Metoda najmniejszych kwadratów

Sprawozdanie z laboratorium 2

Jakub Grześ

15.03.2024

1 Treść zadań

Celem zadania jest zastosowanie metody najmniejszych kwadratów do predykcji, czy nowotwór jest złośliwy (ang. malignant) czy łagodny (ang. benign). Nowotwory złośliwe i łagodne maja różne charakterystyki wzrostu. Istotne cechy to m. in. promień i tekstura. Charakterystyki te wyznaczane są poprzez diagnostykę obrazowa i biopsje. Do rozwiązania problemu wykorzystamy bibliotekę pandas, typ DataFrame oraz dwa zbiory danych: – breast-cancer-train.dat – breast-cancer-validate.dat. Nazwy kolumn znajdują się w pliku breast-cancer.labels. Pierwsza kolumna to identyfikator pacjenta patient ID. Dla każdego pacjenta wartość w kolumnie Malignant/Benign wskazuje klasę, tj. czy jego nowotwór jest złośliwy czy łagodny. Pozostałe 30 kolumn zawiera cechy, tj. charakterystyki nowotworu.

(a) Otwórz zbiory breast-cancer-train.dat i breast-cancer-validate.dat używając funkcji pd.io.parsers.read_csv z biblioteki pandas.
(b) Stwórz histogram i wykres wybranej kolumny danych przy pomocy funkcji hist oraz plot. Pamiętaj o podpisaniu osi i wykresów.
(c) Stwórz reprezentacje danych zawartych w obu zbiorach dla liniowej i kwadratowej metody najmniejszych kwadratów (łącznie 4 macierze). Dla reprezentacji kwadratowej użyj tylko podzbioru dostępnych danych, tj. danych z kolumn radius (mean), perimeter (mean), area (mean), symmetry (mean).
(d) Stwórz wektor b dla obu zbiorów (tablice numpy 1D-array o rozmiarze identycznym jak rozmiar kolumny Malignant/Benign odpowiedniego zbioru danych). Elementy wektora b to 1 jeśli nowotwór jest złośliwy, -1 w przeciwnym wypadku. Funkcja np.where umożliwi zwięzłe zakodowanie wektora b.
(e) Znajdź wagi dla liniowej oraz kwadratowej reprezentacji najmniejszych kwadratów przy pomocy macierzy A zbudowanych na podstawie zbioru breast-cancer-train.dat. Potrzebny będzie także wektor b zbudowany na podstawie zbioru breast-cancer-train.dat. Uwaga. Problem najmniejszych kwadratów należy rozwiązać stosując równanie normalne (tj. nie używając funkcji scipy.linalg.lstsq). Rozwiązując równanie normalne należy użyć funkcji solve, unikając obliczania odwrotności macierzy funkcja scipy.linalg.pinv.
(f) Oblicz współczynniki uwarunkowania macierzy, cond($A^T A$), dla liniowej i

kwadratowej metody najmniejszych kwadratów.
(g) Sprawdź jak dobrze otrzymane wagi przewidują typ nowotworu (łagodny czy złośliwy). W tym celu pomnóż liniowa reprezentacje zbioru breast-cancer-validate.dat oraz wyliczony wektor wag dla reprezentacji liniowej. Następnie powtórz odpowiednie mnożenie dla reprezentacji kwadratowej. Zarówno dla reprezentacji liniowej jak i kwadratowej otrzymamy wektor p. Zakładamy, ze jeśli p[i] > 0, to i-ta osoba (prawdopodobnie) ma nowotwór złośliwy. Jeśli p[i] $>$ 0 to i-ta osoba (prawdopodobnie) ma nowotwór łagodny. Porównaj wektory p dla reprezentacji liniowej i kwadratowej z wektorem b (użyj reguł p[i] > 0 oraz p[i] < 0). Oblicz liczbę fałszywie dodatnich (ang. false-positives) oraz fałszywie ujemnych (ang. false-negatives) przypadków dla obu reprezentacji. Przypadek fałszywie dodatni zachodzi, kiedy model przewiduje nowotwór złośliwy, gdy w rzeczywistości nowotwór był łagodny. Przypadek fałszywie ujemny zachodzi, kiedy model przewiduje nowotwór łagodny, gdy w rzeczywistości nowotwór był złośliwy.

2 Metoda rozwiązania problemu

2.1 Reprezentacja danych

Przygotowano macierze reprezentujące parametry zdrowotne pacjentów. Dane reprezentowano w dwóch postaciach. Liniowej,

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

oraz kwadratowej, gdzie dla uniknięcia dużej wielkości macierzy powodującej dużą ilość obliczeń, uwzględniono jedynie 4 parametry t.j. radius (mean), perimeter (mean), area (mean), symmetry (mean).

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

Takie macierze przygotowano zarówno dla treningowego zestawu danych, jak i walidacyjnego. Wygenerowano dodatkowo 2 wektory, których elementami były 1, jeśli nowotwór dla danego pacjenta okazał się złośliwy, oraz -1 w przeciwnym przypadku.

2.2 Znalezienie wektora wag

Dążymy do znalezienia takiego wektora w dla którego: $Aw \approx b$, gdzie b jest wektorem rozwiązań, a A macierzą cech.

Funkcją kosztu jest $||Aw - b||^2$ i chcemy ją minimalizować. Rozpoczynając od równania $Aw = b$, mnożymy obie strony równania z lewej strony przez transponowaną macierz A, otrzymując $A^T Aw = A^T b$. Jeśli macierz $A^T A$ jest odwracalna, możemy pomnożyć obie strony równania przez inwersję tej macierzy, otrzymując $w = (A^T A)^{-1} A^T b$.

2.3 Współczynnik uwarunkowania

Współczynnik uwarunkowania mówiący o tym, jak błąd danych wejściowych jest wzmacniany przez obliczenia, został obliczony funkcją biblioteczną Numpy linalg.cond.

3 Przygotowanie danych, obliczenia i fragmenty algorytmu

Po wczytaniu danych oraz ich organizacji w odpowiednie macierze przygotowano przykładowy histogram oraz wykres dla kolumny 'area (mean)'.

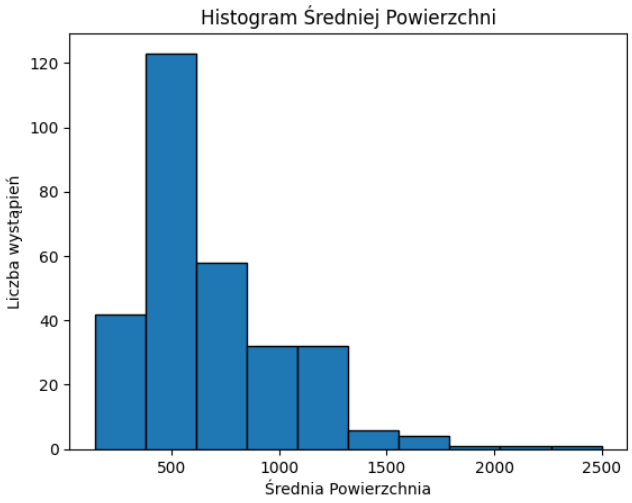

Figure 1: Histogram danych dla kolumny 'area (mean)'

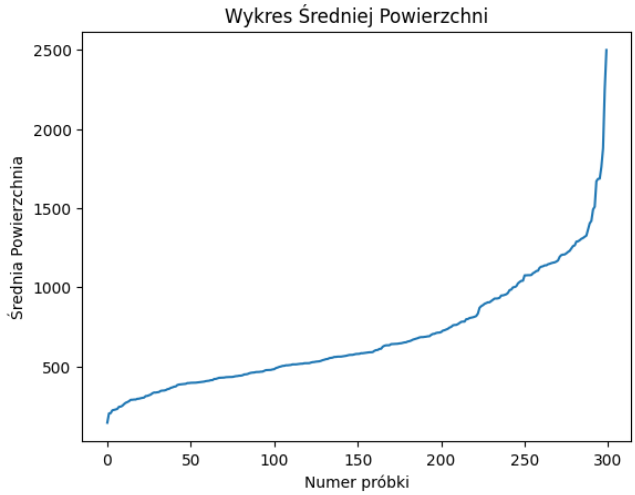

Figure 2: Wykres danych dla kolumny 'area (mean)'

3.1 Obliczanie współczynnika uwarunkowania macierzy

Podając macierze postaci $A^T A$ odpowiednio dla reprezentacji liniowej i kwadratowej jako argumenty funkcji linalg.cond, otrzymano następujące wyniki.

Współczynnik uwarunkowania dla reprezentacji liniowej: 1.8092e+12.
Współczynnik uwarunkowania dla reprezentacji kwadratowej: 9.0568e+17.

Współczynniki uwarunkowania są bardzo wysokie dla obu reprezentacji, jednak dla reprezentacji kwadratowej jest znacznie wyższy niż dla liniowej. Oznacza to, że ta reprezentacja jest bardziej wrażliwa na błędy reprezentacji danych wejściowych. Wskaźnik uwarunkowania macierzy w równaniu $Ax = b$ mówi o tym, jak zmiana normy macierzy A wpłynie na normę rozwiązania x oraz jak ma się stosunek błędu względnego x do błędu względnego b a to właśnie rozwiązanie tego typu równania jest przedmiotem tego laboratorium.

3.2 Obliczenie wektora wag

Korzystając z funkcji linalg.solve rozwiązującej układ równań $Ax = b$ ze względy na x, oraz wykorzystując równanie $(A^T A)w = A^T b$, możemy obliczyć wektor w podając jako argumenty odpowiednio macierze $(A^T A)$ i $A^T b$. Ten wektor jest wrażliwy na wysoki współczynnik uwarunkowania macierzy $(A^T A)$.

```
w_linear = np.linalg.solve(linear_train.T @ linear_train,
    linear_train.T @ b_train)
w_quadratic = np.linalg.solve(quadratic_train.T @
   quadratic_train, quadratic_train.T @ b_train)
```

3.3 Dokładność przewidywań

Celem sprawdzenia jak dobrze udaje się przewidzieć, korzystając z metody najmniejszych kwadratów rodzaj nowotworu, pomnożono wektory wag przez reprezentacje kwadratowe i liniowe, ale tym razem dla zbiorów walidacyjnych, które nie były brane pod uwagę podczas budowy wektora wag. Przyjęto, że jeśli i-ty element wektora jest większy od 0 to pacjent i-ty, prawdopodobnie cierpi na nowotwór złośliwy i odpowiednio dla łagodnego. Następnie porównano rezultaty z kolumną diagnoz pobraną ze zbioru walidacyjnego i sprawdzono, w jakim procencie przypadków udało się poprawnie przewidzieć rodzaj nowotworu.

```
# Wyliczenie wektora p
p_linear = w_linear @ linear_validate_feature.T
p_quadratic = w_quadratic @ quadratic_validate_feature.T

# Interpretacja wynikow
predictions_linear = np.where(p_linear > 0, 1, -1)
predictions_quadratic = np.where(p_quadratic > 0, 1, -1)

# Oblieczanie trafnosci
accuracy_linear = np.where(predictions_linear ==
    b_validate, 1, 0).mean()
accuracy_quadratic = np.where(predictions_quadratic ==
    b_validate, 1, 0).mean()
```

Dokładność predykcji dla reprezentacji liniowej: 96.92%.
Dokładność predykcji dla reprezentacji kwadratowej: 92.31%.

Jak widać odnotowana wysoką skuteczność dla obu reprezentacji, przy czym nieznacznie wyższą dla liniowej. Obliczono także liczbę wyników fałszywie pozytywnych i negatywnych dla obu zestawów.

Reprezentacja	Liniowa	Kwadratowa
Przypadki fałszywie pozytywne	6	15
Przypadki fałszywie negatywne	2	5

Table 1: Analiza przypadków fałszywie pozytywnych i negatywnych

4 Wnioski

Mimo stosunkowo prostej metody udało się opracować efektywną strategię przewidywania typu nowotworu. Warto zwrócić uwagę na istotność doboru cech, w przypadku reprezentacji liniowej uwzględniono 30 czynników, natomiast dla reprezentacji kwadratowej ograniczono się jedynie do 4, co skutkowało jedynie nieco niższą skutecznością przy znaczącym ograniczeniu liczby obliczeń.

Kluczowym czynnikiem przyczyniającym się do wysokiej skuteczności był obszerny zestaw danych treningowych oraz ograniczenie problemu do przewidywania jednej cechy - typu nowotworu. Wysoki współczynnik uwarunkowania dla obu reprezentacji sugeruje, że główną słabością reprezentacji kwadratowej jest jej wysoka wrażliwość na dokładność danych wejściowych. W sytuacji, gdy dostępnych byłoby mniej danych lub cele analizy byłyby bardziej złożone, model mógłby wykazać mniejszą dokładność. Dlatego w przyszłych pracach istotne będzie zwrócenie szczególnej uwagi na jakość danych, wyselekcjonowanie tych najbardziej istotnych, poszukiwanie wartościowych kombinacji czynników oraz dobór odpowiedniej metody do danego problemu.

Bibliografia

[1] Marcin Kuta, *Least squares method*

[2] Qingkai Kong, Timmy Siauw, Alexandre Bayen *Python Programming and Numerical Methods*