Data: 12.03.2025

Laboratorium 2

Autorzy: Kosman Mateusz, Ludwin Bartosz

Temat laboratorium

Celem niniejszego zadania było zastosowanie metody najmniejszych kwadratów do predykcji typu nowotworu, czyli określenia, czy badany guz jest złośliwy, czy łagodny. W ramach sprawozdania przeprowadzono analizę danych diagnostycznych, obejmujących między innymi cechy takie jak promień, obwód, pole powierzchni czy symetria. Do opracowania modelu wykorzystamy zarówno liniową, jak i kwadratową reprezentację problemu, przy czym zastosowano różne metody wyznaczania wag – rozwiązanie równania normalnego oraz metodę opartą na rozkładzie SVD. Poprzez wyniki eksperymentów ocenimy dokładność klasyfikacji poprzez analizę macierzy pomyłek, co pozwoli na wyciągnięcie wniosków dotyczących wpływu współczynnika uwarunkowania macierzy na stabilność numeryczną wyznaczonych wag.

Treść

a. Otwórz zbiory breast-cancer-train.dat i breast-cancer-validate.dat używając funkcji z biblioteki pandas.
b. Stwórz histogram (z rozróżnieniem na typ nowotworu) i wykres wybranej kolumny danych przy pomocy funkcji hist oraz plot. W przypadku wykresu posortuj wartości kolumny od najmniejszej do największej.
c. Stwórz reprezentacje danych zawartych w obu zbiorach dla liniowej i kwadratowej metody najmniejszych kwadratów. 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. Elementy wektora b to 1 jeśli nowotwór jest złośliwy, -1 w przeciwnym wypadku.
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.

f. Znajdź odrębny zbiór wag dla reprezentacji liniowej, używając funkcji która stosuje rozkład SVD do rozwiązania problemu oraz zbiór wag dla zregularyzowanej reprezentacji liniowej, rozwiązując równanie normalne oraz stosując λ = 0.01.
g. Oblicz współczynniki uwarunkowania macierzy, dla liniowej kwadratowej metody najmniejszych kwadratów, wyznaczone na zbiorze treningowym.
h. Sprawdź jak dobrze otrzymane wagi przewidują typ nowotworu (łagodny czy złośliwy).

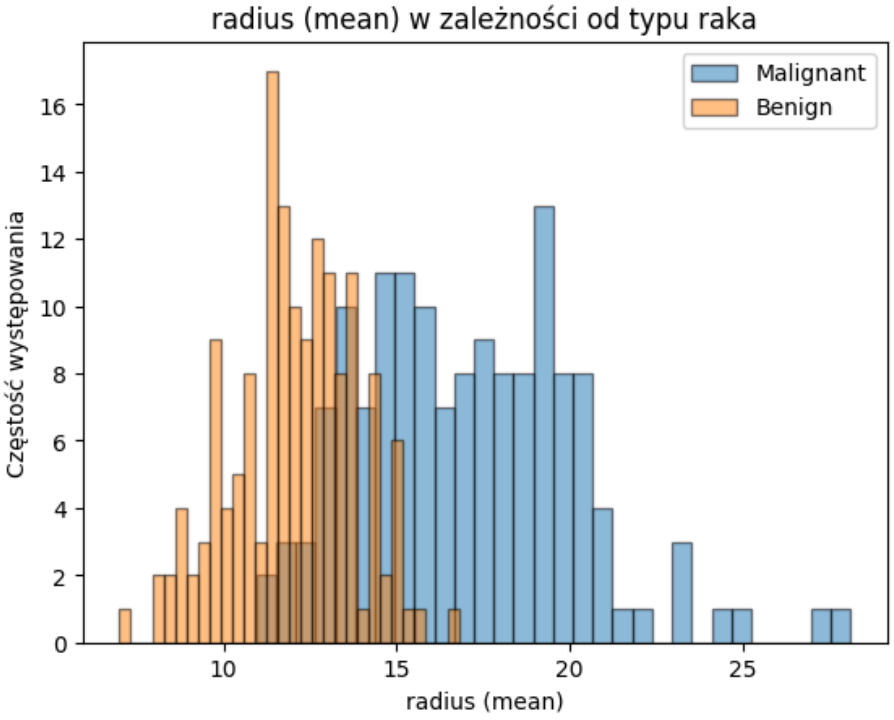
Rozwiązanie

Analiza danych

Pierwszym krokiem w drodze do stworzenia modelu i jego oceny jest analiza danych pozyskanych z plików:

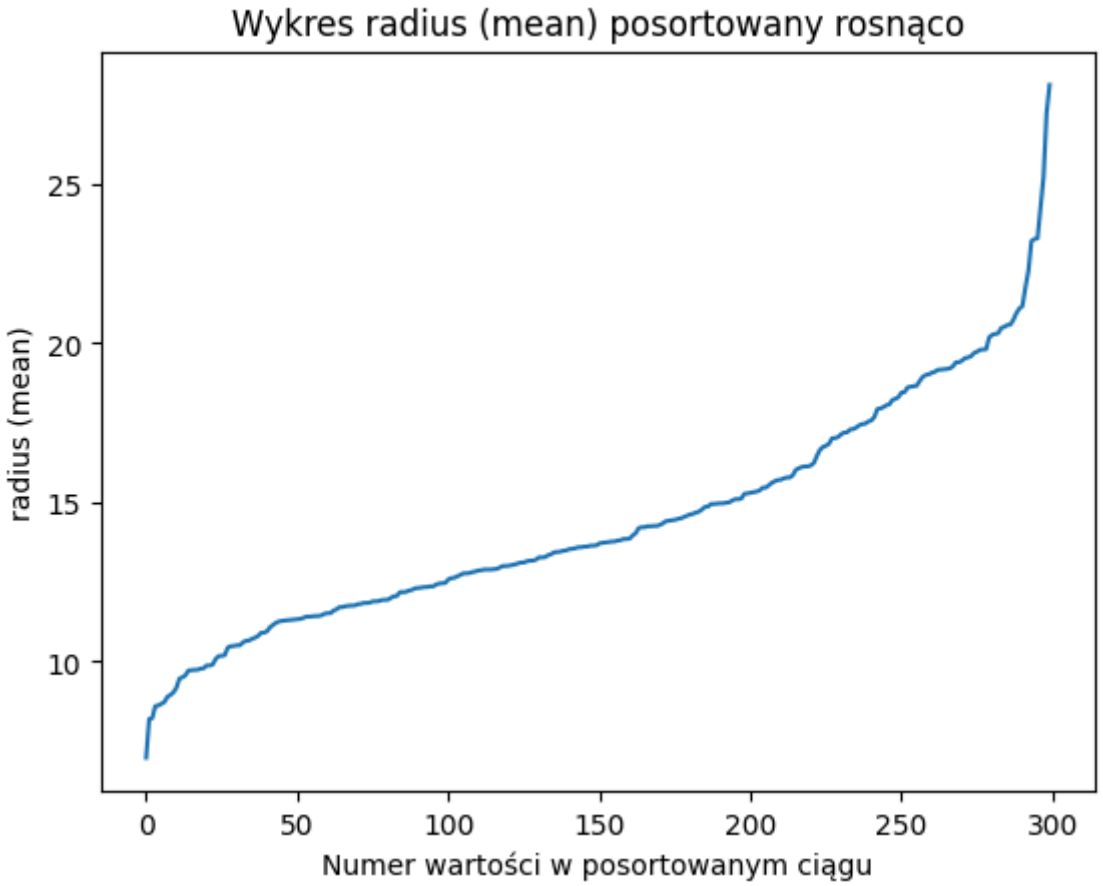
- *breast-cancer-train.dat* - zbiór danych treningowych dla modelu.
- *breast-cancer-validate.dat* - zbiór służący walidacji wyników modelu.
- *breast-cancer.labels* - zbiór etykiet dla poszczególnych kolumn w powyższych zbiorach.

Przeanalizujmy histogram promienia raka (średnia) w zależności od jego typu (złośliwy czy łagodny):



Wykres 1

Oraz wykres z wartościami wyżej wymienionej cechy posortowanych w kolejności rosnącej:



Wykres 2

Na wykresie nr 1 widać wyraźne rozróżnienie obu wersji choroby w zależności od rozmiaru - rak złośliwy ma tendencję do posiadania większego promienia niż łagodny. Po kształcie wykresu 2 możemy wywnioskować natomiast że wartości skrajne (takie jak poniżej 5 oraz powyżej 23) zdarzają się z bardzo niskim prawdopodobieństwem.
Dostrzegalny podział zaobserwowany na rysunku 1 oraz ograniczona liczba skrajnych wartości pozwalają wysunąć nam wniosek, iż model ma istotną szansę wykazywać się dużą skutecznością w rozróżnianiu typu raka, gdyż wartości są skoncentrowane wokół charakterystycznych dla poszczególnych grup przedziałów, co umożliwi precyzyjne wyznaczenie progów decyzyjnych przy klasyfikacji nowych przypadków, a wartości skrajne mogące zakłamywać wynik mają nikłą szansę wystąpienia.

Metoda najmniejszych kwadratów

– Po uprzedniej analizie wstępnej i konkluzji o dobrych rokowaniach modelu przystępujemy do jego utworzenia. W tym celu tworzymy macierz liniową oraz kwadratową reprezentacji danych. W przypadku kwadratowej użyjemy tylko podzbioru cech - *'radius (mean)', 'perimeter (mean)', 'area (mean)', 'symmetry (mean)'.*

– Następnie stworzymy wektor b dla obu zbiorów. Będziemy trzymać się oznaczenia że jeśli nowotwór jest złośliwy to wartość dla danego indeksu przyjmuje liczbę 1, -1 w przeciwnym wypadku.

– W kolejnym kroku znajdujemy odpowiednie wagi dla obu reprezentacji najmniejszych kwadratów. Poszukujemy wektora ω, dla którego iloczyn $A\omega$ jak najlepiej przybliża wektor b. Innymi słowy, chcemy zminimalizować wartość normy $\|Aw-b\|$. Ponieważ macierz A w ogólności nie jest kwadratowa, nie możemy bezpośrednio zastosować funkcji $linalg.solve()$ z biblioteki $numpy$ (odwrotność macierzy A nie istnieje). Aby obejść ten problem, korzystamy z tzw. równania normalnego. Istotą podejścia jest przemnożenie obu stron równania $A\omega = b$ przez macierz transponowaną A^T, dzięki czemu powstaje układ:

$$A^T A\omega = A^T b. \tag{1}$$

Rozwiązanie przybiera wtedy postać:

$$\omega = (A^T A)^{-1} A^T b \tag{2}$$

gdzie $(A^T A)^{-1} A^T$ pełni rolę macierzy pseudoodwrotnej. Taki sposób znajdowania ω w może być jednak wrażliwy na uwarunkowanie macierzy A, więc przy pewnych postaciach A pojawiają się trudności numeryczne.

W programie, zamiast bezpośrednio obliczać $(A^T A)^{-1}$, możemy skorzystać z odpowiedniej funkcji rozwiązującej układ (1) co jest równoważne wyliczeniu wzoru (2).

W celu uzyskania porównania znajdziemy również odrębny zbiór wag używając rozkładu *SVD* do rozwiązania problemu oraz zbiór wag dla zregularyzowanej reprezentacji liniowej, rozwiązując równanie normalne. Oto kod który realizuje szukanie obu zbiorów:

```python
def compute_weights(self, A, b):
    """
    Oblicza wagi metodą najmniejszych kwadratów przy użyciu równania normalnego (A^T A) w = A^T b
    """
    ATA = A.T @ A
    ATb = A.T @ b
    weights = np.linalg.solve(ATA, ATb)
    return weights

def compute_weights_SVD(self, A, lambda_coeff, b):
    """
    Oblicza wagi metodą najmniejszych kwadratów przy użyciu równania normalnego (A^T A + λI) w = A^T b
    korzystając z funkcji scipy.linalg.lstsq stosującej rozkład SVD.
    """
    ATA = A.T @ A
    ATA += lambda_coeff*np.identity(A.shape[1])
    ATb = A.T @ b
    weights = scipy.linalg.lstsq(ATA, ATb)[0]
    return weights
```

Sprawdzanie trafności przewidywań

Wskaźnik uwarunkowania macierzy

Wskaźnik uwarunkowania macierzy definiujemy jako:

$$\kappa(A) = cond(A) = ||A|| \cdot ||A^{-1}||$$

Ma on własność:

$$cond(A^T A) = cond(A)^2$$

Wyznaczamy go korzystając z funkcji *np.linalg.cond* biblioteki *numpy*.

Wyznaczony wskaźnik uwarunkowania dla macierzy metody liniowej wynosi $\sim 1.34 \times 10^6$, podczas gdy dla macierzy metody kwadratowej wynosi on $\sim 9.51 \times 10^8$.

Wskaźnik uwarunkowania określa stosunek błędu reprezentacji numerycznej danych wejściowych do błędu wyniku, stąd duży wskaźnik uwarunkowania oznacza, że błąd wyniku również będzie znaczny. Widząc, że wskaźnik uwarunkowania dla macierzy metody kwadratowej jest ok. 700 razy większy od wskaźnika uwarunkowania dla macierzy metody liniowej, można oczekiwać, że korzystanie z metody liniowej da trafniejszy wynik.

Macierz pomyłek i dokładność metody

Znaczenie komórek w podanych tabelach:

- True/False - wartości prawdziwe dostarczone z zestawem danych
- Positive/Negative - wartość przewidziana na podstawie obliczonych wag

Dokładność jest liczona ze wzoru:

$acc = \frac{TP+TN}{TP+TN+FP+FN}$, gdzie

- TP – liczba przypadków prawdziwie dodatnich
- TN – liczba przypadków prawdziwie ujemnych
- FP – liczba przypadków fałszywie dodatnich
- FN – liczba przypadków fałszywie ujemnych

Macierz metody liniowej:

- Dokładność: 0.969

	Positive	Negative
True	58	2
False	6	194

Tabela 1 - macierz pomyłek metody liniowej

Macierz metody liniowej z rozkładem SVD:

- Dokładność: 0.969

	Positive	Negative
True	58	2
False	6	194

Tabela 2 - macierz pomyłek metody liniowej z rozkładem SVD

Macierz metody liniowej z rozkładem SVD oraz reprezentacją zregularyzowaną:

- Dokładność: 0.973

	Positive	Negative
True	57	3
False	4	196

Tabela 3 - macierz pomyłek metody liniowej z rozkładem SVD oraz reprezentacją zregularyzowaną

Macierz metody kwadratowej:

- Dokładność: 0.923

	Positive	Negative

True	55	5
False	15	185

Tabela 4 - macierz pomyłek metody kwadratowej

Wnioski

- Wskaźnik uwarunkowania dla macierzy metody liniowej był niższy niż ten dla macierzy metody kwadratowej, co sugerowało, że metoda liniowa najmniejszych kwadratów zwróci dokładniejszy wynik. Hipoteza ta sprawdziła się, gdyż dokładność metody liniowej (ok. 0.97) jest większa niż dokładność metody kwadratowej (ok. 0.92). W związku z tym wskaźnik uwarunkowania umożliwia postawienie trafnej hipotezy odnośnie dokładności obliczeń.
- W rozważanym przypadku metoda liniowa dała jednakową dokładność niezależnie od zastosowania rozkładu SVD. Zastosowanie reprezentacji zregularyzowanej umożliwiło jednak wyznaczenie trafniejszych wag. Oznacza to, że dla pewnych przypadków użycie rozkładu SVD lub reprezentacji zregularyzowanej może dać bardziej pożądany rezultat.
- Metoda najmniejszych kwadratów dla zadanego problemu umożliwiła nam klasyfikację nowotworu jako złośliwego lub łagodnego mając do dyspozycji wybrane cechy tego nowotworu; dokładność tej metody wyniosła ok. 92.3-97.3% zależnie od zastosowanego algorytmu liczenia wag.

Źródła

- Prezentacje z MS Teams (zespół MOwNiT 2025)
- https://pl.wikipedia.org/wiki/Wska%C5%BAnik_uwarunkowania