Artur Gęsiarz

Marzec 16, 2024

Laboratorium nr 2

MOwNiT – Metoda najmniejszych kwadratów

1. Treść zadania
   1.1. Celem zadania jest zastosowanie metody najmniejszych kwadratów do predykcji, czy nowotwór jest złośliwy czy łagodny. Nowotwory złośliwe i łagodne mają różne charakterystyki wzrostu. Istotne cechy to m. in. promień i tekstura. Charakterystyki te wyznaczane są poprzez diagnostykę obrazową i biopsje.

   Do rozwiązania problemu wykorzystamy bibliotekę ***pandas,*** typ ***DataFrame*** oraz dwa zbiory danych:

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

   Nazwy kolumn znajdują się w pliku ***breast-cancer.labels***. Pierwsza kolumna to identyfikator pacjenta ***patient ID***. Dla każdego pacjenta wartość w kolumnie ***Malignant/Benign*** wskazuje klasę, tj. czy jego nowotwór jest złośliwy czy łagodny. Pozostałe 30 kolumn zawiera cechy, tj. charakterystyki nowotworu.

2. Rozwiązanie zadania
   2.1. Importowanie danych do testowania:

```
column_names = []
with open("../materialy/lab02/dataset/breast-cancer.labels") as f:
    for line in f:
        column_names.append(line.strip())
```

```
train_data = pd.io.parsers.read_csv("../materialy/lab02/dataset/breast-cancer-train.dat",
                                    names=column_names)
```

```
validate_data = pd.io.parsers.read_csv("../materialy/lab02/dataset/breast-cancer-validate.dat",
                                       names=column_names)
```

   2.2. Implementacja rysowania histogramu charakterysytki promienia

```
plt.figure(figsize=(10, 6))
plt.hist(validate_data['radius (mean)'], bins=20, color='skyblue', edgecolor='black')
plt.xlabel('Radius (mean)')
plt.ylabel('Frequency')
plt.xticks(range(0, 30, 2))
plt.yticks(range(0, 60, 10))
plt.title('Histogram of Radius (mean)')
plt.show()
```

2.3. Implementacja wykresu charakterystki promienia

```
plt.figure(figsize=(10, 6))
plt.plot(validate_data['radius (mean)'], color='skyblue')
plt.xlabel('Sample')
plt.ylabel('Radius (mean)')
plt.title('Line chart of Radius (mean)')
plt.show()
```

2.4. Implementacja reprezentacji danych zawartych w obu zbiorach dla liniowej i kwadratowej metody najmniejszych kwadratów

```
linear_train = train_data.drop(columns=["Malignant/Benign"]).values
linear_validate = validate_data.drop(columns=["Malignant/Benign"]).values

quadratic_train = train_data[["radius (mean)", "perimeter (mean)", "area (mean)", "symmetry (mean)"]].values
quadratic_validate = validate_data[["radius (mean)", "perimeter (mean)", "area (mean)", "symmetry (mean)"]]
 .values
```

2.5. Implementacja wektora b dla obu zbiorów

```
b_train = np.where(train_data["Malignant/Benign"] == 'M', 1, -1)
b_validate = np.where(validate_data["Malignant/Benign"] == 'M', 1, -1)
```

2.6. Implementacja wagi dla liniowej oraz kwadratowej metody najmniejszych kwadratow

```
linear_weights = np.linalg.solve(linear_train.T @ linear_train, linear_train.T @ b_train)
quadratic_weights = np.linalg.solve(quadratic_train.T @ quadratic_train, quadratic_train.T @ b_train)
```

2.7. Implementacja obliczenia współczynnika uwarunkowania macierzy $cond(A^T A)$

```
linear_condition_number = np.linalg.cond(linear_train.T @ linear_train)
quadratic_condition_number = np.linalg.cond(quadratic_train.T @ quadratic_train)

print(linear_condition_number, quadratic_condition_number)
```

2.8. Sprawdzenie jak dobrze otrzymane wagi przewidują typ nowotworu

```
linear_predictions = linear_validate @ linear_weights
quadratic_predictions = quadratic_validate @ quadratic_weights

linear_predictions = np.where(linear_predictions > 0, 1, -1)
quadratic_predictions = np.where(quadratic_predictions > 0, 1, -1)

linear_false_positives = np.sum((linear_predictions == 1) & (b_validate == -1))
linear_false_negatives = np.sum((linear_predictions == -1) & (b_validate == 1))

quadratic_false_positives = np.sum((quadratic_predictions == 1) & (b_validate == -1))
quadratic_false_negatives = np.sum((quadratic_predictions == -1) & (b_validate == 1))
```

2.9. Implementacja wykresu porównującego fałszywie dodatnie oraz fałszywie ujemne wyniki testu na nowotwór

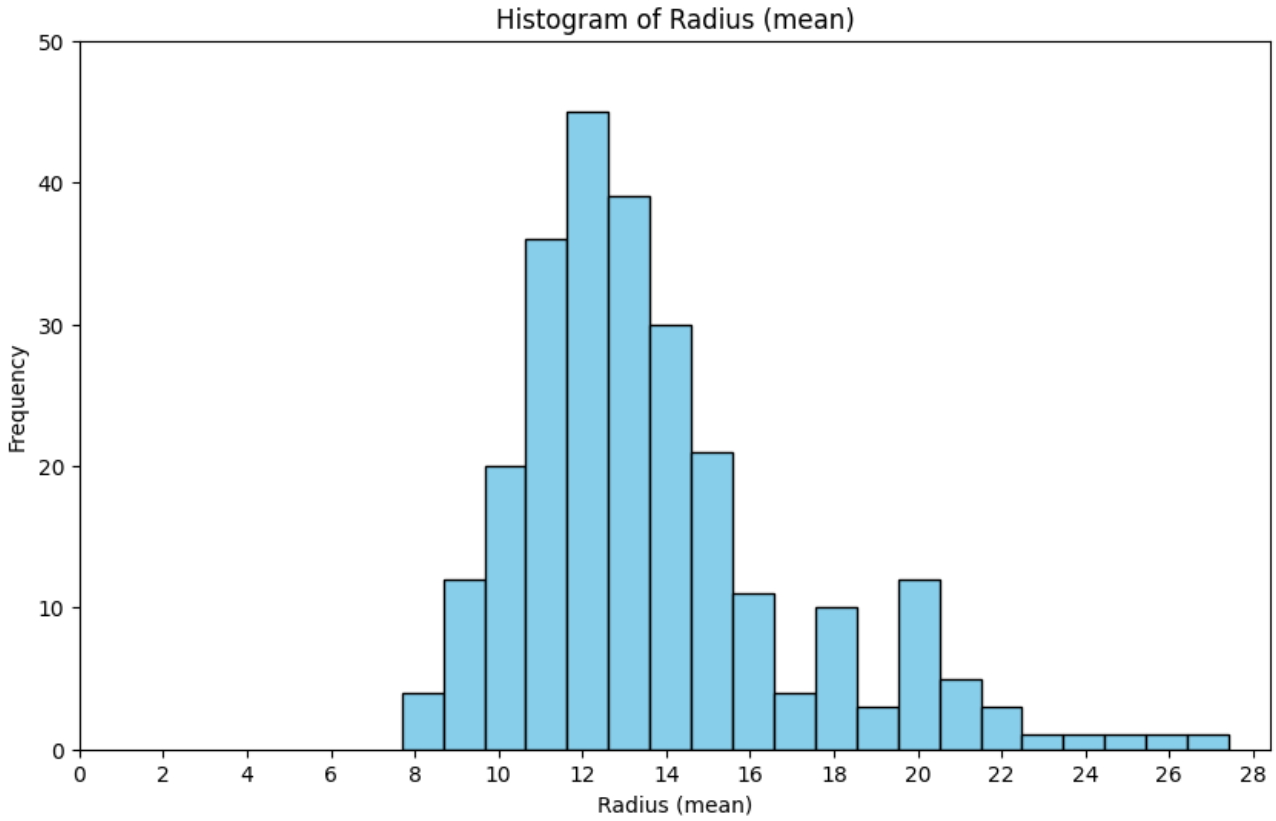
```
methods = ['Linear', 'Quadratic']
false_positives = [linear_false_positives, quadratic_false_positives]
false_negatives = [linear_false_negatives, quadratic_false_negatives]

plt.figure(figsize=(8, 6))
plt.bar(methods, false_positives, color='skyblue', label='False Positives')
plt.bar(methods, false_negatives, color='lightgreen', label='False Negatives', bottom=false_positives)

plt.xlabel('Method')
plt.ylabel('Number of cases')
plt.title('Comparison of false positive and false negative cases')
plt.legend()
plt.show()
```

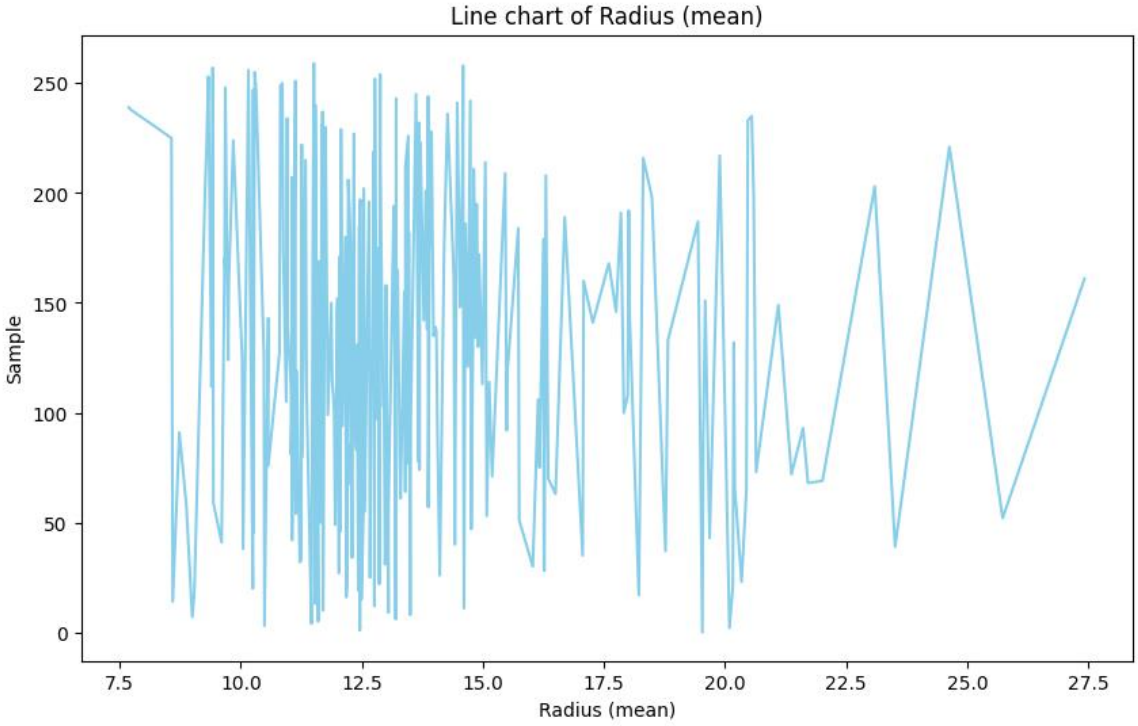
3. Wykresy

3.1. Histogram charakterystki promienia



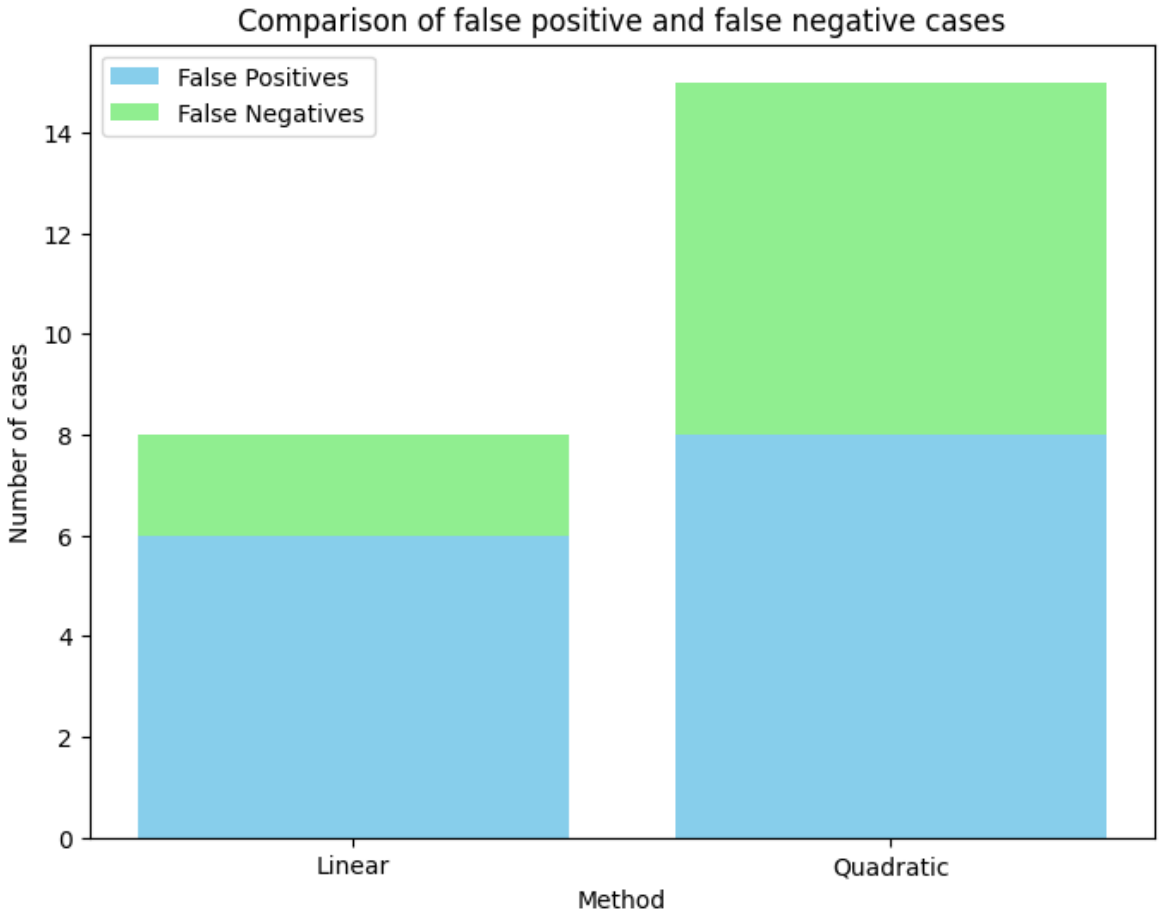
Wykres 1. Histogram opisujący jaki procent badanych ma jaki promień

3.2. Wykres charakterystki promienia



Wykres 2. Wykres opisujący badanych a ich promień choroby

3.3. Wykres porównujący fałszywie dodatnie testy oraz fałszywie ujemne otrzymane poprzez dwie różne metody



Wykres 3. Wykres porównujący metody oraz fałszywie dodatnie/ujemne wyniki testów

4. Tabele

4.1. Wyniki fałszywych dodanie oraz fałszywych ujemnie testów w poszczególnych metodach

Metoda	Fałszywie ujemne testy	Fałszywie dodanie testy
Liniowa	2	6
Kwadratowa	7	8

Tabela 1. Porównanie wyników testów

4.2. Tabela współczynników uwarunkowania macierzy

Metoda	Współczynnik uwarunkowania macierzy
Liniowa	$1{,}14 * 10^{22}$
Kwadratowa	$8{,}29 * 10^{8}$

Tabela 2. Porównanie współczynników uwarunkowania macierzy

5. Wnioski

Analizując wyniki, można zauważyć, że zarówno liniowa, jak i kwadratowa metoda najmniejszych kwadratów miały podobne wyniki w przypadku fałszywie dodanich, ale miały różne wyniki w przypadku fałszywie ujemnych.

Mimo wszystko lepiej poradziła sobie metoda liniowa poniewaz ma mniej przypadkow falszywie ujemnych od metody kwadratowej.
Jednak porównując współczynniki uwarunkowania macierzy znacznie większy ma metoda liniowa od kwadratowej.

Podsumowując, różnice w wynikach między liniową a kwadratową metodą najmniejszych kwadratów mogą wynikać z różnych czynników, takich jak złożoność modelu, rozmiar zestawu treningowego, wybór cech i zakłócenia w danych.

6. Bibliografia

Wykład MOwNiT - prowadzony przez dr. Inż. K. Rycerz
Prezentacje – dr. Inż. M. Kuta

7. Dodatkowe informacje

Rozwiązania zadania znajduje się w pliku ex1.ipynb