Laboratorium 2 – Metoda najmniejszych kwadratów

14.03.2023 **Adam Trybus, Piotr Olszak**

1. **Opis i rozwiązania zadań**
   a) Tak jak polecenie wskazywało, datasety pobieraliśmy za pomocą pd.read_csv z biblioteki pandas

```
# Wczytanie zbioru treningowego
train_data = pd.read_csv('breast-cancer-train.dat', header=None, delimiter=',')
train_labels = pd.read_csv('breast-cancer.labels', header=None)
```

   b) Stworzyliśmy dwa wykresy, które są dokładnie opisane w punkcie 2. Pierwszy wykres - promienia, od obwodu guza, jak i drugi – promienia guza od ilości wystąpień zrobiliśmy za pomocą biblioteki matplotlib.pyplot.

```
# podpis osi i tytuł histogramu
plt.hist(validate_data.iloc[:,2], bins=15)
plt.xlabel('Radius')
plt.ylabel('Count')
plt.title('Histogram of Radius')

# podpis osi i tytuł wykresu
plt.figure()
plt.plot(validate_data.iloc[:,2], validate_data.iloc[:,4], 'k.')
plt.xlabel('Radius')
plt.ylabel('Perimeter')
plt.title('Scatter Plot of Radius vs Perimeter')

# wyświetlenie wykresów
plt.show()
```

   c) Reprezentacja liniową, jak i kwadratową tworzymy przy pomocy funkcji iloc (z biblioteki pandas), która zwraca nam wartości wszystkich wierszy i wybranych przez nas kolumn

```
#tworzenie reprezentacji liniowej
X_train_linear = train_data.iloc[:, 2:].values
X_validate_linear = validate_data.iloc[:, 2:].values

#tworzenie reprezentacji kwadratowej
X_train_quadratic = train_data.iloc[:, [2, 4, 5, 10]].values
X_validate_quadratic = validate_data.iloc[:, [2, 4, 5, 10]].values
```

Dla reprezentacji kwadratowej musieliśmy dodatkowo użyć wzoru z materiałów załączonych do laboratorium.

```
for row in X_train_quadratic:
    quad_train.append([i for i in row] + [i*i for i in row] + [row[0]*row[1],row[0]*row[2],row[0]*row[3],row[1]*row[2],row[1]*row[3],row[2]*row[3]])

for row in X_validate_quadratic:
    quad_train.append([i for i in row] + [i*i for i in row] + [row[0]*row[1],row[0]*row[2],row[0]*row[3],row[1]*row[2],row[1]*row[3],row[2]*row[3]])
```

d) Wektor b został stworzony przy pomocy biblioteki numpy i funkcji where

```
#Aby znaleźć wagi dla liniowej reprezentacji najmniejszych kwadratów, można skorzystać z równania normalnego, które ma postać:

b_vec_train = np.where(train_data.iloc[:, 1] == "M", 1, -1)

b_vec_val = np.where(validate_data.iloc[:, 1] == "M", 1, -1)
```

e) Wagi dla liniowej oraz kwadratowej reprezentacji najmniejszych kwadratów obliczyliśmy wg wzoru z materiałów załączonych do laboratorium

```
# obliczenie wag
A = X_train_linear
b = b_vec_train
w_linear = np.linalg.inv(A.T @ A) @ A.T @ b


# obliczenie wag
A2 = X_train_quadratic
b = b_vec_train
w_quad = np.linalg.inv(A2.T @ A2) @ A2.T @ b
```

f) Współczynniki uwarunkowania cond liczyliśmy przy pomocy biblioteki numpy

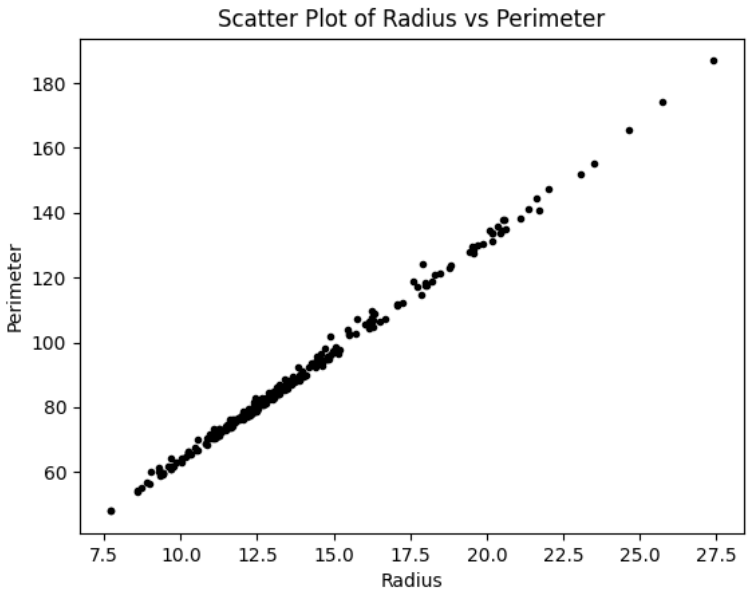
```
#wspolczynniki cond powinny byc bliskie 1
cond = np.linalg.cond(A)
cond2 = np.linalg.cond(A2)
```

g) Predykcje typu nowotworu robimy korzystając z funkcji ‘numpy.sign’ , która zamienia nam wymnożoną macierz na tablicę 1 oraz -1. Następnie porównujemy ile jest ‘false positive’ i ‘false negative’ dla obu reprezentacji przy pomocy funkcji ‘numpy.sum’

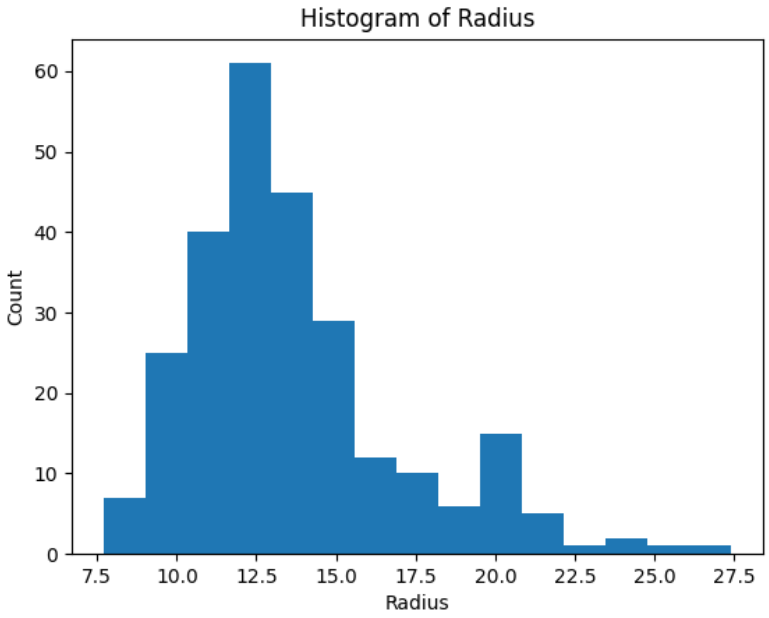
```
y_pred_linear = np.sign(X_validate_linear @ w_linear)
y_pred_quadratic = np.sign(X_validate_quadratic @ w_quad)
y_true = np.where(validate_data.iloc[:, 1] == "M", 1, -1)

fp_linear = np.sum((y_pred_linear == 1) & (y_true == -1))
fn_linear = np.sum((y_pred_linear == -1) & (y_true == 1))

# dla reprezentacji kwadratowej
fp_quadratic = np.sum((y_pred_quadratic == 1) & (y_true == -1))
fn_quadratic = np.sum((y_pred_quadratic == -1) & (y_true == 1))
```

2. Wykresy i wyniki liczbowe

1) Jak widać na poniższym wykresie istnieje silna, pozytywna korelacja między promieniem (radius), a obwodem guza (perimeter) – im większa wartość jednego, tym też rośnie drugie. Obie te cechy pomagają określić rozmiar guza i mogą być przydatne w diagnozowaniu, czy guz jest łagodny czy złośliwy.



2) Wykres promienia guza (radius) od ilości wystąpień. Z wykresu możemy się dowiedzieć, że najwięcej przypadków jest między 10 – 15 cm.



3) Cond (reprezentacja liniowa) = 1345082.9797889264
Cond (reprezentacja kwadratowa) = 28788.960992359596

Fałszywie dodatnie i fałszywie ujemne dla liniowej reprezentacji:
Fałszywie dodatnie: 6
Fałszywie ujemne: 2

Fałszywie dodatnie i fałszywie ujemne dla kwadratowej reprezentacji:
Fałszywie dodatnie: 8
Fałszywie ujemne: 7

3. **Wnioski**

Jesteśmy w stanie z dużą dozą prawdopodobieństwa przewidzieć czy dany guz jest złośliwy czy łagodny.

4. **Bibliografia**

https://jakevdp.github.io/PythonDataScienceHandbook/03.00-introduction-to-pandas.html