Marek Małek, Marcin Serafin 14.03.2024

Laboratorium 02

Metoda najmniejszych kwadratów

# 1 Treść zadania

Celem zadania było zastosowanie metody najmniejszych kwadratów do predykcji, czy nowotwór jest złośliwy czy łagodny. Do rozwiązania należało wykorzystać bibliotekę **pandas** oraz typ **DataFrame**. Dostarczone zostały dwa zbiory danych:

- **breast-cancer-train.dat**
- **breast-cancer-validate.dat**

Oraz plik **breast-cancer.labels**, w którym zostały zawarte nazwy kolumn.

# 2 Rozwiązanie zadania

## 2.1 Wczytanie danych

W celu wczytania danych wykonano następujący fragment kodu:

```
with open("data\\breast-cancer.labels") as f:
    column_names = [line[:(len(line))-1] for line in f.readlines()]

breast_cancer_train = pd.io.parsers.read_csv("data\\breast-cancer-train.dat")
breast_cancer_train.columns = column_names

breast_cancer_validate =  pd.io.parsers.read_csv("data\\breast-cancer-validate.dat")
breast_cancer_validate.columns = column_names
```

Dane zostały wczytane do odpowiednich zmiennych oraz przypisane zostały nazwy charakterystyk do kolumn **DataFramea**

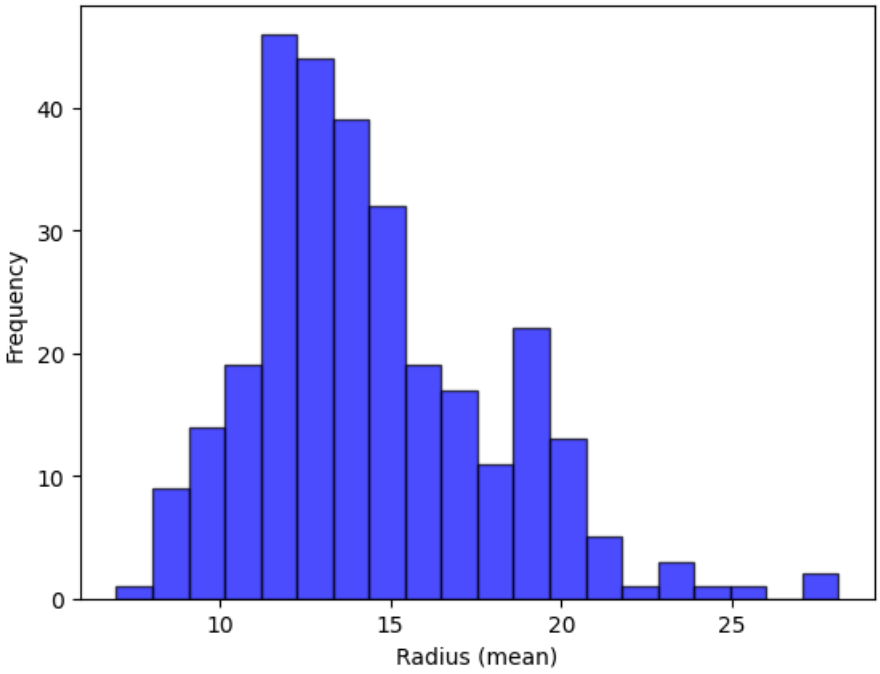
Widok pierwszych pięciu wierszy i pierwszych pięciu kolumn **DataFramea** zbioru danych **breast_cancer_train**:

| | patient ID | Malignant/Benign | radius (mean) | texture (mean) | perimeter (mean) |
|---|---|---|---|---|---|
| 0 | 842517 | M | 20.570000 | 17.770000 | 132.900000 |
| 1 | 84300903 | M | 19.690000 | 21.250000 | 130.000000 |
| 2 | 84348301 | M | 11.420000 | 20.380000 | 77.580000 |
| 3 | 84358402 | M | 20.290000 | 14.340000 | 135.100000 |
| 4 | 843786 | M | 12.450000 | 15.700000 | 82.570000 |

## 2.2 Wizualizacja pojedyńczej charakterystki

### 2.2.1 Histogram

Stworzono histogram klasyfikujący wśród ilu pacjentów wykryto nowotwór z danym średnim promieniem.

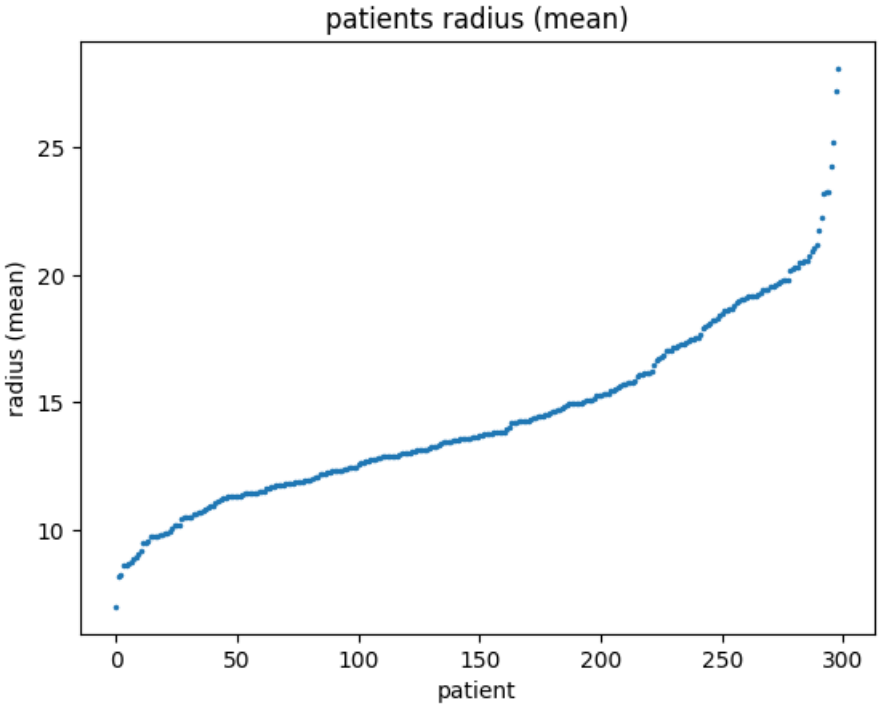



**Wizualizacja** 1: Historgram charakterystki radius (mean)

Na podstawie histogramu widać, że najwięcej pacjentów zostało zaklasyfikowanych do przedziału, gdzie wartość średnia promienia nowotrowru była między 10, a 15.

### 2.2.2 Wykres

Analogicznie do poprzedniego podpunktu stworzono wykres tej samej charakterystki.




**Wizualizacja** 2: Wykres charakterystki radius (mean)

Na podstawie wykresu można zauważyć, że najbardziej rzetelne dane znajdują się na przedziale 10-20. Wartości poza tym zakresem są o wiele rzadsze i można je traktować jako szum.

## 2.3 Reprezentacje danych

Stworzono reprezentacje danych zawartych w zbiorach w oparciu o wzory:

Reprezentacja liniowa:

$$A_{\text{lin}} = \begin{bmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,m} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,m} \\ f_{3,1} & f_{3,2} & \cdots & f_{3,m} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n,1} & f_{n,2} & \cdots & f_{n,m} \end{bmatrix} \tag{1}$$

Reprezentacja kwadratowa (wybrano 4 parametry: **radius (mean)**, **perimeter (mean)**, **area (mean)**, **symmetry (mean)**):

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

## 2.4 Wektor $b$ oraz wagi reprezentacji

W celu znalezienia wag dla liniowych i kwadratowych reprezentacji użyto równania:

$$\boldsymbol{A^T A w = A^T b} \tag{3}$$

gdzie wektor $b$ jest zadany jako:

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} \text{ gdzie } \alpha_i = \begin{cases} 1 & \text{jeśli nowotwór jest złośliwy} \\ -1 & \text{wpp.} \end{cases} \tag{4}$$

Przy tworzeniu wektora $b$ użyto funkcji **np.where**

```
b_train = np.where(breast_cancer_train['Malignant/Benign'] == 'M', 1, -1)
b_validate = np.where(breast_cancer_validate['Malignant/Benign'] == 'M', 1, -1)
```

## 2.5 Współczynnik uwarunkowania macierzy

Przy obliczeniu współczynnika uwarunkowania macierzy $\text{cond}(A)$ oraz $\text{cond}(A^T A)$ wykorzystano równania:

$$\text{cond}(A) = ||A|| \cdot ||A^T|| \tag{5}$$

$$\text{cond}(A^T A) = \text{cond}(A)^2 \tag{6}$$

**DataFrame** reprezentujący obliczone wartości:

| | $\text{cond}(A^T A)$ | $\text{cond}(A)$ |
|---|---|---|
| reprezentacja liniowa | $1.15 \cdot 10^{22}$ | $3.39 \cdot 10^{18}$ |
| reprezentacja najmniejszych kwadratów | $9.02 \cdot 10^{17}$ | $2.56 \cdot 10^{14}$ |

**Tabela** 1: DataFrame współczynników uwarunkowania macierzy

## 2.6 Przewidywanie nowotworu

W celu określenia czy dany nowotwór jest złośliwy czy nie pomnożono reprezentacje liniową oraz kwadratową przez uprzednio wyliczone odpowienie wektory wag. Dla otrzymanego wektora $p$ zliczono liczbę wartości $p_i$ takich, że $p[i] \leq 0$ (wtedy nowotwór prawdopodobnie był łagodny) oraz $p[i] > 0$ (wtedy prawdopodobnie nowotwór był złośliwy). Obliczono też liczbę przewidywań fałszywie dodatnich (przewidziano, że nowotwór był złośliwy, a tak naprawdę był łagodny) oraz fałszywie ujemnych (analogicznie do poprzedniego przykładu). Otrzymane wyniki zestawiono w **DataFramie**

| | Liczba fałszywie ujemnych | Liczba fałszywie dodatnich | Liczba przewidzianych nowotworów złośliwych | Liczba przewidzianych nowotworów łagodnych | Dokładność |
|---|---|---|---|---|---|
| Reprezentacja liniowa | 2 | 6 | 63 | 196 | 96.91% |
| Reprezentacja najmnieszych kwadratów | 5 | 15 | 69 | 190 | 92.27% |

**Tabela** 2: DataFrame klasyfikacji fałszywie ujemnych i fałszywie dodatnich

# 3 Wnioski

Na podsatwie **Tabeli 2** w punktcie **2.6** można stwierdzić, że reprezentacja najmniejszych kwadratów osiągneła gorszy rezulatat, jako że liczba przypadków fałszywie dodatnich oraz fałszywie ujemnych była większa od wyniki reprezentacji liniowej. Podobnie było z dokładnością która była większa dla reprezentacji linowej. Powodem była wąska selekcja parametrów. Jak podano w punkcie **2.3**, reprezentacja liniowa przyjęła wszystkie charakterystkich nowotworu, a reprezentacja kwadratowa tylko wybrane cechy. Jednakże, w punkcie **2.5** współczynnik uwarunkowania macierzy był większy w reprezentacji liniowej. W tym przypadku reprezentacja liniowa jest mniej wrażliwa na błędy ze względu na lepsze dopasowanie, stąd też niższy wspołczynnik uwarnkowania.

# 4 Bibliografia


1. http://heath.cs.illinois.edu/scicomp/notes/cs450_chapt03.pdf
2. https://pythonnumericalmethods.berkeley.edu/notebooks/chapter16.00-Least-Squares-Regression.html