Regresie liniară și polinomială

Noțiunea de *regresie* se referă la o categorie de metode statistice care au ca scop estimarea relațiilor dintre una sau mai multe variabile independente și una sau mai multe variabile dependente. Scopul regresiei este de a determina o funcție prin care se aproximează dependența dintre cele două categorii de variabile. Estimarea acestei dependențe se realizează pornind de la o mulțime finită puncte din spațiul variabilelor (mulțimea datelor de antrenare), care reprezintă cazuri particulare ale interdependenței căutate. Tehnicile prin care se realizează estimarea în cadrul regresiei depind în principal de tipul datelor de intrare/ieșire din setul de date de antrenare și tipul funcției care se estimează.

Regresia liniară

Scopul regresiei liniare este de a determina parametrii unei funcții liniare care descrie relația dintre variabilele dependente și cele independente. În cazul cel mai simplu, se dorește determinarea unei funcții de gradul I care să permită estimarea unei valori de ieșire pentru orice valoare de intrare din domeniul vizat. Estimarea se face pe baza unui set de date cu intrări/ieșiri cunoscute (datele de antrenare).

Exemplu:

Fie următorul set de date, care conține intervalul de timp dedicat studiului pentru susținerea un examen și notele obținute de câțiva studenți la acel examen (Fig. 1):

Nr ore de studiu	**0.5**	**0.75**	**1**	**1.25**	**1.5**	**1.75**	**1.75**	**2**	**2.25**	**2.5**	**2.75**	**3**	**3.25**	**3.5**	**4**	**4.25**	**4.5**	**4.75**	**5**	**5.5**
Nota obținută	**4**	**3**	**2.5**	**1**	**2**	**3.5**	**6**	**4**	**7**	**1.5**	**5**	**2.5**	**5.5**	**3**	**8**	**7**	**7.5**	**6**	**8.5**	**9.5**

Acestea constituie datele de antrenare, scopul fiind de a determina o funcție care să permită estimarea notei obținute pentru orice număr de ore de studiu. În cazul regresiei liniare, această funcție este o dreaptă (o funcție de gradul I).

Folosim următoarele notații:

N – numărul de instanțe din setul de date de antrenare (în exemplu, N = 20)

$x_i, i = 1..N$ – valorile variabilei independente (numărul de ore de studiu)

$y_i, i = 1..N$ – valorile variabilei dependente (nota)

$\overline{x}$ – valoarea medie a variabilei independente

$\overline{y}$ – valoarea medie a variabilei dependente

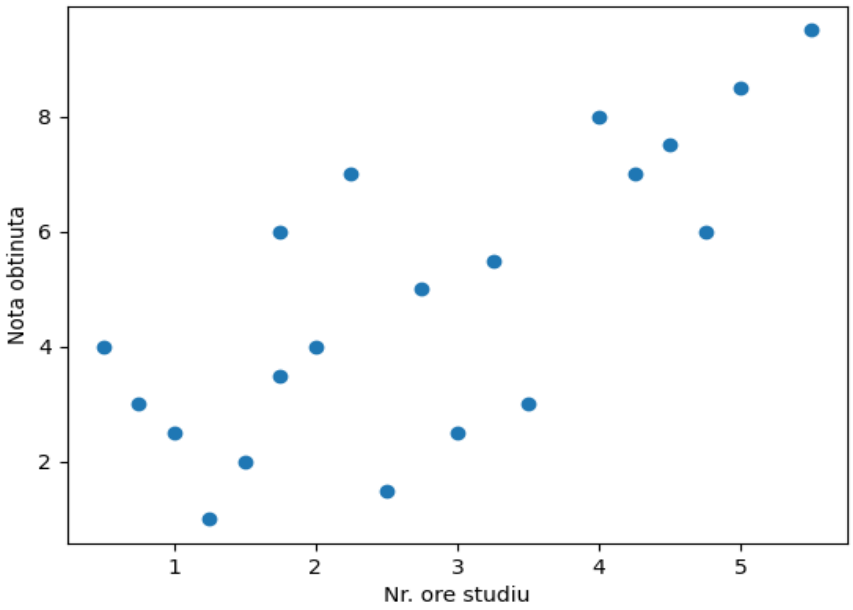

Fig. 1. Datele de antrenare

Trebuie să determinăm dreapta care exprimă cel mai bine dependența notei obținute de numărul de ore de studiu. Fie b_0 și b_1 coeficienții acestei drepte. Funcția pe care trebuie să o estimăm este cea din Ec. (1). Estimarea se referă la determinarea celor mai utile valori ale coeficienților b_0 și b_1. Pentru datele prezentate anterior, funcția care corespunde acestor coeficienți este reprezentată în Fig. 1.

$$y = b_0 + b_1 x \tag{1}$$

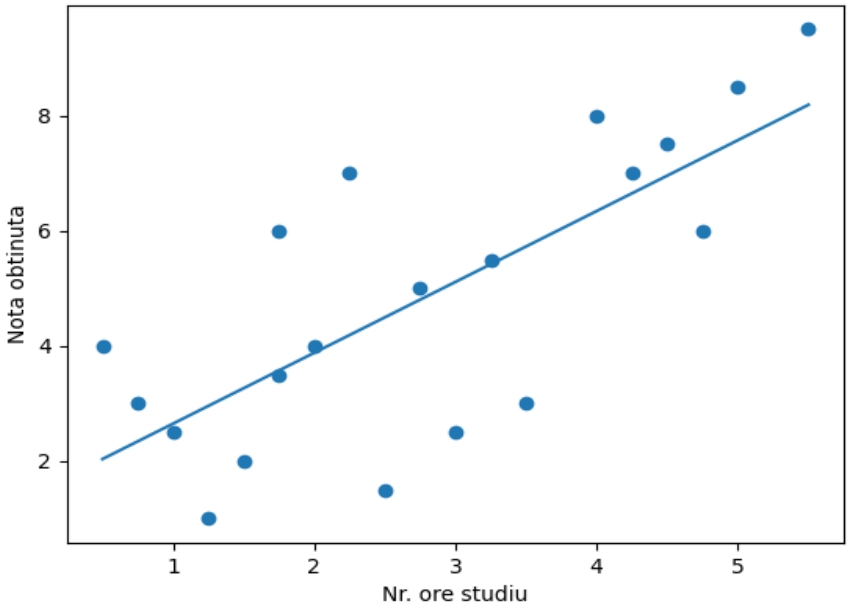

Fig. 2. Funcția liniară care descrie cel mai bine dependența dintre variabila independentă (nr. ore studiu) și cea dependentă (nota obținută)

Soluția analitică

Fie $\hat{y}_i$ valorile returnate de funcția căutată (estimări ale notei obținute) pornind de la valorile cunoscute ale variabilei independente (nr. orelor de studiu) (Ec. (2)).

$$\hat{y}_i = b_0 + b_1 x_i \quad i = 1..N \tag{2}$$

Pentru fiecare x_i obținem o eroare $e_i = y_i - \hat{y}_i$. Dreapta dorită trebuie să minimizeze aceste erori. Altfel spus, această dreaptă are proprietatea că diferențele dintre valorile din setul de date inițial și cele obținute cu funcția care descrie dreapta sunt minime. Astfel, funcția care trebuie minimizată este cea din Ec. (3), care se dezvoltă în Ec. (4):

$$S = \sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2 \tag{3}$$

$$S = \sum_{i=1}^{N} \left(y_i - \left(b_0 + b_1 x_i \right) \right)^2 \tag{4}$$

Așadar, scopul este de a determina coeficieții b_0 și b_1 care minimizează funcția din Ecuația 4. Pentru aceasta, se determină derivatele funcției *S* în raport cu b_0 și b_1 și se caută valorile coeficienților pentru care derivatele sunt nule. Efectuând calculele aferente, rezultă formulele de calcul ale celor doi coeficienți (Ec. (5)).

$$\begin{aligned} b_0 &= \bar{y} - b_1 \bar{x} \\ b_1 &= \frac{\sum_{i=1}^{N} \left(x_i - \bar{x} \right)\left(y_i - \bar{y} \right)}{\sum_{i=1}^{N} \left(x_i - \bar{x} \right)^2} \end{aligned} \tag{5}$$

Odată determinată funcția dorită, se poate estima “corectitudinea” acesteia determinând eroarea medie pătratică (Ec. (6)).

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2 \tag{6}$$

Soluția numerică

Soluția analitică prezentată anterior determină cu precizie valorile coeficienților. În practică, există însă numeroase situații în care coeficienții nu se pot determina prin calcule matematice directe. În aceste cazuri, se pot folosi diverse metode de optimizare numerică. Acestea presupun determinarea valorilor coeficienților pe parcursul mai multor iterații în cadrul cărora coeficienții se ajustează puțin câte puțin, până când valorile lor se apropie suficient de mult de cele ideale. Una dintre cele mai frecvent-întâlnite astfel de metode este ***gradientul descendent (vezi Lab6)***. Pașii sunt următorii:

- Se stabilește un număr de iterații și o rată de învățare *α* (valoare subunitară mică)
- Se inițializează coeficienții (de exemplu cu valori aleatoare)
- Se stabilește modul de calcul al erorii. Se poate folosi *Sum of Squared Errors* (Ec. (7))

$$SSE = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \left(b_0 + b_1 x_i\right)\right)^2 \tag{7}$$

- Pentru fiecare iterație:
  - Se determină eroarea (valoarea SSE).
  - Se determină valorile componentelor gradientului (derivatele parțiale ale erorii în raport cu coeficienții). Calculul gradientului se poate face prin:
    - determinarea analitică a derivatelor parțiale (Ec. (8))

$$\begin{aligned}\frac{\partial SSE}{\partial b_0} &= \sum_{i=1}^{N} -\left(y_i - \left(b_0 + b_1 x_i\right)\right)\\ \frac{\partial SSE}{\partial b_1} &= \sum_{i=1}^{N} -\left(y_i - \left(b_0 + b_1 x_i\right)\right) x_i\end{aligned} \tag{8}$$

    - aproximarea derivatelor folosind diferențe finite (Ec. (9))
      (Δ este o valoare subunitară mică)

$$\begin{aligned}\frac{\partial SSE}{\partial b_0} &= \frac{SSE(b_0 + \Delta) - SSE(b_0)}{\Delta}\\ \frac{\partial SSE}{\partial b_1} &= \frac{SSE(b_1 + \Delta) - SSE(b_1)}{\Delta}\end{aligned} \tag{9}$$

- Se actualizează coeficienții astfel (Ec. (10)):

$$b_0 = b_0 - \alpha \frac{\partial SSE}{\partial b_0}$$
$$b_1 = b_1 - \alpha \frac{\partial SSE}{\partial b_1} \quad (10)$$

- Se determină noua eroare SSE. Dacă aceasta este suficient de apropiată ca valoare de eroarea de la începutul iterației, algoritmul se oprește.

Regresia polinomială

Cazul anterior este o situație particulară, când se caută o funcție de gradul I. În caz general, funcția dorită poate fi un polinom de orice grad (Ec. (11)):

$$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \ldots \quad (11)$$

Ca și în cazul regresiei liniare, coeficienții b_i se pot determina analitic sau prin metoda gradientului descendent.

Soluția analitică

Ec. (11) se poate scrie sub formă matriceală (Ec. (12)).

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \ldots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^m \\ 1 & x_2 & x_2^2 & \ldots & x_2^m \\ 1 & x_3 & x_3^2 & \ldots & x_3^m \\ \ldots & & & & \\ 1 & x_n & x_n^2 & \ldots & x_n^m \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_3 \\ \ldots \\ b_m \end{bmatrix} \quad (12)$$

$$\vec{y} = X\vec{b}$$

Unde *n* este numărul instanțelor (nr valorilor din setul de date de antrenare) și *m* este gradul polinomului. Prin prelucrarea Ec. (12) rezultă expresia vectorului de coeficienți (Ec. (13)).

$$\vec{b} = (X^T X)^{-1} X^T \vec{y} \quad (13)$$

Cerințe:

1) Implementați regresia liniară pentru datele din exemplu

- determinați cei doi coeficienți și afișați dreapta obținută (pentru datele din exemplu, din calculul analitic ar trebui să rezulte $b_0 = 1.419$, $b_1 = 1.23$). Determinați eroarea medie pătratică (pentru datele din exemplu, MSE = 2.709).
- determinați cei doi coeficienți prin metoda gradientului descendent. Calculați eroarea medie pătratică și comparați rezultatele cu cele obținute anterior (ar trebui să rezulte valori foarte apropiate de cele determinate analitic).

2) Implementați regresia polinomială folosind polinoame de gradul 2 ... 8.

- determinați eroarea pentru fiecare funcție polinomială.
- determinați polinomul care se potrivește cel mai bine pe datele din laborator (cel pentru care MSE este minim). În Fig. 3 se prezintă câteva exemple de regresie cu polinoame de diverse grade.

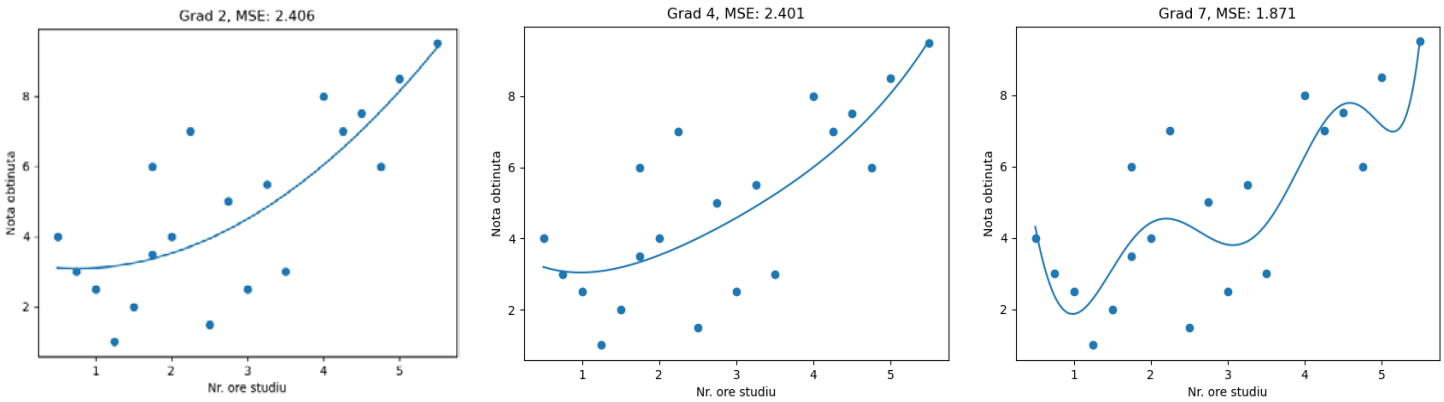

Fig. 3. Regresie folosind polinoame de diverse grade și valorile MSE care rezultă în fiecare caz

Pentru afișarea grafică a datelor și funcțiilor se poate proceda ca în exemplul următor:

```
import numpy as np
import matplotlib.pyplot as plt

x = np.array([1, 3, 4, 6, 3, 6, 7, 2])
y = np.array([5, 2, 5, 7, 1, 4, 3, 5])

def myFunc(x):
    return 0.7 * x ** 2 - 3 * x + 1

plt.scatter(x, y)
xplot = np.arange(min(x), max(x), 0.01)
plt.plot(xplot, myFunc(xplot))
plt.xlabel('x values')
plt.ylabel('y values')
plt.show()
```