Laboratorium 2 - Regresja liniowa metodą najmniejszych kwadratów

Mateusz Podmokły - II rok Informatyka WI

7 marzec 2024

1 Treść zadania

Zadanie 1. Celem zadania jest zastosowanie metody najmniejszych kwadratów do predykcji, czy nowotwór jest złosliwy, czy łagodny. Nowotwory złosliwe i łagodne maja rózne charakterystyki wzrostu.
Do rozwiazania problemu wykorzystamy bibliotekę `pandas`, typ `DataFrame` oraz dwa zbiory danych:

- `breast-cancer-train.dat`
- `breast-cancer-validate.dat`

Zawierają one klasę nowotworu oraz cechy, tj. charakterystyki nowotworu.
Wykorzystamy liniową oraz kwadratową metodę najmniejszych kwadratów. Dla reprezentacji kwadratowej użyjemy tylko podzbioru dostępnych danych, tj. danych z kolumn `radius (mean)`, `perimeter (mean)`, `area (mean)`, `symetry (mean)`. Do reprezentacji liniowej danych metody najmniejszych kwadratów wykorzystamy macierz

$$A_{lin} = \begin{bmatrix} f_{1,1} & \cdots & f_{1,m} \\ f_{2,1} & \cdots & f_{2,m} \\ \vdots & \cdots & \vdots \\ f_{n,1} & \cdots & f_{n,m} \end{bmatrix}$$

Reprezentacja kwadratowa:

$$A_{quad} =$$

$$\begin{bmatrix} f_{1,1}, f_{1,2}, f_{1,3}, f_{1,4}, f_{1,1}^2, f_{1,2}^2, f_{1,3}^2, f_{1,4}^2, f_{1_1}f_{1_2}, f_{1_1}f_{1_3}, f_{1_1}f_{1_4}, f_{1_2}f_{1_3}, f_{1_2}f_{1_4}, f_{1_3}f_{1_4} \\ \vdots \\ f_{n,1}, f_{n,2}, f_{n,3}, f_{n,4}, f_{n,1}^2, f_{n,2}^2, f_{n,3}^2, f_{n,4}^2, f_{n_1}f_{n_2}, f_{n_1}f_{n_3}, f_{n_1}f_{n_4}, f_{n_2}f_{n_3}, f_{n_2}f_{n_4}, f_{n_3}f_{n_4} \end{bmatrix}$$

Wagi możemy obliczyć z równania normalnego:

$$w = (A^T A)^{-1} A^T b$$

Rozwiazujac równanie normalne nalezy uzyc funkcji `solve` unikając obliczania odwrotności macierzy.

2 Specyfikacja użytego środowiska

Specyfikacja:

- Środowisko: Visual Studio Code,
- Język programowania: Python,
- System operacyjny: Microsoft Windows 11,
- Architektura systemu: x64.

3 Rozwiązanie problemu

W realizacji rozwiązania wykorzystane zostały następujące biblioteki:

```
1       import pandas as pd
2       import numpy as np
3       import matplotlib.pyplot as plt
```

Stworzone zostały reprezentacje danych zawartych w obu zbiorach dla liniowej i kwadratowej metody najmniejszych kwadratów (łącznie 4 macierze). Następnie stworzony został wektor b dla obu zbiorów, który reprezentuje rodzaj nowotworu (1 - złośliwy, -1 - łagodny).
Wagi zostały obliczone w następujący sposób:

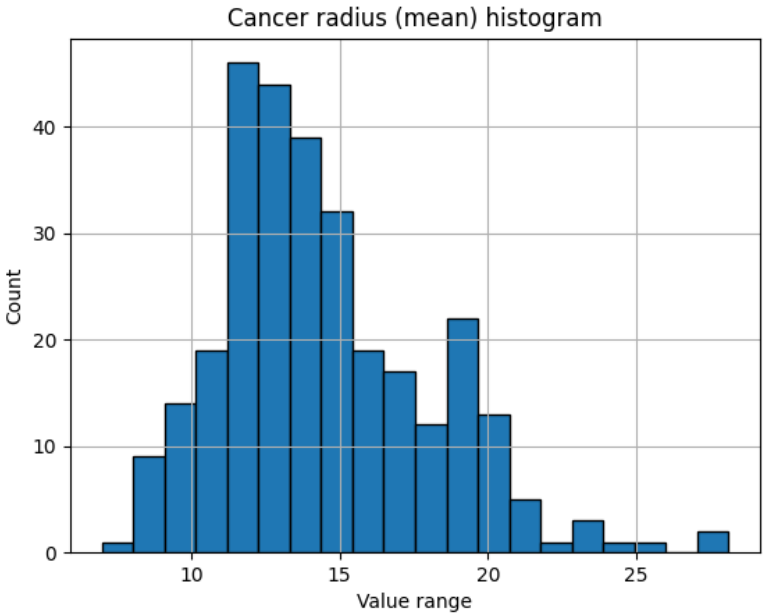
$$\texttt{np.linalg.solve}(A^T A, A^T b)$$

Obliczanie współczynników uwarunkowania macierzy:

$$\texttt{np.linalg.cond}(A^T A)$$

Można sprawdzić skuteczność przewidywań otrzymanych wag. W tym celu obliczamy wektor p przez pomnożenie reprezentacji zbiorów walidacyjnych z odpowiednimi wektorami wag. Jeżeli $p[i] > 0$, to przewidujemy, że i-ta osoba ma nowotwór złośliwy, natomiast jeśli $p[i] \leqslant 0$, to przewidujemy, że ma łagodny. Porównane zostały wektory p dla obu reprezentacji z odpowiednimi wektorami b w celu wekryfikacji poprawności.

4 Przedstawienie wyników



Rysunek 1: Histogram przedstawiający średni promień nowotworu.

Współczynniki uwarunkowania macierzy:

$$cond_{lin} \approx 1.81 \cdot 10^{12}$$
$$cond_{quad} \approx 9.06 \cdot 10^{17}$$

Zatem lepiej uwarunkowana jest macierz w przypadku metody liniowej.

Metoda liniowa

Prawidłowe rozpoznania nowotworu złośliwego : 58/60.
Prawidłowe rozpoznania nowotworu łagodnego : 194/200.
Dokładność: 96.92%.

Metoda kwadratowa

Prawidłowe rozpoznania nowotworu złośliwego : 55/60.
Prawidłowe rozpoznania nowotworu łagodnego : 185/200.
Dokładność: 92.31%.

Wyższą dokładnością charakteryzuje się metoda liniowa.

5 Wnioski

Wyniki pokazują, że regresja liniowa metodą najmniejszych kwadratów, z wysoką dokładnością (prawie 97%), przewiduje typ nowotworu na podstawie wybranych objawów. Macierz metody liniowej jest lepiej uwarunkowana, z czego może wynikać wyższa dokładność tej metody w porównaniu do metody kwadratowej. Może być ona wykorzystywana w analizie różnych zbiorów danych.