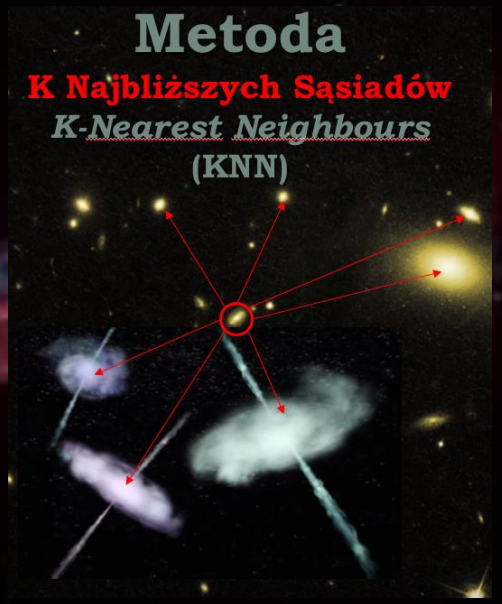
METODY INŻYNIERII WIEDZY

WALIDACJA KRZYŻOWA dla ZAAWANSOWANEGO KLASYFIKATORA KNN

ĆWICZENIA



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

WALIDACJA KRZYŻOWA
k-fold CROSS-VALIDATION

- ✓ Umożliwia wykorzystanie całego zbioru danych zarówno do uczenia, jak również do walidacji modelu.
- ✓ Służy do określenia jakości modelu już w trakcie jego adaptacji / uczenia, w celu wyeliminowania problemu przeuczenia się (*overfitting*).
- ✓ Polega na podziale zbioru uczącego na k równolicznych podzbiorów, z których k-1 jest wykorzystanych do uczenia / adaptacji modelu, a 1 podzbiór służy do walidacji modelu.

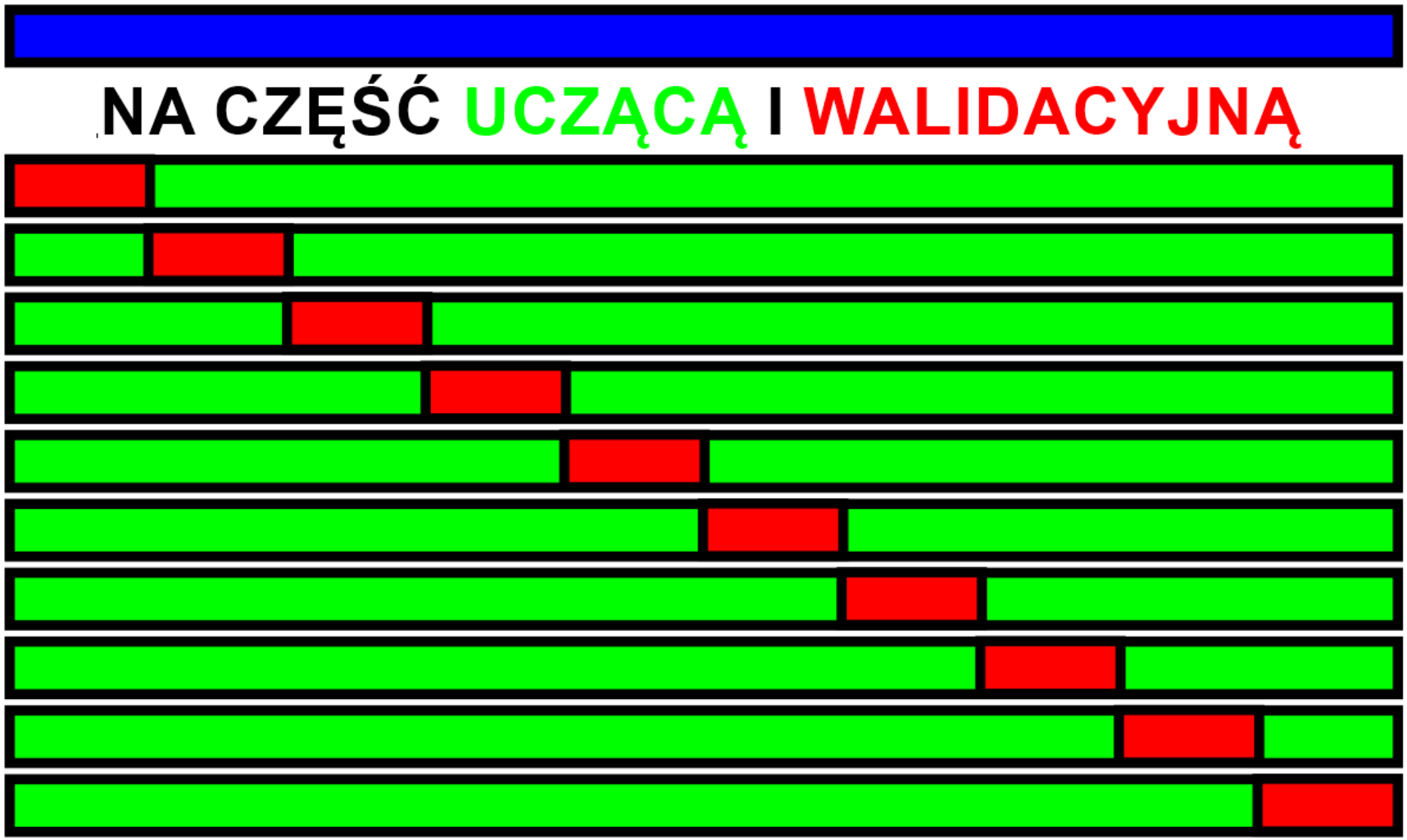
WALIDACJA KRZYŻOWA
DLA KLASYFIKATORA KNN

- ✓ Zbuduj zaawansowany klasyfikator kNN i dobierz dla niego najlepsze k z wykorzystaniem walidacji krzyżowej z doborem losowym lub sekwencyjnym wzorców walidacyjnych proporcjonalnie wybranych dla każdej klasy.
- ✓ Zaawansowany klasyfikator kNN wykorzystuje głosowanie poprzez ważone odległości od klasyfikowanego punktu. Im dalej głosujący się znajduje tym jego głos ma mniejsze znaczenie.
- ✓ Ponadto zaawansowany klasyfikator stosuje normalizację względem ilości reprezentantów danej klasy w zbiorze uczącym dla uniknięcia faworyzowania klas bardziej licznych.
- ✓ Dodatkowo dokonujemy eksperymentowania z doborem ilości podziałów zbioru dla walidacji krzyżowej od 3 do 10.

PRZYKŁAD

10-krotnej walidacji krzyżowej

CAŁY ZBIÓR DANYCH DZIELONY JEST



NA CZĘŚĆ UCZĄCĄ I WALIDACYJNĄ

PARAMETRY METODY WALIDACJI KRZYŻOWEJ

Dobór parametru k zależny jest od wielkości zbioru danych i ich rodzaju. Dla dużych zbiorów danych stosuje się k=3 w celu zmniejszenia ilości adaptacji modelu. Dla mniejszych zbiorów danych zwykle stosuje się większe wartości k, żeby nie uszczuplać zbioru uczącego za bardzo, co mogłoby spowodować budowę słabej jakości modeli. Najczęściej stosuje się k=10.

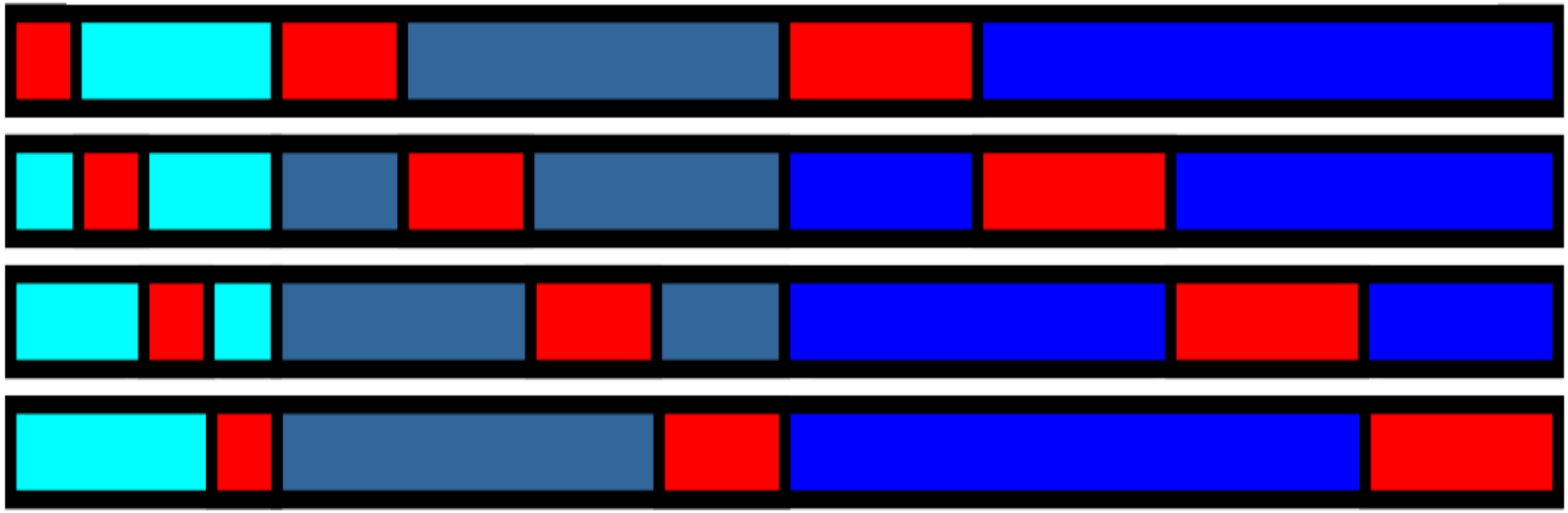
Sposób podziału zbioru danych na k podzbiorów jest niemniej istotny, gdyż jeśli wzorce są posortowane wg klas w zbiorze uczącym, wtedy wybór kolejnych podzbiorów może powodować uwzględnienie w walidacji tylko wzorców jednej klasy, a ponadto znaczne uszczuplenie wzorców uczących dla tej klasy, co jest bardzo niekorzystne z punktu widzenia budowy modelu!

Najlepiej dobierać wzorce **proporcjonalnie** do ich liczności i reprezentacji poszczególnych klas tak, aby były reprezentatywne.

4-krotnej walidacja krzyżowa z proporcjonalnym wyborem wzorców

CAŁY ZBIÓR DANYCH DZIELONY JEST PROPORCJONALNIE

DO LICZNOŚCI KLAS NA CZĘŚĆ UCZĄCĄ I WALIDACYJNĄ



N-krotna walidacja krzyżowa
leave one out cross-validation

N-krotna walidacja krzyżowa zakłada dobór k=N, gdzie N to ilość wzorców zbioru danych uczących.

W takim przypadku tylko 1 wzorzec jest wykluczany ze zbioru uczącego i traktowany jako walidacyjny.

Nauka jest więc powtarzana N-krotnie na zbiorach utworzonych poprzez pominięcie 1 wzorca stosowanego do walidacji.

Usunięcie jednego wzorca zwykle nie daje reprezentatywnych wyników działania odnośnie jakości modelu i jego możliwości generalizacji.

MODYFIKACJE METODY WALIDACJI KRZYŻOWEJ

W celu uzyskania lepszego działania metody, zbiór danych powinien być dzielony w taki sposób, żeby stosunek reprezentantów poszczególnych klas w zbiorze walidacyjnym był mniej więcej taki sam jak w całym zbiorze uczącym.

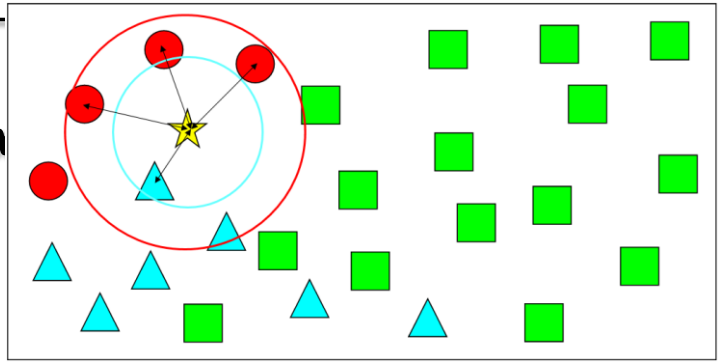
Z tego powodu zaleca się posortowanie zbioru danych uczących w taki sposób, żeby można było z każdej klasy wybierać kolejne podzbiory do walidacji krzyżowej w sposób **reprezentatywny** dla całego zbioru danych uczących.

Czasami stosuje się również uproszczoną metodę selekcji 1 / k wzorców ze zbioru uczącego do walidacji, polegającą na **losowaniu tych wzorców z całego zbioru (*random subsampling*)**, a pozostałe wykorzystywane są do adaptacji / uczenia modelu.

ZALECENIA CO DO IMPLEMENTACJI METODY KNN

1. **Wczytanie danych z zbioru uczącego, np. zbiór Iris, umieszczając dane w tabeli lub 5 kolumnowej tablicy elementów typu float.**
   Zbiór uczący składa się ze zbioru par $<\mathbf{x}^i, y^i>$, gdzie
   $\mathbf{x}^i$ jest zbiorem parametrów wektorów $\mathbf{x}^i = \{x^i_1, \ldots, x^i_n\}$ definiujących obiekty,
   y^i jest indeksem lub nazwą klasy, do której obiekt $\mathbf{x}^i$ należy i którą razem z innymi obiektami tej klasy definiuje.
2. **W podstawowej wersji tej metody dla wybranego k tworzymy podstawową pętlę obliczeniową, w której obliczamy odległość Euklidesa klasyfikowanego wzorca x (zadanego w postaci wektora określonych cech) do wszystkich wzorców uczących:**

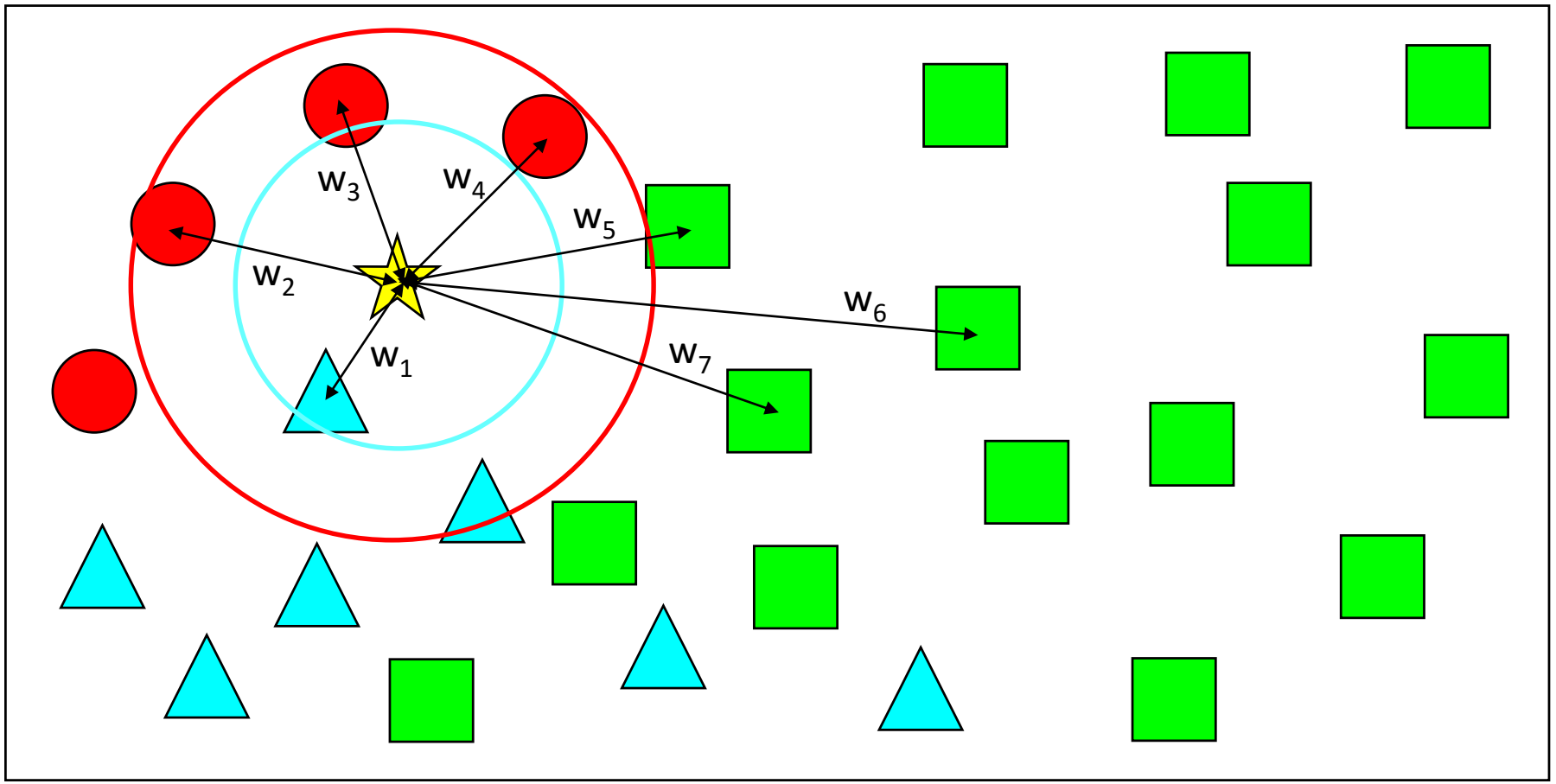
$$\left\| x - x^k \right\|_2 = \sqrt{\sum_{j=0}^{J} (x_j - x$$



3. **Wynikiem jest ta klasa, która powiązana jest z największą ilością wzorców spośród tych k najbliższych. Wzorzec testowy do klasyfikacji podajemy z konsoli.**
4. **W ogólnej wersji metody tworzymy tablicę najbliższych wzorców do dla poszczególnych k, wyznaczamy dla nich zwycięską klasę i najlepsze k.**

MODYFIKACJE METODY K NAJBLIŻSZYCH SĄSIADÓW

- ✓ **Metoda Ważonych Odległości Najbliższych Sąsiadów (Distance Weighted Nearest Neighbors) prowadzi do głosowania na temat klasyfikacji gwiazdki biorąc pod uwagę k najbliższych sąsiadów lub nawet wszystkie wzorce, 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:**



ZAAWANSOWANA METODA K NAJBLIŻSZYCH SĄSIADÓW

- ✓ Z ważeniem odległości Najbliższych Sąsiadów
- ✓ Z normalizacją względem ilości reprezentantów klas: N_1, N_2, N_3,… , N_L
  gdzie $N = N_1 + N_2 + N_3 + \ldots + N_L$ jest ilością wszystkich wzorców uczących
- ✓ Walidacja krzyżową
- ✓ Dobraniem k dla metody kNN, rozpoczynamy implementację od k=N,
  gdzie N to ilość wzorców uczących wszystkich klas
  po ich wyselekcjonowaniu metodą walidacji krzyżowej (cross-validation)

Wykorzystujemy więc znormalizowane sumy ważone odległości Euklidesa:

$$V_l = \frac{\sum_{k=0}^{K_l} \sqrt{\sum_{j=0}^{J} \left(x_j - x_j^k\right)^2}}{N_l}$$

Dla każde klasy wyznaczamy taki współczynnik głosowania (voting coefficient) i wybieramy najmniejszy dla określonego K dla metody KNN, decydujący o klasyfikacji wzorca do określonej klasy l, przy założeniu, iż mamy do czynienia z L klasami.

UWAGI IMPLEMENTACYJNE

- ✓ **W trakcie implementacji możemy posłużyć się tabelą lub tablicą (wtedy wartościom symbolicznym należy nadać wartości liczbowe) do reprezentowania i rozróżnienia wzorców uczących i walidacyjnych oraz zapisania klasy uzyskanej w wyniku działania metody, np.:**

IRIS	ATTRIBUTES				TRAINED	true - learning	EVALUATED
No	leaf-length	leaf-width	petal-length	petal-width	class name	false - validation	class name
1	5,10	3,50	1,40	0,20	Iris-setosa	false	
2	4,90	3,00	1,40	0,20	Iris-setosa	false	
3	4,70	3,20	1,30	0,20	Iris-setosa	false	
4	4,60	3,10	1,50	0,20	Iris-setosa	false	
5	5,00	3,60	1,40	0,20	Iris-setosa	false	
6	5,40	3,90	1,70	0,40	Iris-setosa	true	
7	4,60	3,40	1,40	0,30	Iris-setosa	true	
8	5,00	3,40	1,50	0,20	Iris-setosa	true	
9	4,40	2,90	1,40	0,20	Iris-setosa	true	
10	4,90	3,10	1,50	0,10	Iris-setosa	true	
11	5,40	3,70	1,50	0,20	Iris-setosa	true	
12	4,80	3,40	1,60	0,20	Iris-setosa	true	
13	4,80	3,00	1,40	0,10	Iris-setosa	true	
14	4,30	3,00	1,10	0,10	Iris-setosa	true	
15	5,80	4,00	1,20	0,20	Iris-setosa	true	

UWAGI IMPLEMENTACYJNE

- ✓ **Najlepiej przeglądać tablicę wszystkich wzorców obliczać wszystkie V_l równocześnie dodając odległość do odpowiedniej komórki tablicy V_l:**

K	1	2	3	4	5
V_1					
V_2					
V_3					

- ✓ **Warto zastosować tablicę najbliższych sąsiadów, którą aktualizujemy w trakcie przeglądania zbioru uczącego i obliczania odległości Euklidesa do wybranego wzorca testowego lub walidacyjnego. Dzięki temu możemy równocześnie wyznaczyć K najbliższych sąsiadów dla dowolnego K.**
  Do określenia najbliższych wzorców stosujemy indeks wzorca z tablicy głównej podobnie klucz główny w relacyjnych bazach danych :

K najbliższych	Odległość Euklidesa	Indeks wzorca	Wynik klasyfikacji
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			