

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:

```
1 column_names = []
2
3 with open("data\\breast-cancer.labels") as f:
4     for line in f.readlines():
5         column_names.append(line[:len(line)-1])
6
7 breast_cancer_train = pd.io.parsers.read_csv("data\\breast-cancer-train.dat")
8 breast_cancer_train.columns = column_names
9
10 breast_cancer_validate = pd.io.parsers.read_csv("data\\breast-cancer-validate.dat")
11 breast_cancer_validate.columns = column_names
```

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

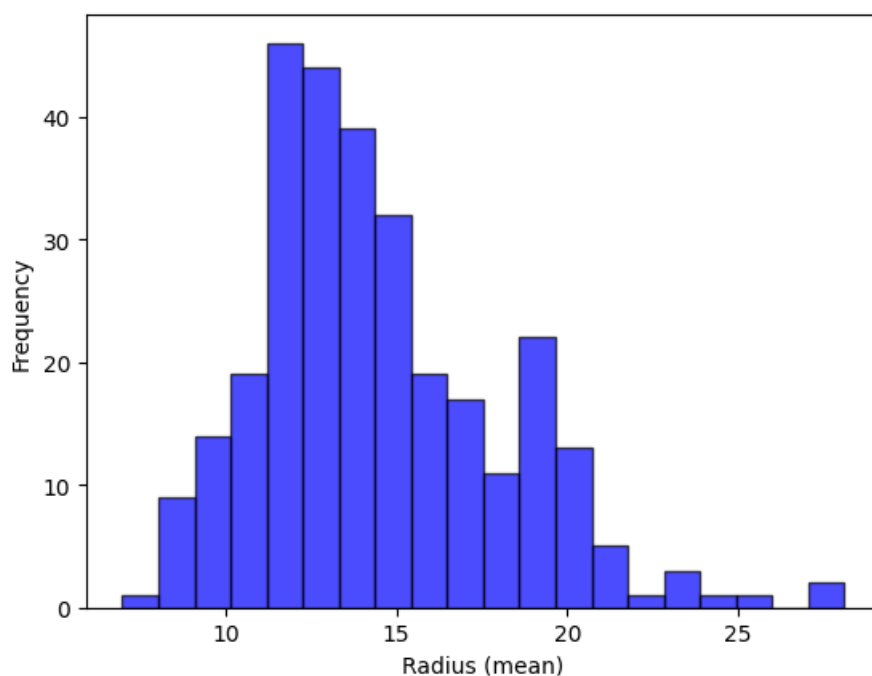
Widok pierwszych pięciu wierszy i pierwszych pięciu kolumn **DataFramea** zbioru danych **breast\_cancer\_train**:

	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 pojedynczej charakterystyki

### 2.2.1 Histogram

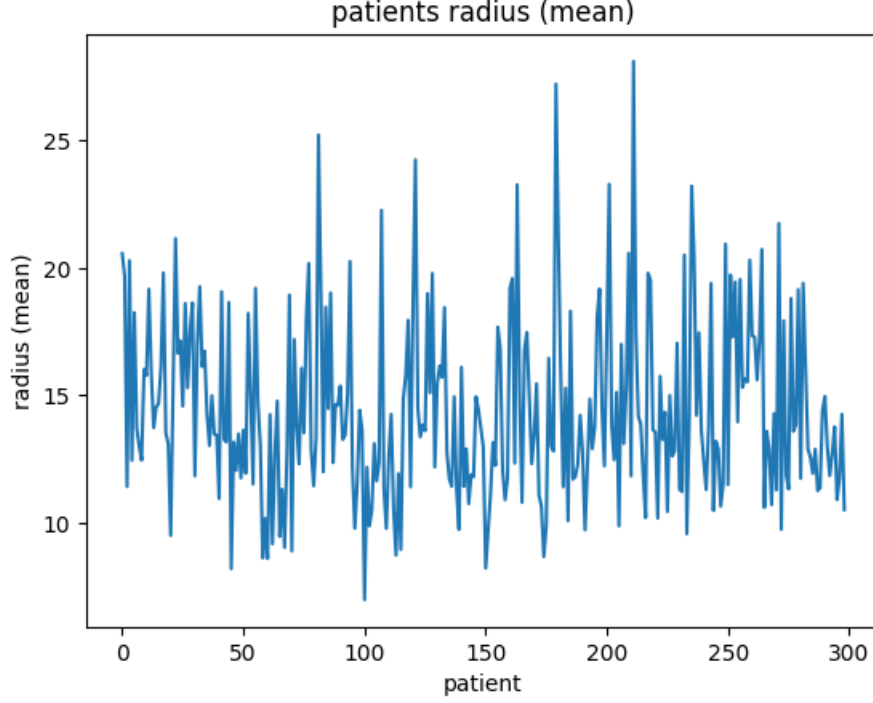
Stworzono histogram klasyfikujący wśród ilu pacjentów wykryto nowotwór z danym średnim promieniem.



**Wizualizacja 1:** Histogram charakterystyki radius (mean)

### 2.2.2 Wykres

Analogicznie do poprzedniego podpunktu stworzono wykres tej samej charakterystyki.



Wizualizacja 2: Wykres charakterystyki radius (mean)

## 2.3 Reprezentacje danych

Stworzono reprezentacje danych zawartych w zbiorach w oparciu o wzory:

Reprezentacja liniowa:

$$A_{\text{lin}} = \begin{bmatrix} f_{1,1} & f_{1,2} & \dots & f_{1,m} \\ f_{2,1} & f_{2,2} & \dots & f_{2,m} \\ f_{3,1} & f_{3,2} & \dots & f_{3,m} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n,1} & f_{n,2} & & f_{n,m} \end{bmatrix} \quad (1)$$

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

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

## 2.4 Wektor $b$ oraz wagi reprezentacji

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

$$A^T A w = A^T b \quad (3)$$

gdzie wektor  $b$  jest zadany jako:

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

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

```
1 b_train = np.where(breast_cancer_train['Malignant/Benign'] == 'M', 1, -1)
2 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:

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

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

**DataFrame** reprezentujący obliczone wartości:

	$\text{cond}(A)$	$\text{cond}(A^T A)$
reprezentacja liniowa	3392717729702346240.000000	11518506744199914717184.000000
reprezentacja najmniejszych kwadratów	256020503885671.218750	902853865047191680.000000

**Tabela 1:** DataFrame współczynników uwarunkowania macierzy

### 2.5.1 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 **DataFrame**

	Liczba fałszywie ujemnych	Liczba fałszywie dodatnich	Liczba przewidzianych nowotworów złośliwych	Liczba przewidzianych nowotworów łagodnych
Reprezentacja liniowa	2	6	63	196
Reprezentacja najmniejszych kwadratów	5	15	69	190

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

### 3 Wnioski

### 4 Bibliografia

1. [http://heath.cs.illinois.edu/scicomp/notes/cs450\\_chapt03.pdf](http://heath.cs.illinois.edu/scicomp/notes/cs450_chapt03.pdf)
2. <https://pythonnumericalmethods.berkeley.edu/notebooks/chapter16.00-Least-Squares-Regression.html>