

Lab02 Metoda najmniejszych kwadratów

Patryk Blacha, Radosław Szepielak

14 marca 2025

Spis treści

1	Wprowadzenie teoretyczne	1
1.1	Metoda najmniejszych kwadratów w klasyfikacji	1
2	Metodologia	1
2.1	Zbiór danych i przygotowanie	1
2.2	Proces przetwarzania	2
3	Analiza numeryczna	2
3.1	Stabilność obliczeniowa	2
4	Wyniki eksperymentalne	3
4.1	Porównanie skuteczności	3
4.2	Szczegółowa analiza błędów	3
5	Dyskusja	3
5.1	Interpretacja wyników	3
5.2	Ograniczenia metody	3

1 Wprowadzenie teoretyczne

1.1 Metoda najmniejszych kwadratów w klasyfikacji

Metoda najmniejszych kwadratów (MNK) została zastosowana do problemu klasyfikacji binarnej poprzez odpowiednie przekształcenie etykiet. Dla macierzy cech \mathbf{A} i wektora \mathbf{b} rozwiązanie wyraża się wzorem:

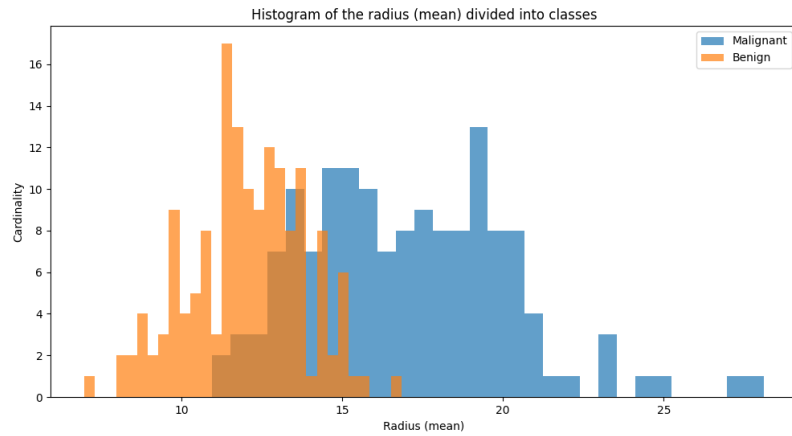
$$\mathbf{w} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}$$

gdzie \mathbf{w} zawiera optymalne wagi modelu.

2 Metodologia

2.1 Zbiór danych i przygotowanie

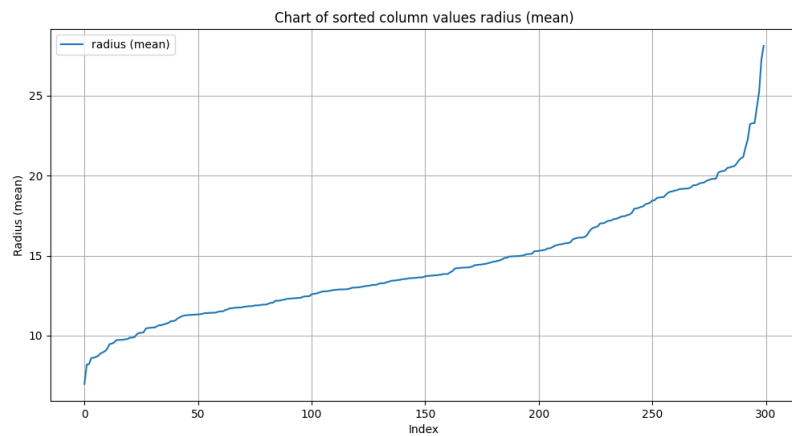
- Zbiór Breast Cancer Wisconsin zawierający 30 cech diagnostycznych
- Podział na treningowy (300 próbek) i walidacyjny (260 próbek)
- Kodowanie etykiet: 1 (złośliwy), -1 (łagodny)



Rysunek 1: Rozkład wartości cechy 'radius (mean)' z podziałem na klasy

2.2 Proces przetwarzania

1. Normalizacja brakująca (dane już wystandaryzowane)
2. Generowanie cech kwadratowych dla wybranych atrybutów
3. Budowa macierzy projektowych dla modeli liniowych i nieliniowych



Rysunek 2: Posortowane wartości cechy 'radius (mean)'

3 Analiza numeryczna

3.1 Stabilność obliczeniowa

- Współczynnik uwarunkowania dla modelu liniowego: 2.10×10^{12}
- Współczynnik uwarunkowania dla modelu kwadratowego: 1.07×10^{18}

4 Wyniki eksperymentalne

4.1 Porównanie skuteczności

Model	Dokładność
Liniowy (SVD)	97.31%
Kwadratowy	91.92%

Tabela 1: Porównanie efektywności modeli

4.2 Szczegółowa analiza błędów

- Model liniowy: 5 błędów I rodzaju (FP), 2 błędy II rodzaju (FN)
- Model kwadratowy: 16 błędów I rodzaju, 5 błędów II rodzaju
- Główna przyczyna błędów: nakładanie się rozkładów cech dla klas

5 Dyskusja

5.1 Interpretacja wyników

- Wysoka dokładność modelu liniowego sugeruje liniową separowalność danych
- Spadek dokładności dla modelu kwadratowego wskazuje na:
 - * Przeuczenie na szumach
 - * Problemy numeryczne związane z wysokim współczynnikiem uwarunkowania

5.2 Ograniczenia metody

- Wrażliwość na współliniowość cech
- Brak automatycznej selekcji cech
- Problemy z interpretacją wag przy wysokiej korelacji zmiennych

Pełne parametry modeli

```
1 Wagi modelu liniowego (pierwsze 5):  
2 - Bias: -6.172  
3 - radius_mean: -0.608  
4 - texture_mean: 0.025  
5 - perimeter_mean: 0.078  
6 - area_mean: 0.00058
```

Literatura

- Materiały pomocnicze do laboratorium zamieszczone na platformie Teams (lab02/lab2-intro.pdf)