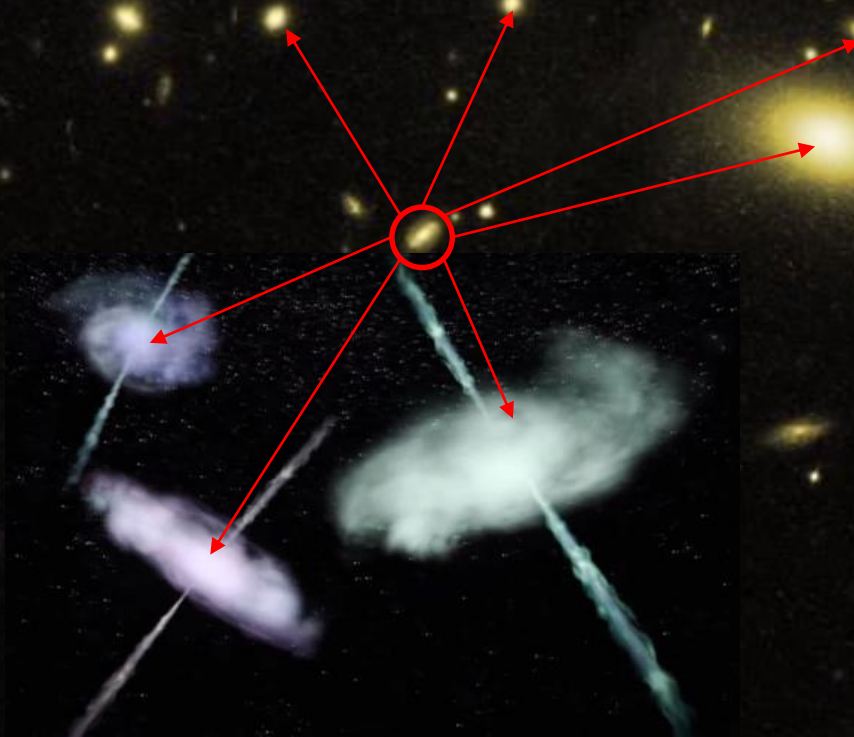


METODY INŻYNIERII WIEDZY ĆWICZENIA

Metoda

K Najbliższych Sąsiadów
K-Nearest Neighbours
(KNN)



Adrian Horzyk

Akademia Górniczo-Hutnicza

*Wydział Elektrotechniki, Automatyki, Informatyki
i Inżynierii Biomedycznej*

Katedra Automatyki i Inżynierii Biomedycznej

Laboratorium Biocybernetyki

30-059 Kraków, al. Mickiewicza 30, paw. C3/205

horzyk@agh.edu.pl, Google: Adrian Horzyk

ZALECENIA CO DO IMPLEMENTACJI METODY KNN

1. Wczytaj dane dla przykładowego zbioru uczącego zbiór Iris, umieszczając dane w tabeli lub 6 kolumnowej tablicy elementów typu float. **Zbiór uczący** składa się ze zbioru par $\langle \mathbf{x}^i, y^i \rangle$, gdzie $\mathbf{x}^i$ jest zbiorem parametrów wektorów $\mathbf{x}^i = \{x_1^i, \dots, x_n^i\}$ definiujących obiekty, y^i jest indeksem lub nazwą klasy, do której obiekt $\mathbf{x}^i$ należy i którą razem z

DANE UCZĄCE wczytywane z pliku tekstowego				
długość liścia	szerokość liścia	długość płatk	szerokość płatk	nazwa klasy
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3,0	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
...	...	...	...	...
7,0	3,2	4,7	1,4	Iris-versicolor
6,4	3,2	4,5	1,5	Iris-versicolor
...	...	...	...	...
6,2	3,4	5,4	2,3	Iris-virginica
5,9	3,0	5,1	1,8	Iris-virginica



TABLICA FLOAT W PROGRAMIE REPREZENTUJĄCA ZBIÓR DANYCH						
	Atrybut 1	Atrybut 1	Atrybut 1	Atrybut 1	Nr klasy	Nr klasy wynikowej
ID	0	1	2	3	4	5
0	5,1	3,5	1,4	0,2	1	1
1	4,9	3,0	1,4	0,2	1	2
2	4,7	3,2	1,3	0,2	1	1
...	...	...	...	...	...	...
50	7,0	3,2	4,7	1,4	2	2
51	6,4	3,2	4,5	1,5	2	3
...	...	...	...	...	...	...
148	6,2	3,4	5,4	2,3	3	3
149	5,9	3,0	5,1	1,8	3	2

poprawnie
błąd klasyfikacji metody
poprawnie
poprawnie
błąd klasyfikacji metody
błąd klasyfikacji metody
poprawnie

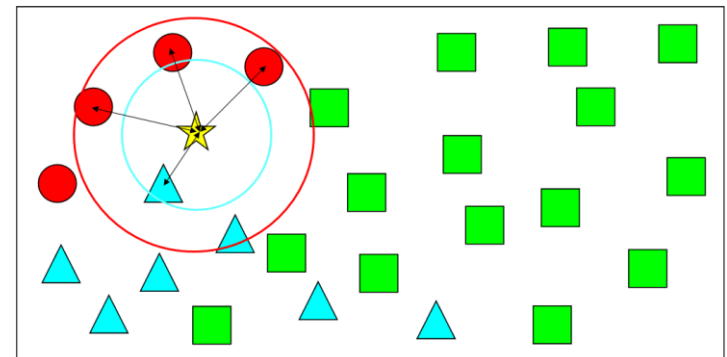
2. Pierwsze cztery kolumny reprezentują atrybuty: długość i szerokość listka i płatek kwiatu, natomiast 5. kolumna służy do zapisu numeru klasy: 1 – Iris-Setosa, 2 – Iris-Versicolor, 3 – Iris-Virginica. Ostatnia 6. kolumna służy do zapisu wyniku działania metody, czyli zapisania numeru klasy, do której metoda przypisała dany wzorec.

3. W podstawowej wersji tej metody dla wybranego k tworzymy podstawową pętlę obliczeniową, w której obliczamy odległość Euklidesa klasyfikowanego wzorca $\mathbf{x}$ (zadanego w postaci wektora określonych cech) do wszystkich wzorców uczących. Celem przyspieszenia działania metody kNN możemy pominąć obliczanie pierwiastka bo ci sami najbliżsi sąsiedzi zostaną wybrani dla sumy kwadratów jak dla odległości Euklidesa.

$$\|\mathbf{x} - \mathbf{x}^k\|_2 = \sqrt{\sum_{j=0}^J (x_j - x_j^k)^2}$$

$$\|\mathbf{x} - \mathbf{x}^i\|^2 = \sum_j (x_j - x_j^i)^2$$

4. Celem jest określenie najlepszego k oraz procentu wzorców poprawnie sklasyfikowanych, przy czym k nie może być większe niż liczność najmniej licznej klasy ze zbioru uczącego!
5. Należy zbudować pętlę obliczeniową, w której dla kolejnych wzorców szukamy k najbliższych sąsiadów do wybranego wzorca.



ZALECENIA CO DO IMPLEMENTACJI METODY KNN

6. Korzystamy tutaj z pomocniczej tablicy rankingowej, w której trzymamy k-najbliższych sąsiadów (id w tablicy wzorców) zgodnie z odległością Euklidesa lub sumę kwadratów:

$$\|x - x^k\|_2 = \sqrt{\sum_{j=0}^J (x_j - x_j^k)^2} \quad \|\mathbf{x} - \mathbf{x}^i\|^2 = \sum_j (x_j - x_j^i)^2$$

TABLICA FLOAT W PROGRAMIE REPREZENTUJĄCA ZBIÓR DANYCH							
	Atrybut 0	Atrybut 1	Atrybut 2	Atrybut 3	Nr klasy	Nr klasy wynikowej	
ID	0	1	2	3	4	5	
0	5,1	3,5	1,4	0,2	1	1	poprawnie
1	4,9	3,0	1,4	0,2	1	2	błąd klasyfikacji metody
2	4,7	3,2	1,3	0,2	1	1	poprawnie
	...	...	...	...	...	...	
50	7,0	3,2	4,7	1,4	2	2	poprawnie
51	6,4	3,2	4,5	1,5	2	3	błąd klasyfikacji metody
	...	...	...	...	...	...	
148	6,2	3,4	5,4	2,3	3	3	błąd klasyfikacji metody
149	5,9	3,0	5,1	1,8	3	2	poprawnie

2 kolumnowa TABLICA RANKINGOWA K najbliższych sąsiadów		
K	ID wzorca uczącego	Jego odległość Euklidesa do badanego
1	4	0,01000
2	2	0,01500
3	14	0,01800
4	5	0,02300
	...	
19	7	0,19200
20	6	0,20300

Do tablicy rankingowej dodajemy ID porównywanego wzorca w odpowiedniej pozycji przez proste wstawianie lub wstawianie połówkowe, jeśli jego odległość jest bliższa niż tych obecnie przechowywanych w tablicy

7. W pętli obliczeniowej przechodzimy po wszystkich wzorcach uczących, z którego każdy po kolei traktujemy jako testowy, wyznaczając dla niego najbliższych sąsiadów w tabeli rankingowej. Wobec tego w tej pętli jest kolejna pętla, w której przechodzimy po pozostałych 149 wzorcach, których odległość do badanego wzorca badany i spośród nich K najbliższych zapisujemy do tablicy rankingowej, podmieniając je w trakcie pętli na te najbliższe w zależności od obliczonej odległości Euklidesa. Jeśli odległość ta jest mniejsza niż najdalszego w tej tablicy, wtedy jego wyszukujemy pozycję w liście poprzez proste wstawianie lub wstawianie połówkowe, a następnie przesuwamy wzorce w liście rankingowej. Ostatni najdalszy wypada z listy.
8. Tworzymy sobie funkcję, która oblicza nam odległość Euklidesa dla dwóch wskazanych wzorców o indeksach i, j z tablicy, sumując kwadraty odejmych od siebie odpowiadających wartości dla poszczególnych atrybutów:
- Tworzymy pętlę
- ```
for (j=0; j<4; j++)
 sumaKwadratow += (T[k][j] - T[l][j]) * (T[k][j] - T[l][j]);
odlegloscEuklidesa = Math.Sqrt((T[k][j] - T[l][j]) * (T[k][j] - T[l][j]));
```

# ZALECENIA CO DO IMPLEMENTACJI METODY KNN

9. Po obliczeniu odległości Euklidesa sprawdzamy, czy jest ona mniejsza od odległości dla ostatniego wzorca w tablicy rankingowej R.
10. Jeśli nie, to przechodzimy do następnego wzorca i obliczamy odległość Euklidesa dla niego.
11. Jeśli tak, to algorytmem sortowania przez proste wstawianie lub algorytmem sortowania przez wstawianie połówkowe (dla bardziej ambitnych), wyszukujemy indeks w tablicy rankingowej R, idąc np. od najbardziej oddalonego wzorca w górę tablicy tak długo dopóki następny wzorzec nie będzie miał odległości mniejszej niż ta obliczona obecnie lub nie dojdziemy do początku tablicy.
12. Po wyznaczeniu indeksu w tablicy rankingowej R, dokonujemy przesunięcia wzorców o dłuższych odległościach w dół, przy czym ostatni (o najdłuższej odległości) wypada z niej. W miejscu wyznaczonego indeksu wpisujemy nowe ID wzorca oraz jego odległość od badanego.
13. W taki sposób wyznaczamy tablicę rankingową dla K najbliższych sąsiadów. Ambitni mogą zamiast tablicy zbudować listę rankingową, która nie wymaga przesuwania elementów w tablicy, więc działa szybciej!
14. Po wyznaczeniu tablicy rankingowej i konkretnego K, które badamy, możemy przejść do głosowania K najbliższych wzorców (na podstawie ich klas) na temat tego wzorca, którego klasę próbujemy określić. Głosowanie przebiega na zasadzie zwykłej większości głosów, a głosy zliczamy na podstawie numerów klas określonych dla porównywanych wzorców uczących.

Zliczanie głosów możemy wykonać za pomocą małej pomocniczej tablicy:

| Iris-setosa | Iris-versicolor | Iris-virginica |
|-------------|-----------------|----------------|
| 1           | 2               | 3              |
| 15          | 93              | 41             |

**Zwycięska klasa!**

15. W ogólnej wersji metody tworzymy tablicę najbliższych wzorców do dla poszczególnych k, wyznaczamy dla nich zwycięską klasę i najlepsze k.

| 2 kolumnowa TABLICA RANKINGOWA K najbliższych sąsiadów |                    |                                      |
|--------------------------------------------------------|--------------------|--------------------------------------|
| K                                                      | ID wzorca uczącego | Jego odległość Euklidesa do badanego |
| 1                                                      | 4                  | 0,01000                              |
| 2                                                      | 2                  | 0,01500                              |
| 3                                                      | 14                 | 0,01800                              |
| 4                                                      | 5                  | 0,02300                              |
|                                                        | ...                |                                      |
| 19                                                     | 7                  | 0,19200                              |
| 20                                                     | 6                  | 0,20300                              |

# ZALECENIA CO DO IMPLEMENTACJI METODY KNN

16. Wszystko umieszczamy w pętli obliczeniowej i obliczamy te odległości i tabele Rankingowe dla kolejno wszystkich wzorców na zasadzie każdy względem pozostałych wzorców uczących i dla wszystkich badanych  $K = 1, \dots, 20$ :

| TABLICA FLOAT W PROGRAMIE REPREZENTUJĄCA ZBIÓR DANYCH |           |           |           |           |          |                    |
|-------------------------------------------------------|-----------|-----------|-----------|-----------|----------|--------------------|
|                                                       | Atrybut 0 | Atrybut 1 | Atrybut 2 | Atrybut 3 | Nr klasy | Nr klasy wynikowej |
| ID                                                    | 0         | 1         | 2         | 3         | 4        | 5                  |
| 0                                                     | 5,1       | 3,5       | 1,4       | 0,2       | 1        | 1                  |
| 1                                                     | 4,9       | 3,0       | 1,4       | 0,2       | 1        | 2                  |
| 2                                                     | 4,7       | 3,2       | 1,3       | 0,2       | 1        | 1                  |
| ...                                                   | ...       | ...       | ...       | ...       | ...      | ...                |
| 50                                                    | 7,0       | 3,2       | 4,7       | 1,4       | 2        | 2                  |
| 51                                                    | 6,4       | 3,2       | 4,5       | 1,5       | 2        | 3                  |
| ...                                                   | ...       | ...       | ...       | ...       | ...      | ...                |
| 148                                                   | 6,2       | 3,4       | 5,4       | 2,3       | 3        | 3                  |
| 149                                                   | 5,9       | 3,0       | 5,1       | 1,8       | 3        | 2                  |

Wzorec \* od którego obliczamy odległość Euklidesa

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

| TABLICA FLOAT W PROGRAMIE REPREZENTUJĄCA ZBIÓR DANYCH |           |           |           |           |          |                    |
|-------------------------------------------------------|-----------|-----------|-----------|-----------|----------|--------------------|
|                                                       | Atrybut 0 | Atrybut 1 | Atrybut 2 | Atrybut 3 | Nr klasy | Nr klasy wynikowej |
| ID                                                    | 0         | 1         | 2         | 3         | 4        | 5                  |
| 0                                                     | 5,1       | 3,5       | 1,4       | 0,2       | 1        | 1                  |
| 1                                                     | 4,9       | 3,0       | 1,4       | 0,2       | 1        | 2                  |
| 2                                                     | 4,7       | 3,2       | 1,3       | 0,2       | 1        | 1                  |
| ...                                                   | ...       | ...       | ...       | ...       | ...      | ...                |
| 50                                                    | 7,0       | 3,2       | 4,7       | 1,4       | 2        | 2                  |
| 51                                                    | 6,4       | 3,2       | 4,5       | 1,5       | 2        | 3                  |
| ...                                                   | ...       | ...       | ...       | ...       | ...      | ...                |
| 148                                                   | 6,2       | 3,4       | 5,4       | 2,3       | 3        | 3                  |
| 149                                                   | 5,9       | 3,0       | 5,1       | 1,8       | 3        | 2                  |

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

Wzorec \* od którego obliczamy odległość Euklidesa

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

| TABLICA FLOAT W PROGRAMIE REPREZENTUJĄCA ZBIÓR DANYCH |           |           |           |           |          |                    |
|-------------------------------------------------------|-----------|-----------|-----------|-----------|----------|--------------------|
|                                                       | Atrybut 0 | Atrybut 1 | Atrybut 2 | Atrybut 3 | Nr klasy | Nr klasy wynikowej |
| ID                                                    | 0         | 1         | 2         | 3         | 4        | 5                  |
| 0                                                     | 5,1       | 3,5       | 1,4       | 0,2       | 1        | 1                  |
| 1                                                     | 4,9       | 3,0       | 1,4       | 0,2       | 1        | 2                  |
| 2                                                     | 4,7       | 3,2       | 1,3       | 0,2       | 1        | 1                  |
| ...                                                   | ...       | ...       | ...       | ...       | ...      | ...                |
| 50                                                    | 7,0       | 3,2       | 4,7       | 1,4       | 2        | 2                  |
| 51                                                    | 6,4       | 3,2       | 4,5       | 1,5       | 2        | 3                  |
| ...                                                   | ...       | ...       | ...       | ...       | ...      | ...                |
| 148                                                   | 6,2       | 3,4       | 5,4       | 2,3       | 3        | 3                  |
| 149                                                   | 5,9       | 3,0       | 5,1       | 1,8       | 3        | 2                  |

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

Pozostałe wzorce uczące do których obliczana jest ta odległość

Wzorec \* od którego obliczamy odległość Euklidesa

| Iris-setosa | Iris-versicolor | Iris-virginica | Dla K |
|-------------|-----------------|----------------|-------|
| 1           | 2               | 3              | 1     |
| 1           | 0               | 0              | 1     |

Zwycięska klasa!

| Iris-setosa | Iris-versicolor | Iris-virginica | Dla K |
|-------------|-----------------|----------------|-------|
| 1           | 2               | 3              | 2     |
| 0           | 2               | 0              | 2     |

Zwycięska klasa!

| Iris-setosa | Iris-versicolor | Iris-virginica | Dla K |
|-------------|-----------------|----------------|-------|
| 1           | 2               | 3              | 3     |
| 0           | 2               | 1              | 3     |

Zwycięska klasa!

| Iris-setosa | Iris-versicolor | Iris-virginica | Dla K |
|-------------|-----------------|----------------|-------|
| 1           | 2               | 3              | 4     |
| 0           | 3               | 1              | 4     |

Zwycięska klasa!

| Iris-setosa | Iris-versicolor | Iris-virginica | Dla K |
|-------------|-----------------|----------------|-------|
| 1           | 2               | 3              | 5     |
| 1           | 3               | 1              | 5     |

Zwycięska klasa!

| Iris-setosa | Iris-versicolor | Iris-virginica | Dla K |
|-------------|-----------------|----------------|-------|
| 1           | 2               | 3              | 6     |
| 1           | 3               | 2              | 6     |

Zwycięska klasa!

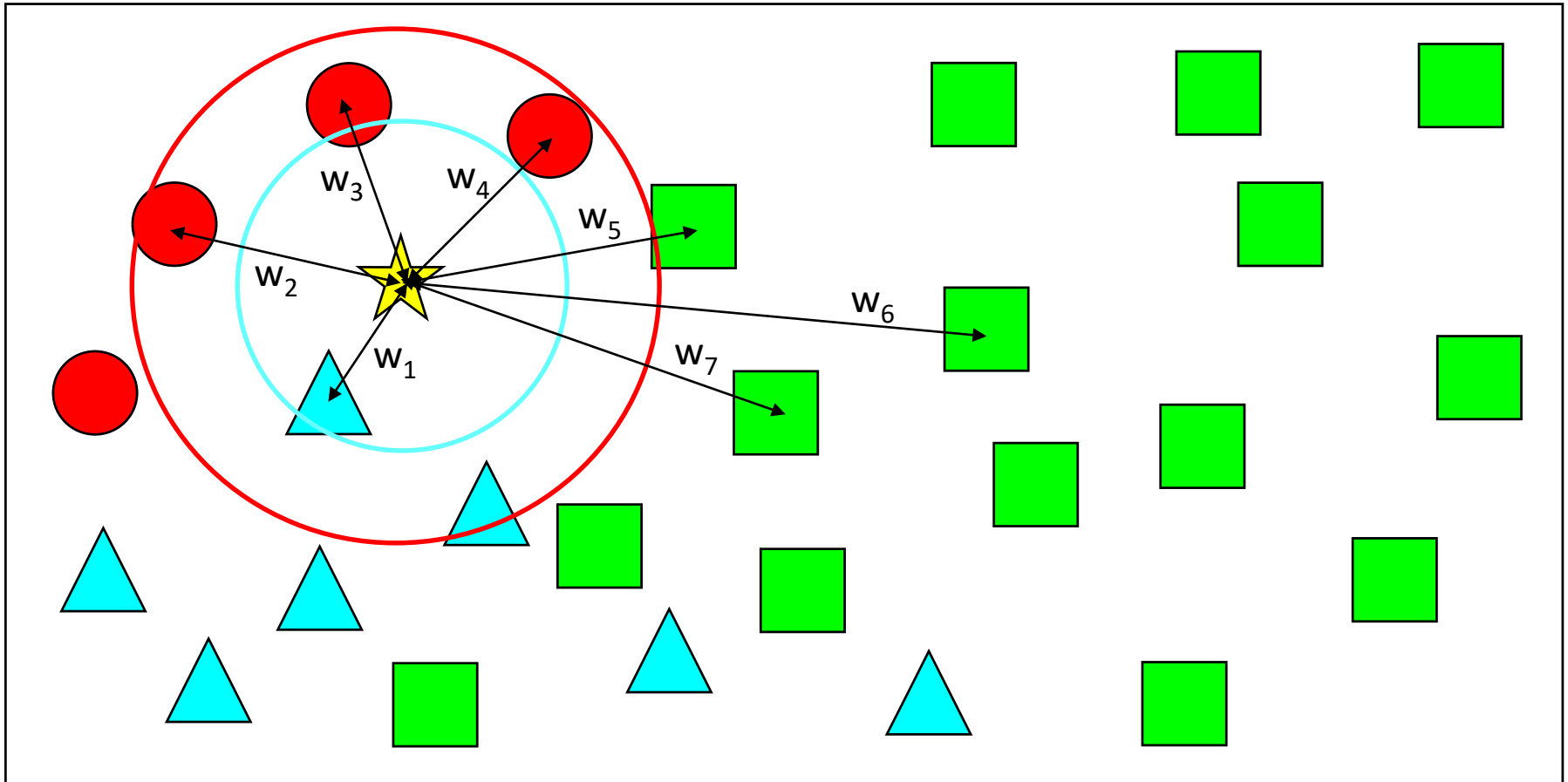
| Iris-setosa | Iris-versicolor | Iris-virginica | Dla K |
|-------------|-----------------|----------------|-------|
| 1           | 2               | 3              | 20    |
| 1           | 15              | 4              | 20    |

Zwycięska klasa!

# MODYFIKACJE METODY K NAJBLIŻSZYCH SĄSIADÓW

## Metoda Ważonych Odległości Najbliższych Sąsiadów (Distance Weighted Nearest Neighbors)

proceedzi do głosowania na temat klasyfikacji gwiazdki biorąc pod uwagę k najbliższych sąsiadów, lecz ich **głosy są ważone w zależności od ich odległości** (dla wybranej metryki) do gwiazdki: im dalej jest głosujący wzorzec tym ma mniejszą wagę. A więc wzorce położone najbliżej będą miały największy wpływ na wynik klasyfikacji. Można wziąć też pod uwagę wszystkie wzorce i ich odległości od wzorca klasyfikowanego, lecz wtedy trzeba wagę podzielić przed licznosc klasy, do której należy, gdyż w odwrotnym przypadku uprzywilejowane byłyby klasy najbardziej liczne!



# WAŻENIE ODLEGŁOŚCI ORAZ NORMALIZACJA WZGLĘDEM LICZNOŚCI POSZCZEGÓLNYCH KLAS

## Jak dobrać wagi dla Metody Ważonych Odległości Najbliższych Sąsiadów?

Należy wziąć pod uwagę taką zależność wagi od odległości, żeby waga malała wraz ze zwiększającą się odległością. Możliwości jest kilka (proszę poeksperymentować), np.:

$$w = \frac{d_{max} - d}{d_{max} \cdot N_{class}}$$

$$w = \frac{1}{d \cdot N_{class}}$$

$$w = \frac{1}{\sqrt[m]{d} \cdot N_{class}}$$

Gdzie:

$d$  – odległość Euklidesa badanego wzorca od wzorca uczącego

$m$  – stopień pierwiastka: 2, 3, 4...

$d_{max}$  – to maksymalna odległość wzorca spośród  $k$ -najbliższych (badanych)

$N_{class}$  – to liczność klasy „class”, do której należy wzorec uczący względem którego wyznaczamy odległość i który bierze udział w tym ważonym odległością „głosowaniu”.

KONKURS NA NAJLEPSZĄ ZMODYFIKOWANĄ METODĘ KNN!