Tomasz Głąbek, 18.03.2025

MOWNIT laboratorium 2

Metoda najmniejszych kwadratów

Opis ćwiczenia

Celem zadania jest zastosowanie metody najmniejszych kwadratów do predykcji, czy nowotwór jest złośliwy (ang. malignant) czy łagodny (ang. benign). Nowotwory złośliwe i łagodne mają różne charakterystyki wzrostu. Istotne cechy to m. in. promień i tekstura. Charakterystyki te wyznaczane są poprzez diagnostykę obrazową i biopsje.

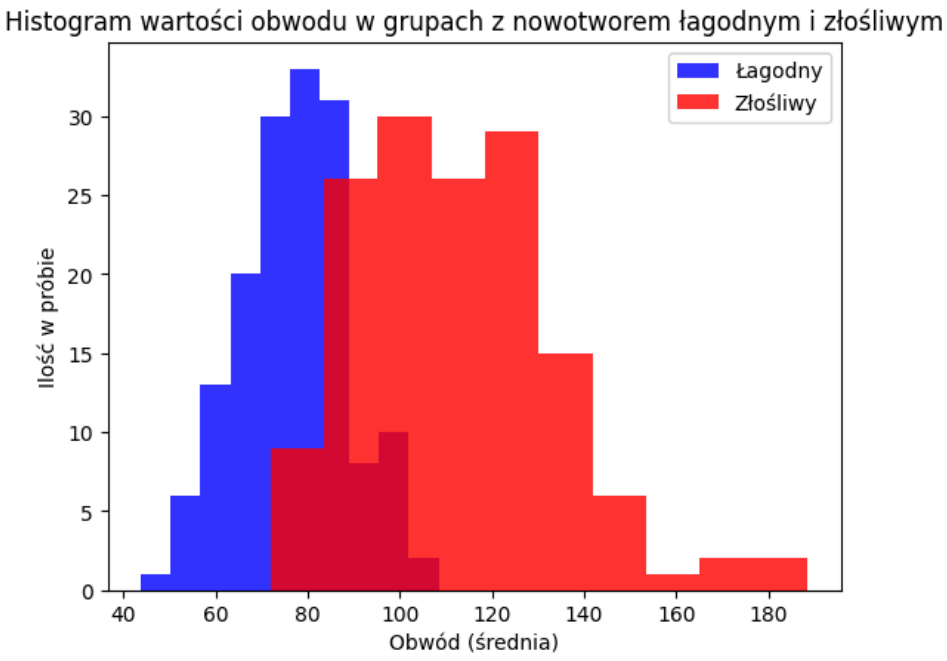
Zbiory danych wykorzystywane w zadaniu:

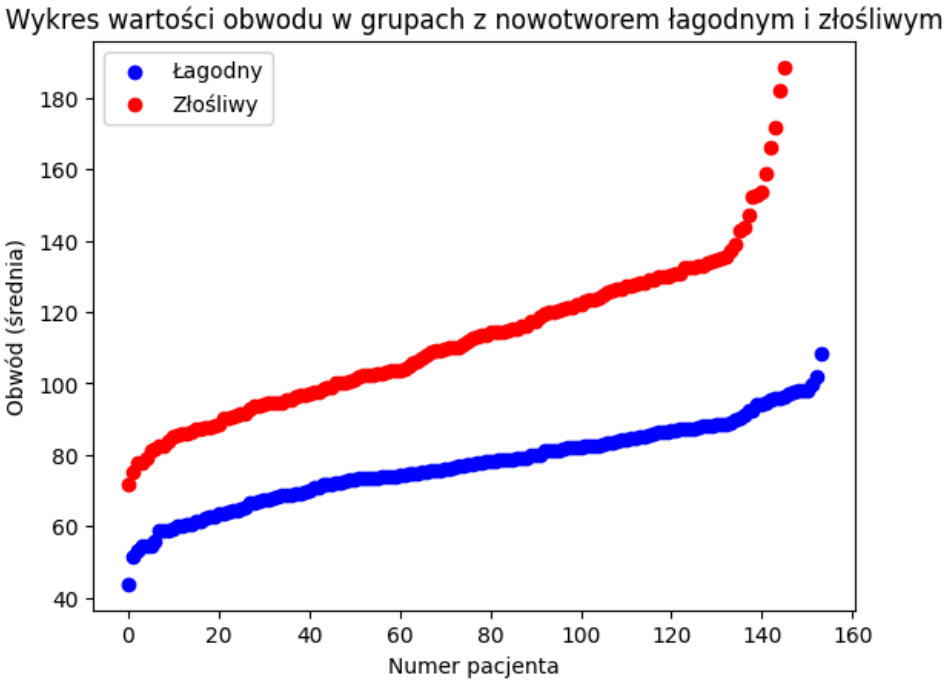
- breast-cancer-train.dat - dane służące do zbudowania modelu przewidywań
- breast-cancer-validate.dat - dane służące do weryfikacji modelu

Wykonanie zadań

Przykładowy histogram oraz wykres

Narysowany histogram oraz wykres na podstawie wybranej kolumny z danych, w tym przypadku jest to kolumna z obwodem





Metoda najmniejszych kwadratów

Macierz A

Tworzymy macierz A reprezentującą nasze dane, więc z tego co wczytaliśmy należy usunąć kolumny z id pacjenta oraz odpowiedzią jaki to typ nowotworu, ponieważ w naszym modelu nie chcemy brać pod uwagę tych parametrów.

Macierz A, reprezentacja kwadratowa

W reprezentacji kwadratowej musimy trochę przekształcić macierz A, różnice są pokazane na poniższym obrazku.

Linear representation

$$A_{\text{lin}} = \begin{bmatrix} f_{1,1} & f_{1,2} & f_{1,3} & f_{1,4} \\ f_{2,1} & f_{2,2} & f_{2,3} & f_{2,4} \\ \vdots & \vdots & \vdots & \vdots \\ f_{n,1} & f_{n,2} & f_{n,3} & f_{n,4} \end{bmatrix}$$

Quadratic representation

$$A_{\text{quad}} = \begin{bmatrix} f_{1,1}, f_{1,2}, f_{1,3}, f_{1,4}, f_{1,1}^2, f_{1,2}^2, f_{1,3}^2, f_{1,4}^2, f_{1,1}f_{1,2}, f_{1,1}f_{1,3}, f_{1,1}f_{1,4}, f_{1,2}f_{1,3}, f_{1,2}f_{1,4}, f_{1,3}f_{1,4} \\ \vdots \\ f_{n,0}, f_{n,1}, f_{n,2}, f_{n,3}, f_{n,0}^2, f_{n,1}^2, f_{n,3}^2, f_{n,4}^2, f_{n,1}f_{n,2}, f_{n,1}f_{n,3}, f_{n,1}f_{n,4}, f_{n,2}f_{n,3}, f_{n,2}f_{n,4}, f_{n,3}f_{n,4} \end{bmatrix}$$

W moim programie odpowiada za to funkcja `create_quad_rep_matrix(np.ndarray)`, która jako argument przyjmuje liniową reprezentację macierzy A.

Wektor b
Wektor b zawiera informacje o faktycznym typie nowotworu, jest budowany w następujący sposób: 1 jeśli nowotwór jest złośliwy, −1 w przeciwnym wypadku.

Wektor wag w
Chcemy wyznaczyć wektor wag, który będzie mówił jaki wpływ na wynik mają poszczególne parametry. Powinien on zatem spełniać: $Aw \approx b$. Takiego równanie może nie dać się bezpośrednio rozwiązać, ponieważ macierz A nie musi być kwadratowa, korzystamy zatem z równania normalnego.

$$Aw \approx b \tag{1}$$

$$A^T Aw \approx A^T b \tag{2}$$

$$w \approx (A^T A)^{-1} A^T b \tag{3}$$

Ostatnie równanie pozwala bezpośrednio wyliczyć wektor w, jednak zawiera ono kosztowną operację obliczania macierzy odwrotnej, zatem w moim programie korzystam z równania 2. Rozwiązuje je przy użyciu funkcji `np.linalg.solve` do samego rozwiązania oraz `np.linalg.matmul` do mnożenia macierzy.

Porównanie z metodą SVD
Wektor w można również wyliczyć funkcją `scipy.linalg.lstsq`, która pod spodem korzysta ze stabilniejszej numerycznie, ale bardziej kosztownej obliczeniowo, metody SVD. W tym przypadku dane są na tyle małe, że nie widać różnic czasowych, wyniki również są prawie identyczne.

Współczynnik uwarunkowania macierzy
Dla kwadratowej macierzy może być wyliczony ze wzoru $\text{cond}(A) = \|A\|\|A^{-1}\|$, w ogólnym wypadku liczymy $\text{cond}(A^T A)$. Można go policzyć funkcją `np.linalg.cond`. Dla reprezentacji liniowej wyniósł on około $1.8 * 10^{12}$, dla kwadratowej $9.057 * 10^{17}$. Tak duże współczynniki uwarunkowania oznaczają, że nasze zadanie jest źle uwarunkowane, czyli przy małych wahaniach w danych wejściowych otrzymamy bardzo zaburzony wynik.

Przewidywania
Tym razem ponownie korzystamy z równania $Aw \approx p$, jednak tym razem mamy już policzony wektor wag w i szukamy wektora p, który odpowie na pytanie, czy model przewiduje nowotwór złośliwy, czy łagodny. Obliczenie tego jest bardzo proste, wystarczy funkcja `np.linalg.matmul`. Obliczony wektor interpretujemy w następujący sposób:

- $p[i] > 0$ - przewidujemy nowotwór złośliwy
- $p[i] \leq 0$ - przewidujemy nowotwór łagodny

	Nowotwór złośliwy	Nowotwór łagodny
Przewidywany nowotwór złośliwy	58	6
Przewidywany nowotwór łagodny	2	194

Tabela 1: Macierz pomyłek modelu liniowego

Dokładność modelu liniowego: acc $\approx 96.923\%$

	Nowotwór złośliwy	Nowotwór łagodny
Przewidywany nowotwór złośliwy	55	15
Przewidywany nowotwór łagodny	5	185

Tabela 2: Macierz pomyłek modelu kwadratowego

Dokładność modelu kwadratowego: acc $\approx 92.308\%$

Z mojej analizy wynika, że model kwadratowy jest mniej dokładny, jednak wyniki mógł zaburzyć mały rozmiar próby.

Wnioski

Ćwiczenie pokazało, że odpowiednio zastosowana metoda najmniejszych kwadratów pozwala na bardzo dobre predykcje niektórych rzeczy, nawet pomimo wysokiego współczynnika uwarunkowania problemu. W analizie wyszła też lepsza dokładność liniowej metody najmniejszych kwadratów od kwadratowej, jednak nie mogę być pewny tego wyniku, przez niski rozmiar próby badawczej (około 300). Można poczynić ciekawą obserwację, że modele oparte za równo na liniowej jak i kwadratowej metodzie, zwracają fałszywy pozytywny wynik dużo częściej niż fałszywy negatywny. Nie jestem w stanie stwierdzić z czego to wynika, jednak wydaje się to korzystna zależność z perspektywy praktycznego zastosowania.

Bibliografia

1. Wprowadzenie do ćwiczenia `lab2-intro.pdf` zamieszczone na platformie Teams.