Praca domowa 5
Miary odległości

Damian Jankowski s188597

20 maja 2023

1 Wstęp

Celem pracy domowej było zapoznanie się z wybranymi miarami odległości. Miary odległości odgrywają ważną rolę w uczeniu maszynowym. Są one wykorzystywane do obliczania odległości między obserwacjami w celu klasyfikacji nowych danych. W zależności od wybranej miary, wyniki mogą się znacznie różnić. Dlatego też ważne jest, aby wybrać miarę odpowiednią do danego problemu.

Dla przykładowych danych należało obliczyć odległości następującymi miarami:

- miejska (Manhattan)
- euklidesowa
- Czebyszewa

Każda z tych miar jest rozszerzeniem tzw. odległości Minkowskiego, która jest zdefiniowana następująco:

$$L_m(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^m\right)^{\frac{1}{m}} \tag{1}$$

gdzie:

- x, y - wektory danych
- n - liczba wymiarów
- m - parametr miary

1.1 Odległość miejska (Manhattan)

Szczególny przypadek odległości Minkowskiego, dla $m = 1$, wyrażona wzorem:

$$L_1(x,y) = \sum_{i=1}^{n} |x_i - y_i| \tag{2}$$

Daje podobne wyniki jak odległość euklidesowa, ale odstające wartości są stłumione przez brak podnoszenia ich do kwadratu.

1.2 Odległość euklidesowa

Jest to odległość Minkowskiego dla $m = 2$, wyrażona wzorem:

$$L_2(x,y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \tag{3}$$

Najczęściej wykorzystywana miara odległości. Wadą tej miary jest duża wrażliwość na wartości odstające z racji kwadratowego wykładnika.

1.3 Odległość Czebyszewa

Wyrażona wzorem:

$$L_\infty(x, y) = \max_{i=1}^{n} |x_i - y_i| = \lim_{m\to\infty} L_m(x, y) \tag{4}$$

Wprowadzona przez Pafnutija Czebyszewa. Jest to szczególny przypadek odległości Minkowskiego, dla $m = \infty$. Wartość odległości jest równa największej różnicy między wartościami wektorów.

W szachach odległość Czebyszewa jest równa liczbie ruchów króla, potrzebnych do przejścia z jednego pola na drugie.

2 Zadanie

Do przećwiczenia wyznaczania tych miar wybrałem przykładowy zbiór danych:

$$X = \begin{bmatrix} -3.5 \\ 4.4 \\ -2.2 \end{bmatrix} \quad Y = \begin{bmatrix} 0.1 \\ -2.5 \\ 3.9 \end{bmatrix} \tag{5}$$

Obliczenia prezentuję poniżej:

$$X = \begin{bmatrix} -3{,}5 \\ 4{,}4 \\ -2{,}2 \end{bmatrix} \quad Y = \begin{bmatrix} 0{,}1 \\ -2{,}5 \\ 3{,}9 \end{bmatrix} \quad L_m = \left(\sum_{i=1}^{n} |x_i - y_i|^m \right)^{\frac{1}{m}}$$

Odległość miejska (Manhattan):

$$L_1 = |-3{,}5 - 0{,}1| + |4{,}4 + 2{,}5| + |-2{,}2 - 3{,}9| =$$

$$= 3{,}6 + 6{,}9 + 6{,}1 = 16{,}6$$

Odległość Euklidesowa:

$$L_2 = \sqrt{(-3{,}5 - 0{,}1)^2 + (4{,}4 - (-2{,}5))^2 + (-2{,}2 - 3{,}9)^2} =$$

$$= \sqrt{12{,}96 + 47{,}61 + 37{,}21} \approx 9{,}89$$

Odległość Czebyszewa:

$$L_\infty = \max(3{,}6; 6{,}9; 6{,}1) = 6{,}9$$

By sprawdzić poprawność obliczeń napisałem krótki skrypt w języku Python:

```python
X = [ -3.5, 4.4, -2.2 ]
Y = [ 0.1, -2.5, 3.9 ]

def manhattan_distance(x, y):
    return sum(abs(a - b) for a, b in zip(x, y))

def euclidean_distance(x, y):
    return round(sum((a - b) ** 2 for a, b in zip(x, y)) ** .5, 2)

def chebyshev_distance(x, y):
    return max(abs(a - b) for a, b in zip(x, y))

print("Manhattan distance: ", manhattan_distance(X, Y))
print("Euclidean distance: ", euclidean_distance(X, Y))
print("Chebyshev distance: ", chebyshev_distance(X, Y))
```

Wynik działania programu:

```
Manhattan distance:  16.6
Euclidean distance:  9.89
Chebyshev distance:  6.9
```