
- **Liniar în parametri**: niciun parametru nu apare ca exponent sau este înmulțit/împărțit cu un alt parametru
- **Liniar în variabila predictor** (X): X apare doar la prima putere

Exemplu

- Predicția prețului pentru pizza
  - Care este prețul unei pizza de 20 inch?

Diameter	Price
6	7
8	9
10	13
14	17.5
18	18



Calculare

x	y	x^2	xy
3	8	9	24
9	6	81	54
5	4	25	20
3	2	9	6
∑x = 20	**∑y = 20**	**$\sum x^2 = 124$**	**∑xy = 104**

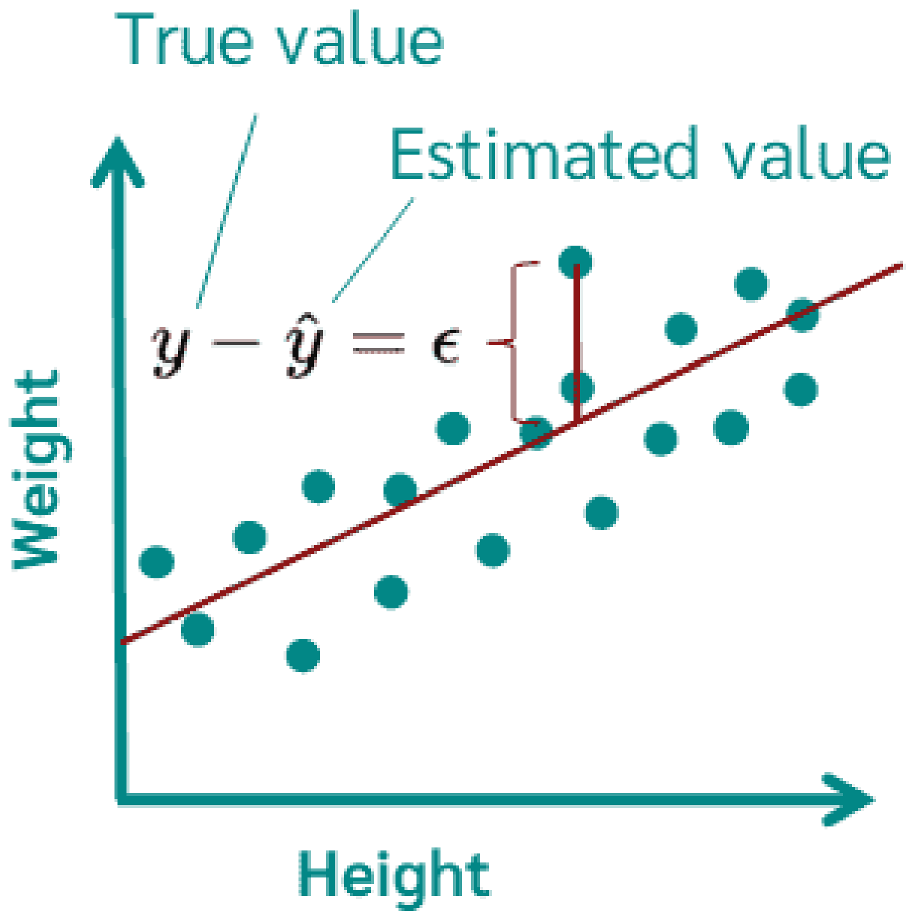
x	y
3	8
9	6
5	4
3	2

$$\boldsymbol{a = \frac{\sum y \sum x^2 - \sum x \sum xy}{n \sum x^2 - (\sum x)^2} = \frac{20 * 124 - 20 * 104}{4 * 124 - 20^2} = \frac{400}{96} = 4.17}$$

$$\boldsymbol{b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} = \frac{4 * 104 - 20 * 20}{4 * 124 - 20^2} = \frac{16}{96} = 0.166}$$

$$\boldsymbol{y = 4.17 + 0.166x}$$

Eroare



Error epsilon

$$y = b \cdot x + a + \boxed{\epsilon}$$

Eroare

- Graficul *Scatterplot* arată că punctele nu sunt pe o linie și astfel, pe lângă relație, descriem și eroarea $\boldsymbol{\varepsilon}$:

$$\boldsymbol{y_i = a + \ b_1 \, X_i + \ \varepsilon_i}$$

  - distanța dintre valoarea estimată și valoarea adevărată trebuie să fie cât mai mică (implicit și eroarea $\boldsymbol{\varepsilon}$)
- Coeficientul de regresie b poate avea acum semne diferite:
  - b > 0: există o corelație pozitivă între x și y (cu cât x mai mare, cu atât y mai mare)
  - b < 0: există o corelație negativă între x și y
  - b = 0: nu există nicio corelație între x și y

Regresia Liniară multiplă

$$\mathbf{y} = \boldsymbol{a} + \boldsymbol{b_1} \boldsymbol{X_1} + \boldsymbol{b_2} \boldsymbol{X_2} + \cdots + \boldsymbol{b_n} \boldsymbol{X_n} + \boldsymbol{\varepsilon}$$

, unde y = variabila dependentă, x – variabila independentă, $\boldsymbol{a}$ – intercept, $\boldsymbol{b_1}, \boldsymbol{b_2}, \ldots, \boldsymbol{b_n}$ – pantele, $\boldsymbol{\varepsilon}$ – termenul de eroare

- Coeficienții pot fi interpretați ca la regresia simplă
- Mai multe variabile independente
- O singură variabilă dependentă
- Utilizată adeasea în cercetări de piață

Regresia Liniară multiplă

- Predicție Y după x_1 și x_2

$$X = \begin{pmatrix} 1 & 1 & 4 \\ 1 & 2 & 5 \\ 1 & 3 & 8 \\ 1 & 4 & 2 \end{pmatrix} \quad y = \begin{pmatrix} 1 \\ 6 \\ 8 \\ 12 \end{pmatrix}$$

x1 Product 1 Sales	x2 Product 2 Sales	Y Weekly Sales
1	4	1
2	5	6
3	8	8
4	2	12

$$y = a + b_1 x_1 + b_2 x_2$$

$$\mathrm{b} = ((X^T X)^{-1} X^T)\, Y \text{ , unde } a = \begin{pmatrix} a \\ b_1 \\ b_2 \end{pmatrix}$$

Exemplu calcul

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 4 & 5 & 8 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 4 \\ 1 & 2 & 5 \\ 1 & 3 & 8 \\ 1 & 4 & 2 \end{pmatrix} = \begin{pmatrix} 4 & 10 & 19 \\ 10 & 30 & 46 \\ 19 & 46 & 109 \end{pmatrix}$$

$$(X^T X)^{-1} = \begin{pmatrix} 4 & 10 & 19 \\ 10 & 30 & 46 \\ 19 & 46 & 109 \end{pmatrix}^{-1} = \begin{pmatrix} 3.15 & -0.59 & -0.30 \\ -0.59 & 0.2 & 0.016 \\ -0.3 & 0.016 & 0.054 \end{pmatrix}$$

$$(X^T X)^{-1} X^T = \begin{pmatrix} 3.15 & -0.59 & -0.30 \\ -0.59 & 0.2 & 0.016 \\ -0.3 & 0.016 & 0.054 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 4 & 5 & 8 & 2 \end{pmatrix} = \begin{pmatrix} 0.05 & 0.47 & -1.02 & 0.19 \\ -0.32 & -0.098 & 0.155 & 0.26 \\ -0.065 & 0.05 & 0.185 & -0.125 \end{pmatrix}$$

$$= ((X^T \; X)^{-1} X^T) \, Y = \begin{pmatrix} 0.05 & 0.47 & -1.02 & 0.19 \\ -0.32 & -0.098 & 0.155 & 0.26 \\ -0.065 & 0.05 & 0.185 & -0.125 \end{pmatrix} \begin{pmatrix} 1 \\ 6 \\ 8 \\ 12 \end{pmatrix} = \begin{pmatrix} \mathbf{-1.69} \\ \mathbf{3.48} \\ \mathbf{-0.05} \end{pmatrix}$$

$$\boldsymbol{y = -1.69 + 3.48x_1 - 0.05x_2}$$

Regresie cu valiabile categoriale

- Presupunem că toate variabilele sunt variabile continue
- Variabile categoriale:
  - Variabile ordinale - codifică datele cu valori continue
    - Evaluation: Excellent (5), Very good (4), Good (3), Poor (2), Very poor (1)
  - Variabile nominale – de utilizează *dummy variables*
    - Department: Computer, Biology, Physics

	Computer	Biology	Physics
Computer	1	0	0
Biology	0	1	0
Physics	0	0	1

Funcție cost

- Funcția de cost este necesară pentru a calcula diferența dintre valorile reale și cele prezise
- Pentru modelul de regresie liniară, funcția de cost va fi minimul Erorii Pătratice Medie Rădăcină (engl. *Root Mean Square Error* – **RMSE**) a modelului, obținută prin scăderea valorilor prezise din valorile reale
- Este o valoare numerică care indică succesul sau eșecul unui anumit model, fără a fi nevoie să înțelegem funcționarea interioară a unui model

Tipuri de funcție cost

- **Eroare Medie** (*Mean Error*) – aceste erori pot fi negative sau pozitive
  - prin urmare, se pot anula reciproc în timpul însumării, iar **eroarea medie a modelului poate fi zero**
- **Eroare medie pătratică** (*Mean Squared Error* - MSE) - MSE reprezintă diferența medie pătrată dintre predicții și rezultatele așteptate

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2$$

n – nr. de puncte
y_i - valoarea reală
$\hat{y}_i$ - valoarea prezisă

Tipuri de funcție cost

- **Eroare absolută medie** (*Mean Absolute Error* – MAE) – MAE calculează valoarea absolută a diferenței (nu există posibilitate de erori negative)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|$$

n – nr. de puncte
y_i - valoarea reală
$\hat{y}_i$ - valoarea prezisă

- Root Mean Squared Error - unul dintre cei doi indicatori principali de performanță pentru un model de regresie

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n}}$$

Coeficient de determinare

- Coeficientul de determinare (R^2) este proporția variației variabilei dependente care este datorată variabilelor independente

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

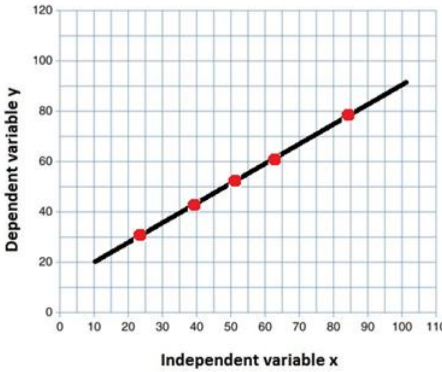
$\boldsymbol{y_i}$ – valoarea reală
$\boldsymbol{\hat{y}_i}$ - valoarea estimate de model
$\boldsymbol{\bar{y}_i}$ - valoarea medie

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}$$

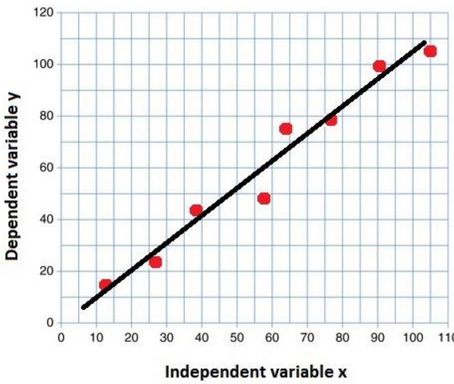
Coeficient de determinare

- Exemplu de valori pentru R^2

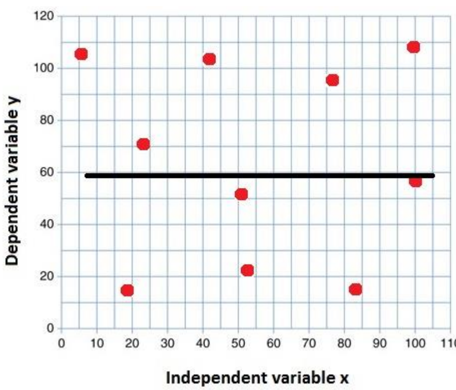
$R^2 = 1$ Toată variația valorilor y este explicată de valorile x



$R^2 = 0.83$ 83 % din variația valorilor y este explicată de valorile x



$R^2 = 0$ Variația valorilor y nu este influențată deloc de valorile x



Regresie Liniară pentru clasificare

- Pentru clasificare binară
  - se codifică etichetele clasei ca y=0,1 sau {–1,1}
  - se aplică: $\mathbf{y} = \mathbf{X} * \mathbf{b} + \mathbf{e}$
  - se determină de care clasă este mai apropiată predicția
    - dacă clasa 1 este codificată la 1 și clasa 2 este – 1

$$\begin{array}{ll} class\ 1 & if\ f(x) \geq 0 \\ class\ 2 & if\ f(x) < 0 \end{array}$$

    - **Modelel liniare NU sunt optimizate pentru clasificare**

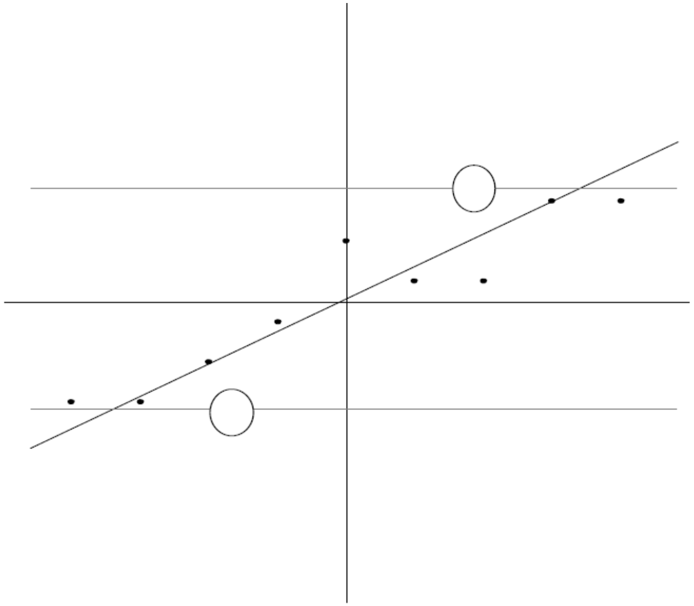
⇒ Regresie logistică

Regresie Logistică

- Prezice rezultatele pe o variabilă de rezultat binară
  - de exemplu, dacă un pacient are sau nu o boală
  - dacă un nou solicitant va reuși sau nu
  - rezultatul nu este continuu sau distribuit normal
- când avem un răspuns de tip variabilă binară
  - codificăm „boală” ca 1 și „fără boală” ca 0, putem doar să potrivim o linie prin acele puncte așa cum am face cu regresia liniară? Posibil! Dar există anumite probleme.

Regresie Liniară: Probleme

- Problema potrivirii unei linii regulate de regresie la o variabilă dependentă binară
  - linia pare să simplifice prea mult relația
  - oferă predicții care nu pot fi valori observabile ale lui Y pentru valorile extreme ale lui X
  - abordarea este analogă cu potrivirea unui model liniar la probabilitatea evenimentului



- Produce predicții neobservabile pentru valori extreme ale variabilei dependente

Regresie Logistică

- Se începe cu ecuația Regresiei Liniare
- Se vor determina coeficienții de regresie

Variabilă dependentă

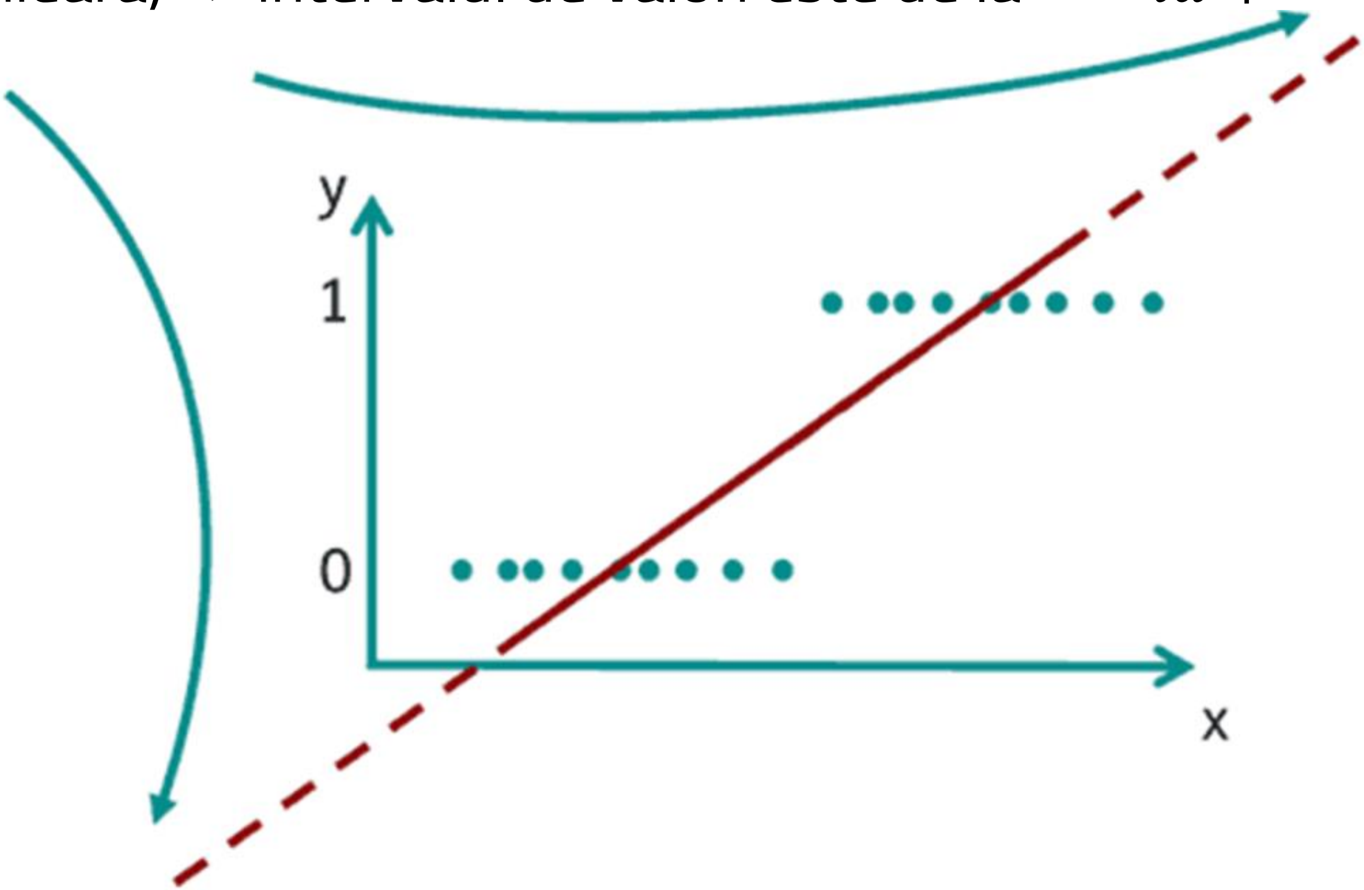
Variabile independente

$$\boxed{\hat{y}} = \boxed{b_1} \cdot \boxed{x_1} + \boxed{b_2} \cdot \boxed{x_2} + \ldots + \boxed{b_k} \cdot \boxed{x_k} + \boxed{a}$$

Coeficienți de regresie

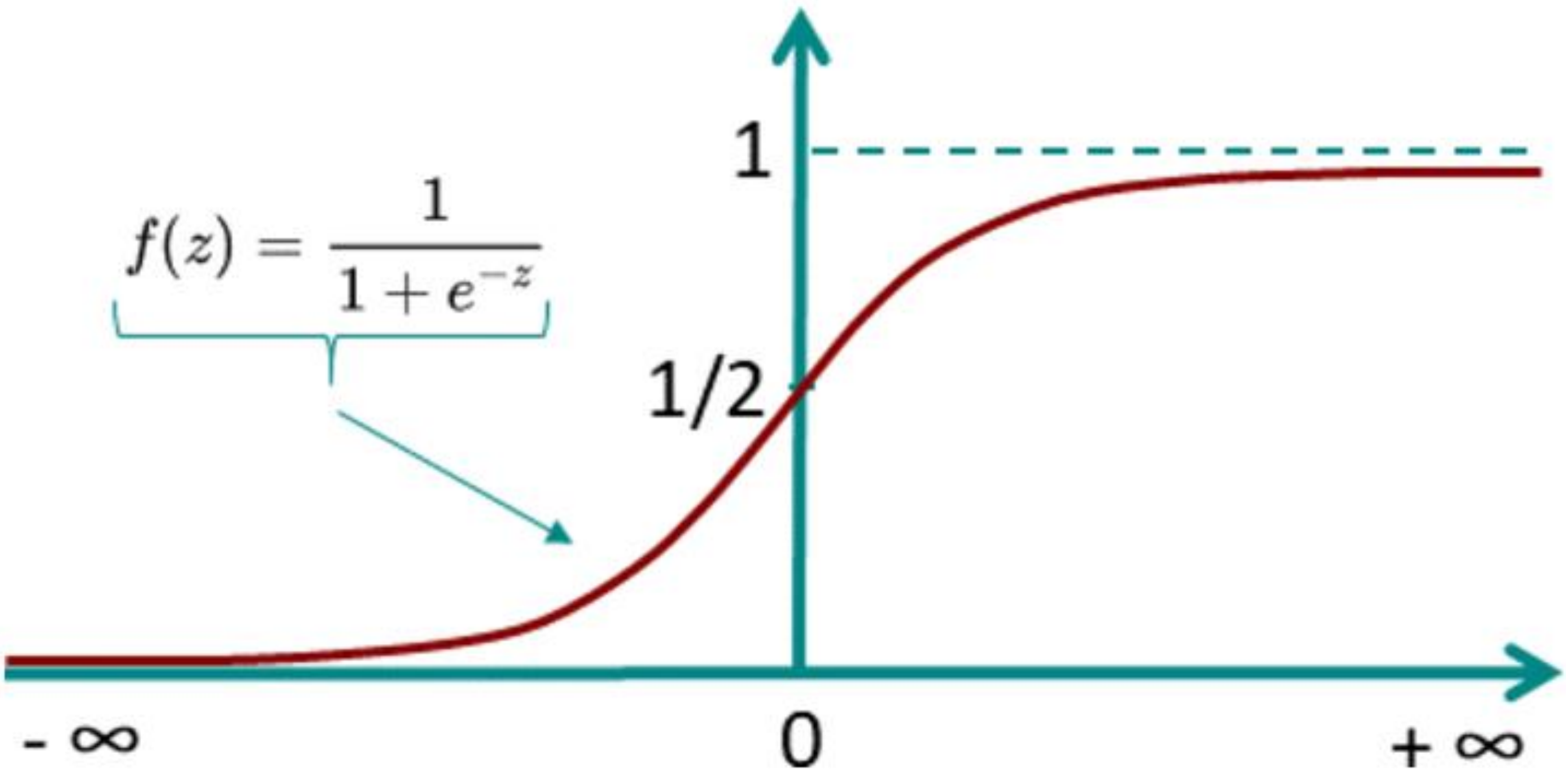
Regresie Logistică

- Rezultatul calcului ar fi un model liniar (similar cu Regresia Liniează) -> intervalul de valori este de la $-\infty$ $la + \infty$



Funcția logistică

- Modelul logistic se bazează pe **funcția logistică**
- Are valori doar între 0 și 1



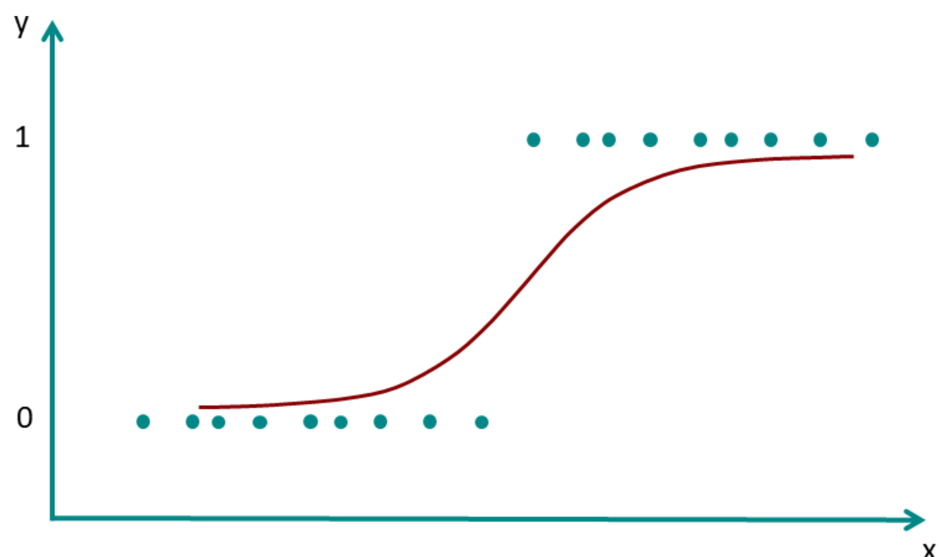
Funcția logistică

- Ecuația funției logistice

$$\hat{y} = \boxed{b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a}$$

$$f(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-\left(\boxed{b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a}\right)}}$$

- Graficul funcției va arăta astfel:



Abordare probabilistică

- Funcția logistică este perfectă pentru a descrie probabilitatea P(y=1|x)
- Probabilitatea ca pentru valori date ale variabilei independente, variabila dependentă binară y să fie 0 sau 1 este dată de ecuația:

$$P(y=1|x_1, x_2, \dots, x_n) = \frac{1}{1+\ e^{-(b_1x_1+b_2x_2+\cdots+b_nx_n+a)}}$$

$$P(y=0|x_1, x_2, \dots, x_n) = 1\ - P(y=1)$$
$$= 1 - \frac{1}{1+\ e^{-(b_1x_1+b_2x_2+\cdots+b_nx_n+a)}}$$

Exemplu

- Exemplu - tabelul arată numărul de ore de studiu pentru un student și dacă a promovat examenul (1) sau nu (0)

Study Hours (x_k)	0.5	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass (y_k)	0	0	0	0	0	0	0	1	0	1	0	1	0	1	1	1	1	1	1

Ecuație de regresie: $\boldsymbol{\hat{y}_i = a + b_1 X_i}$

Determinăm valorile pentru:
$\boldsymbol{a}$ = -4.1
$\boldsymbol{b_1}$ = 1.5

De exemplu, pentru x = 2: $\boldsymbol{\hat{y}_i = a + b_1 * 2 = -4.1 + 1.5 * 2 = -1.1}$

Aplicăm funcția logistică: $P(y = 1|x) = \frac{1}{1+e^{-\hat{y}}} = \frac{1}{1+e^{1.1}} \approx \mathbf{0.25} \rightarrow \textcolor{red}{\boldsymbol{clasa\ 0}}$

Similar pentru x = 4: $\boldsymbol{\hat{y}_i = a + b_1 * 4 = -4.1 + 1.5 * 4 = 1.9}$

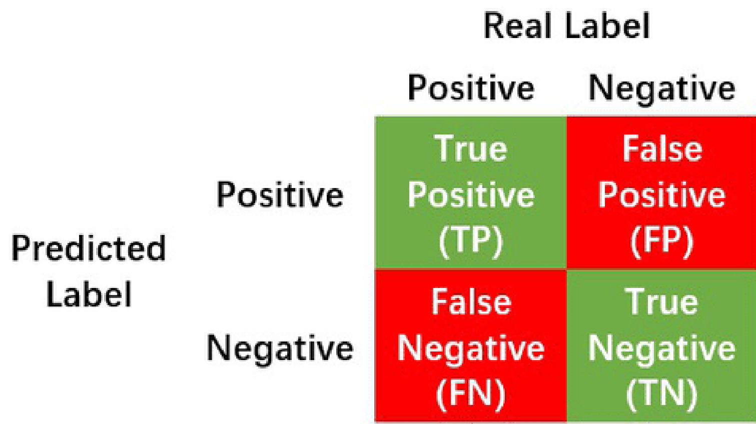
$$P(y = 1|x) = \frac{1}{1+e^{-\hat{y}}} = \frac{1}{1+e^{-1.9}} \approx \mathbf{0.87} \rightarrow \textcolor{red}{\boldsymbol{clasa\ 1}}$$

Regresie Logistică Multinomială

- De regulă regresia logistică este aplicată în probleme de clasificare binară
- Regresia logistică multinomială este o metodă de clasificare care generalizează regresia logistică la problema de clasificare multiclasă
- Modul de funcționare este același ca și în regresia logistică, singura diferență fiind că variabilele dependente sunt mai degrabă categorice decât binare, adică există n rezultate posibile și nu doar două

Indicatori de performanță

- Fiind algoritm de clasificare, se aplică aceeași indicatori de performanță specifici algoritmilor de clasificare:
  - Acuratețe
  - Precizie
  - Recall
  - Scor F1



$$\text{Precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FN}}$$

$$\text{Accuracy} = \frac{\sum \text{TP} + \text{TN}}{\sum \text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

ÎNTREBĂRI ?