Maciej Borowiec
19.03.2025

Laboratorium 2
Metoda najmniejszych kwadratów

1. Wprowadzenie

Celem ćwiczenia było zrobienie zadania, w którym należało skorzystać z metody najmniejszych kwadratów. Należało wyznaczyć charakterystyczne wielkości dla tej metody. Następnie używając równań na macierzach wyznaczyć wektory wag dla danych cech danych i użyć ich do predykcji wyników dla każdego zbioru danych. Na koniec te wartości należało porównać z wartościami prawdziwymi, by zobaczyć efektywność tej metody.

2. Zadanie

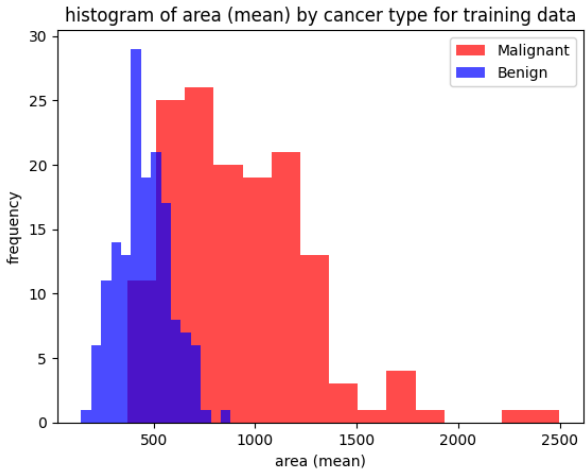
W zadaniu mamy do analizy dane na temat nowotworów piersi dla różnych pacjentów. Mamy użyć metodę najmniejszych kwadratów do predykcji czy dany nowotwór jest złośliwy (ang. malignant), czy łagodny (ang. benign). Dane są dane w 3 plikach:

- **breast-cancer-train.dat** - dane treningowe
- **breast-cancer-validate.dat** - dane walidujące
- **breast-cancer.labels** - nazwy kolumn cech nowotworu

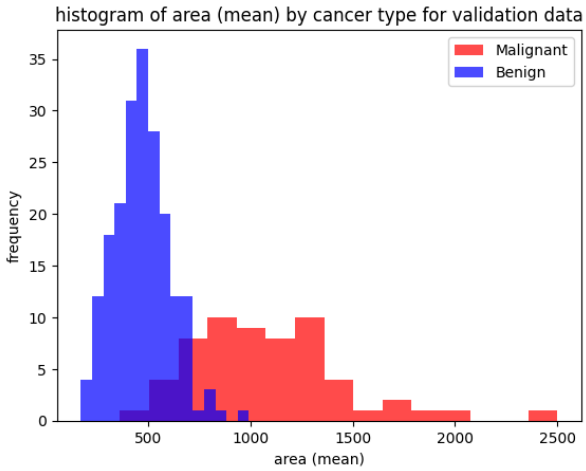
Wczytujemy te dane, by później ich użyć do analizy.

2.1. Przedstawienie danych na wykresie

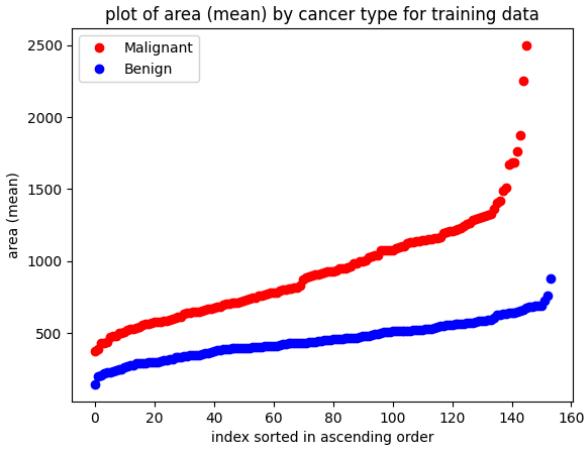
Na początku przedstawimy przykładową cechę nowotworu na histogramach i wykresach w zależności od typu nowotworu dla obu zbiorów danych. Za przedstawioną cechę wybrałem średnie pole powierzchni (area (mean)). Na rysunkach poniżej przedstawionych (rys. 1-4) łatwo zauważyć, że średnie pole powierzchni jest znacząco większe dla nowotworów złośliwych. Możemy również zauważyć, że dane walidujące mają dużo więcej pacjentów z nowotworem łagodnym niż z nowotworem złośliwym, gdzie pacjenci w grupie treningowej mają mniej więcej tę samą częstotliwość występowania obu typów nowotworu.



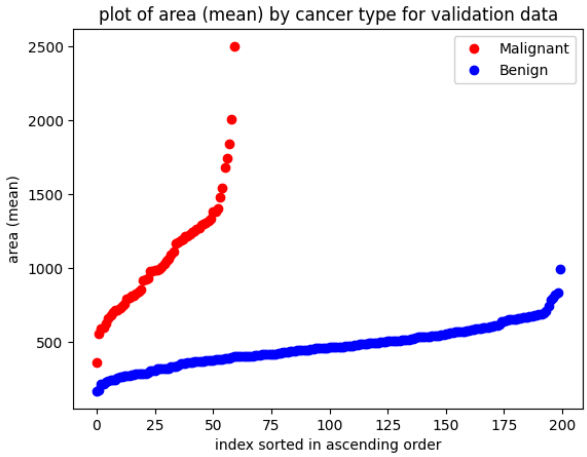
Rys. 1. Histogram średniego pola powierzchni w zależności od typu nowotworu dla danych testowych



Rys. 2. Histogram średniego pola powierzchni w zależności od typu nowotworu dla danych walidujących



Rys. 3. Wykres średniego pola powierzchni w zależności od typu nowotworu dla danych testowych



Rys. 4. Wykres średniego pola powierzchni w zależności od typu nowotworu dla danych walidujących

2.2. Metoda najmniejszych kwadratów

2.2.1. Utworzenie potrzebnych reprezentacji zbiorów danych

Na początek należało stworzyć reprezentacje liniowe i kwadratowe macierzy dla obu zbiorów danych. Przy tworzeniu reprezentacji kwadratowej należało się ograniczyć jedynie do czterech parametrów nowotworu (*radius (mean), perimeter (mean), area (mean), symmetry (mean)*), gdzie przy reprezentacji liniowej użyto wszystkich 30 parametrów.

Należało również stworzyć wektory o wartościach reprezentujących czy dany pacjent ma nowotwór złośliwy czy łagodny (zwany dalej wektorem typów) dla obu zbiorów danych. Oznaczono 1 jako nowotwór złośliwy, a -1 jako nowotwór łagodny. Wektor typów dla danych treningowych będzie użyty w równaniu macierzowym przy wyznaczaniu wag, a wektor danych walidacyjnych przy walidacji otrzymanych wyników.

2.2.2. Znalezienie wektorów wag

Następnie należało znaleźć wektory wag w zależności od użytej metody. Znaleziono podane wektory wag:

- wektor dla reprezentacji liniowej - z równania $A^T Aw = A^T y$, gdzie
  - A – macierz danej reprezentacji dla danych testowych
  - w – wektor wag
  - y – wektor typów dla danych testowych
- wektor dla reprezentacji kwadratowej - również z równania podanego wyżej
- wektor dla reprezentacji liniowej z użyciem SVD - przy użyciu funkcji **scipy.linalg.lstsq**
- wektor dla zregularyzowanej reprezentacji liniowej - użyto zmodyfikowanego równania: $(A^T A + \lambda I)w = A^T y$, gdzie za λ przyjęto 0.01

2.2.2. Obliczenie współczynników uwarunkowania

Do sprawdzenia jakości uwarunkowania użyłem współczynnika uwarunkowania dla obu reprezentacji macierzy danych treningowych. Do obliczeń została wykorzystana funkcja **numpy.linalg.cond**. Otrzymane współczynniki zostały podane w tabeli poniżej:

Reprezentacja macierzy	**Współczynnik uwarunkowania**
Liniowa	$1.809248 \cdot 10^{12}$
Kwadratowa	$9.056815 \cdot 10^{17}$

Tab. 1. Wartości współczynnika uwarunkowania dla obu reprezentacji macierzy danych treningowych

Z tabeli (tab. 1) widać, że reprezentacja liniowa jest lepiej uwarunkowana niż reprezentacja kwadratowa. Obie macierze jednak mają duży wskaźnik uwarunkowania co w zależności od jakości danych może prowadzić do dużych błędów numerycznych.

2.2.3. Predykcja typu nowotworu

Na koniec użyjemy tych danych do predykcji typu nowotworu i porównamy wartości otrzymane z wartościami oczekiwanymi, by ocenić precyzję metody. Aby to zrobić mnożymy daną reprezentację macierzy przez wektor otrzymanych wag. Otrzymujemy wektor p i zakładamy, że jeśli $p[i] > 0$ to i-ta osoba ma nowotwór złośliwy, a w przeciwnym przypadku nowotwór łagodny. Porównujemy wyniki w tym wektorze z wektorem typów danych walidacyjnych, by otrzymać dokładność metody. Otrzymane wyniki są przedstawione w poniżej przedstawionych macierzach pomyłek:

Reprezentacja liniowa		
	Faktycznie pozytywny	Faktycznie negatywny
Predykcyjnie pozytywny	TP - 58	FP - 6
Predykcyjnie negatywny	FN - 2	TN - 194
Dokładność: 96.923%		

Tab. 2. Macierz pomyłek dla reprezentacji liniowej

Reprezentacja kwadratowa		
	Faktycznie pozytywny	Faktycznie negatywny
Predykcyjnie pozytywny	TP - 55	FP - 15
Predykcyjnie negatywny	FN - 5	TN - 185
Dokładność: 92.308%		

Tab. 3. Macierz pomyłek dla reprezentacji kwadratowej

Reprezentacja liniowa (SVD)		
	Faktycznie pozytywny	Faktycznie negatywny
Predykcyjnie pozytywny	TP - 58	FP - 6
Predykcyjnie negatywny	FN - 2	TN - 194
Dokładność: 96.923%		

Tab. 4. Macierz pomyłek dla reprezentacji liniowej (SVD)

Zregularyzowana reprezentacja liniowa		
	Faktycznie pozytywny	Faktycznie negatywny
Predykcyjnie pozytywny	TP - 55	FP - 1
Predykcyjnie negatywny	FN - 5	TN - 199
Dokładność: 97.692%		

Tab. 5. Macierz pomyłek dla zregularyzowanej reprezentacji liniowej

Z danych macierzy pomyłek (tab. 2-5) można wywnioskować, że zregularyzowana reprezentacja liniowa ma największą precyzję w tym przypadku. W tym przypadku, bardziej stabilna numerycznie dekompozycja SVD nie miała wpływu na dokładność reprezentacji liniowej. Natomiast reprezentacja kwadratowa okazała się mieć najmniejszą dokładność z wszystkich.

5. Podsumowanie

W podanym zadaniu należało zapoznać i zastosować metodę najmniejszych kwadratów, która jest bardzo popularnym narzędziem używanym w regresji liniowej. Zadanie pokazało różne modele, które można zastosować w tej metodzie i to, że wyniki mogą się różnić dokładnością w zależności od wybranego modelu. W przypadku tego zadania reprezentacja kwadratowa okazała się być najmniej dokładna co mogło być spowodowane ograniczeniem się do jedynie czterech cech przy użyciu tej reprezentacji. Jednakże nie zawsze tak musi być i należy dobierać stosowne rozwiązania w zależności od danego problemu.

6. Blbliografia

- Materiały zamieszczone wraz zadaniem