

Laboratorium 02 - Metoda najmniejszych kwadratów

Dawid Żak

Szymon Hołysz

2025-03-18

Table of contents

Treść zadania	1
Próbka ze zbioru danych	1
Ponizej znajduje się histogram i wykres posortowany rosnąco dla cechy radius (mean) dla próbek złośliwych i łagodnych	1
Tworzenie wektora wag	3
Wykorzystanie rozkładu SVD	3
Współczynniki uwarunkowania	3
Predykcja	4
Wnioski	4

Treść zadania

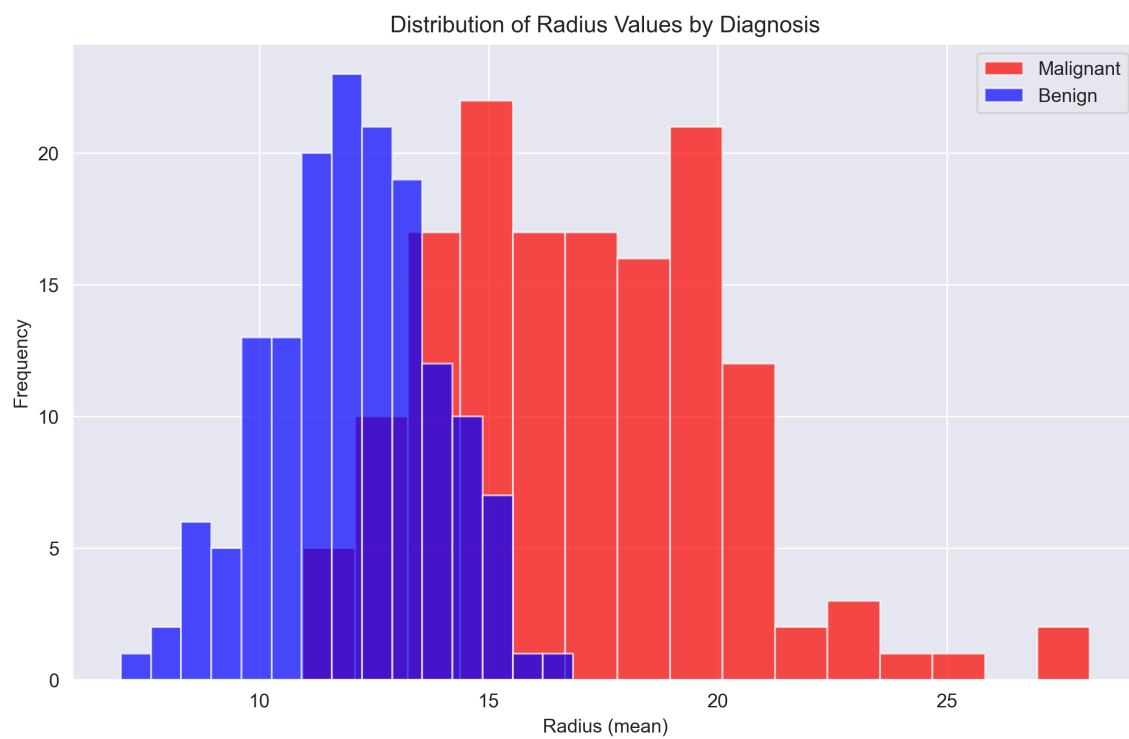
Celem zadania jest zastosowanie metody najmniejszych kwadratów do predykcji, czy nowotwór jest złośliwy (ang. malignant) czy łagodny (ang. benign). Nowotwory złośliwe i łagodne mają różne charakterystyki wzrostu. Istotne cechy to m. in. promień i tekstura. Charakterystyki te wyznaczone są poprzez diagnostykę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

Próbka ze zbioru danych

	patient ID	Malignant/ Benign	radius (mean)	texture (mean)	perimeter (mean)	area (mean)
0	842302	M	17.99	10.38	122.80	1001.0
1	842517	M	20.57	17.77	132.90	1326.0
2	84300903	M	19.69	21.25	130.00	1203.0
3	84348301	M	11.42	20.38	77.58	386.1
4	84358402	M	20.29	14.34	135.10	1297.0

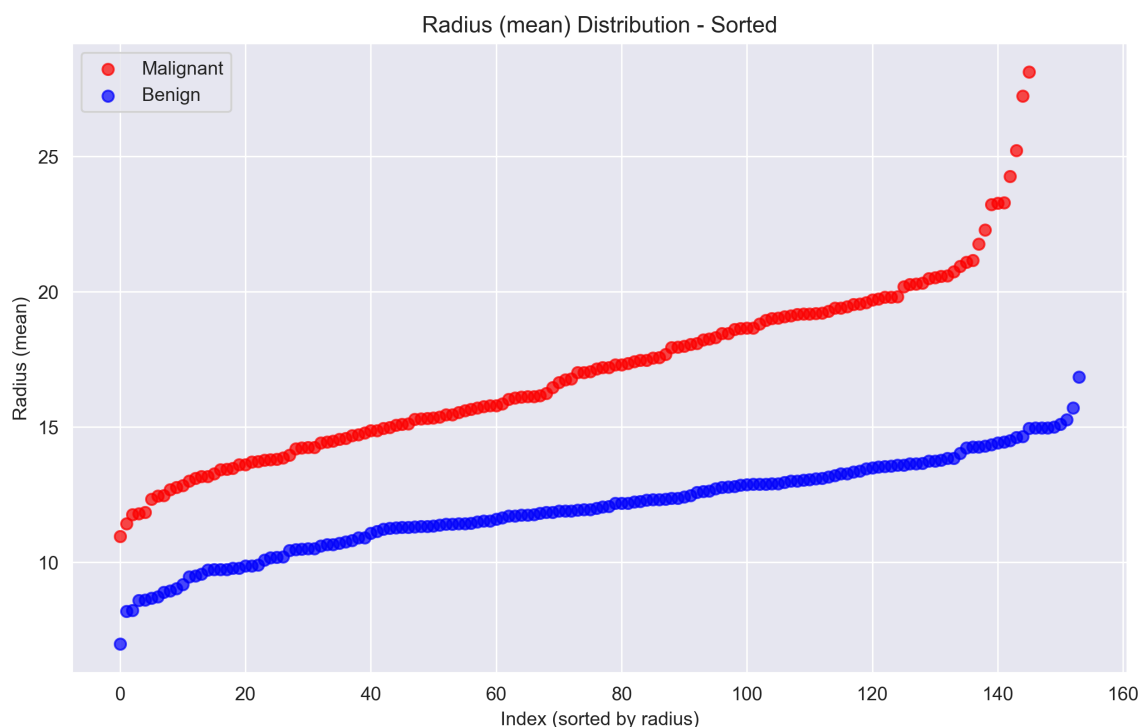
Ponizej znajduje się histogram i wykres posortowany rosnąco dla cechy radius (mean) dla próbek złośliwych i łagodnych

Możemy zauważyć, że próbki złośliwe posiadają większe odchylenie standardowe niż próbki łagodne. Wykresy przypominają wyglądem rozkład normalny.



Rysunek 1. Histogram dla cechy radius (mean) dla próbek złośliwych i łagodnych

Z wykresu posortowanego rosnąco wynika, że próbki złośliwe posiadają w większości przypadków wyższą wartość cechy radius (mean) niż próbki łagodne.



Rysunek 2. Wykres posortowany rosnąco dla cechy radius (mean) dla próbek złośliwych i łagodnych

Stworzyliśmy reprezentację danych zawartych w obu zbiorach dla liniowej i kwadratowej metody najmniejszych kwadratów. (Łącznie 4 macierze)

Stworzyliśmy wektor b , dla obu zbiorów danych, który zawiera wartości 1 dla próbek złośliwych i -1 dla próbek łagodnych.

Tworzenie wektora wag

Do stworzenia wektora wag wykorzystaliśmy wzór:

$$A^T A w = A^T y$$

do wyliczenia wagi w wykorzystaliśmy funkcję `np.linalg.solve` z biblioteki `numpy`.

Wykorzystanie rozkładu SVD

Do alternatywnego wyznaczenia wektora wag wykorzystaliśmy rozkład SVD o wartości λ równej 0.01

Współczynniki uwarunkowania

Do wyliczenia współczynnika uwarunkowania wykorzystaliśmy funkcję `np.linalg.cond` z biblioteki `numpy`. Dla poszczególnych metod otrzymaliśmy następujące wyniki:

- Współczynnik uwarunkowania liniowy wynosi - $1.8092 \cdot 10^{12}$
- Współczynnik uwarunkowania kwadratowy wynosi - $9.0568 \cdot 10^{17}$

Wartości współczynnika uwarunkowania dla obu zbiorów są bardzo duże, co oznacza, że te macierze są źle uwarunkowane.

Predykcja

Poniżej znajduje się tabela z wynikami predykcji dla obu zbiorów danych.

Metoda	TP	TN	FP	FN	Accuracy
Liniowa	58	194	6	2	96.92%
Kwadratowa	55	185	15	5	92.30%

Gdzie:

- TP - True Positive (prawdziwie dodatnie)
- TN - True Negative (prawdziwie ujemne)
- FP - False Positive (fałszywie dodatnie)
- FN - False Negative (fałszywie ujemne)
- $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

Wnioski

- Współczynnik uwarunkowania wskazuje, jak uwarunkowana jest macierz. Im wyższy współczynnik tym gorzej uwarunkowana macierz i tym mniej stabilny numerycznie jest model. Obliczone współczynniki wskazują, że reprezentacja liniowa jest lepiej uwarunkowana i stabilniejsza numerycznie niż reprezentacja kwadratowa.
- W powyższej tabeli przedstawione są wyniki przewidywań na podstawie wag uzyskanych różnymi metodami. Wynika z nich, że w tym przypadku (przewidywanie złośliwości nowotworu dla określonych parametrów) bardziej skuteczny jest model oparty o reprezentację liniową. Wyniki modelu opartego o reprezentację liniową nie zależą od sposobu uzyskania wag, we wszystkich trzech przypadkach ilość pomyłek jest identyczna.