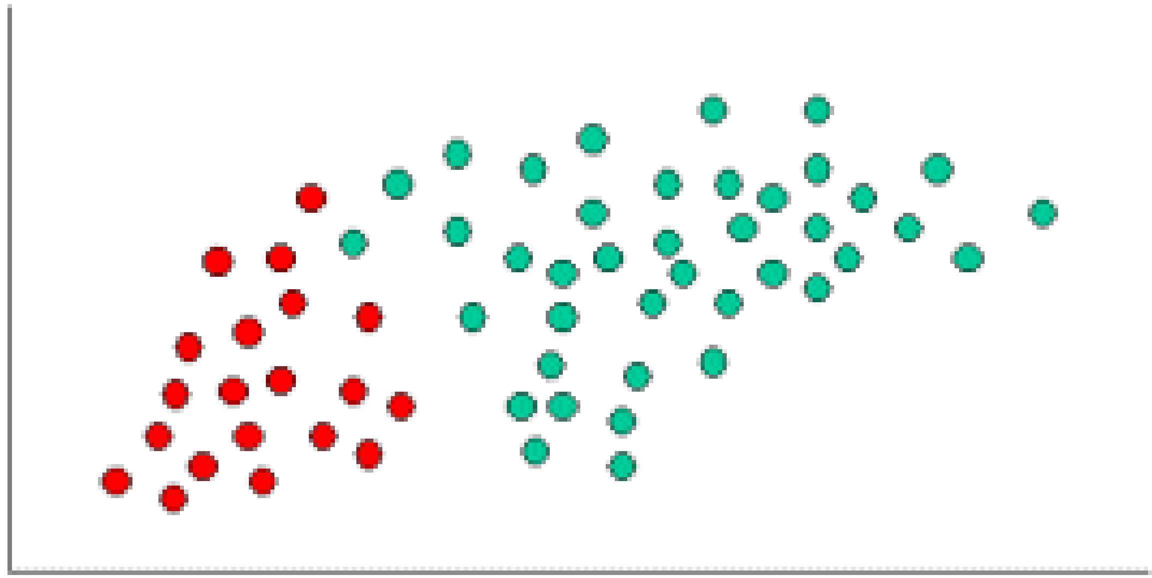
Klasyfikatory: k-NN oraz naiwny Bayesa

Agnieszka Nowak – Brzezińska

Wykład IV

Naiwny klasyfikator Bayesa

- Naiwny klasyfikator bayesowski jest prostym probabilistycznym klasyfikatorem.
- Zakłada się wzajemną niezależność zmiennych niezależnych (tu naiwność)
- Bardziej opisowe może być określenie- „model cech niezależnych”.
- Model prawdopodobieństwa można wyprowadzić korzystając z twierdzenia Bayesa.
- W zależności od rodzaju dokładności modelu prawdopodobieństwa, naiwne klasyfikatory bayesowskie można „uczyć” bardzo skutecznie w trybie uczenia z nadzorem.



- Jeśli wiemy, że kulek czerwonych jest 2 razy mniej niż zielonych (bo czerwonych jest 20 a zielonych 40) to prawdopodobieństwo tego, że kolejna (nowa) kulka będzie koloru zielonego jest dwa razy większe niż tego, że kulka będzie czerwona.
- Dlatego możemy napisać, że znane z góry prawdopodobieństwa:

$$\text{Prawdopodobieństwo zielonych} = \frac{\text{liczba zielonych}}{\text{liczba wszystkich kulek (zielonych i czerwonych)}}$$

$$\text{Prawdopodobieństwo czerwonych} = \frac{\text{liczba czerwonych}}{\text{liczba wszystkich kulek (zielonych i czerwonych)}}$$

$$\text{Prawdopodobieństwo zielonych} = \frac{\text{liczba zielonych}}{\text{liczba wszystkich kulek (zielonych i czerwonych)}}$$

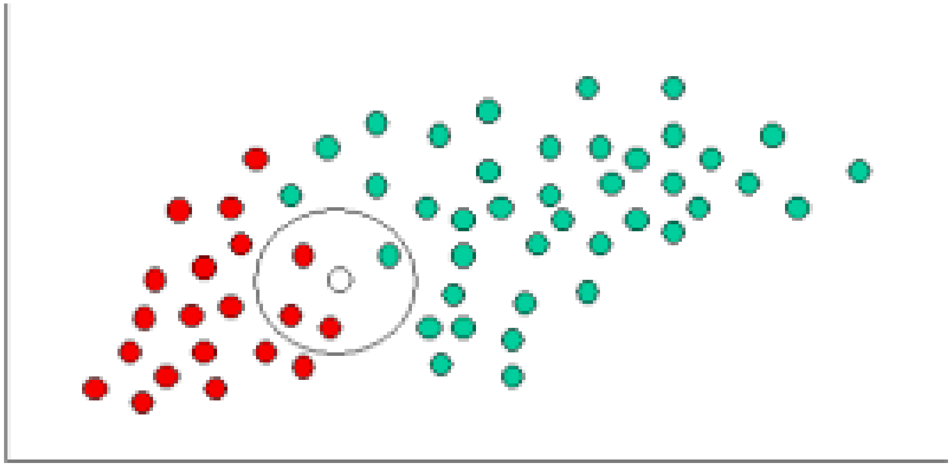
$$\text{Prawdopodobieństwo czerwonych} = \frac{\text{liczba czerwonych}}{\text{liczba wszystkich kulek (zielonych i czerwonych)}}$$

Jeśli więc czerwonych jest 20 a zielonych 40, to razem wszystkich jest 60. Więc

$$\text{Prawdopodobieństwo zielonych} = \frac{40}{60} = 0.66$$

$$\text{Prawdopodobieństwo czerwonych} = \frac{20}{60} = 0.33$$

Więc teraz gdy mamy do czynienia z nową kulką (na rysunku – biała):



- To spróbujmy ustalić jaka ona będzie. Dokonujemy po prostu klasyfikacji kulki do jednej z dwóch klas: zielonych bądź czerwonych.
- Jeśli weźmiemy pod uwagę sąsiedztwo białej kulki takie jak zaznaczono, a więc do 4 najbliższych sąsiadów, to widzimy, że wśród nich są 3 kulka czerwone i 1 zielona.
- Obliczamy liczbę kulek w sąsiedztwie należących do danej klasy : zielonych bądź czerwonych z wzorów:

$$\text{kulka X jest zielona w swoim sąsiedztwie} = \frac{\text{liczba kulek zielonych w sąsiedztwie kulki X}}{\text{liczba wszystkich kulek zielonych}}$$

$$\text{kulka X jest czerwona w swoim sąsiedztwie} = \frac{\text{liczba kulek czerwonych w sąsiedztwie kulki X}}{\text{liczba wszystkich kulek czerwonych}}$$

W naszym przypadku, jest dziwnie, bo akurat w sąsiedztwie kulki X jest więcej kulek czerwonych niż zielonych, mimo, iż kulek zielonych jest ogólnie 2 razy więcej niż czerwonych. Dlatego zapiszemy, że

$$\text{kulka X jest zielona w swoim sąsiedztwie} = \frac{1}{40}$$

$$\text{kulka X jest czerwona w swoim sąsiedztwie} = \frac{3}{20}$$

Dlatego ostatecznie powiemy, że

Prawdopodobieństwo że kulka X jest zielona = prawdopodobieństwo kulki zielonej * prawdopodobieństwo, że kulka X jest zielona w swoim sąsiedztwie

$$= \frac{40}{60} * \frac{1}{40} = \frac{1}{60}$$

Prawdopodobieństwo że kulka X jest czerwona = prawdopodobieństwo kulki czerwonej * prawdopodobieństwo, że kulka X jest czerwona w swoim sąsiedztwie =

$$\frac{20}{60} * \frac{3}{20} = \frac{1}{20}$$

Ostatecznie klasyfikujemy nową kulkę X do klasy kulek czerwonych, ponieważ ta klasa dostarcza nam większego prawdopodobieństwa posteriori.

Algorytm *k* najbliższych sąsiadów (lub algorytm k-nn z ang. *k nearest neighbours*)

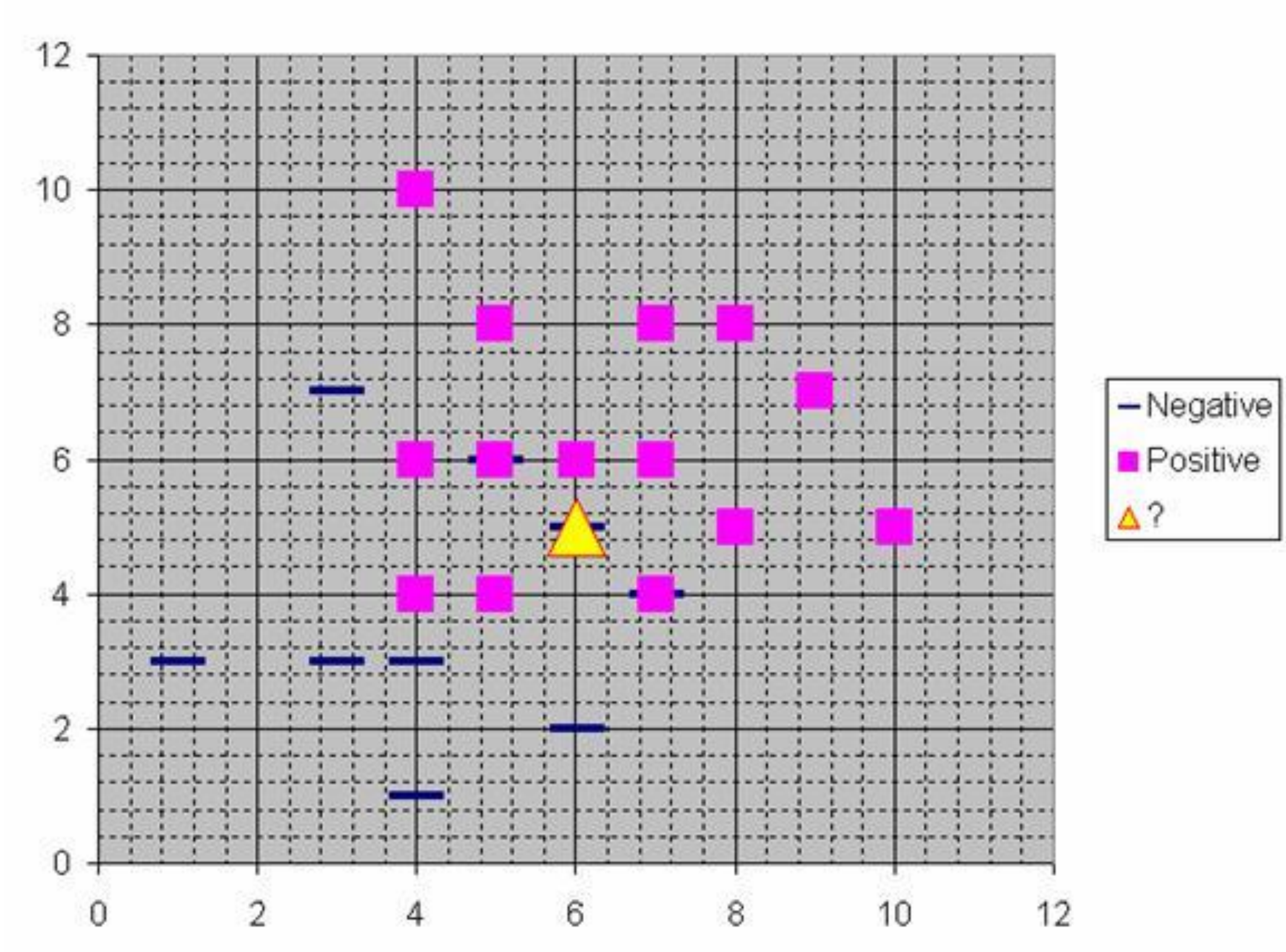
– jeden z algorytmów regresji nieparametrycznej używanych w statystyce do prognozowania wartości pewnej zmiennej losowej. Może również być używany do klasyfikacji.

-

Założenia

- Dany jest zbiór uczący zawierający obserwacje z których każda ma przypisany wektor zmiennych objaśniających oraz wartość zmiennej objaśnianej *Y*.
- Dana jest obserwacja *C* z przypisanym wektorem zmiennych objaśniających dla której chcemy prognozować wartość zmiennej objaśnianej *Y*.

Do której klasy przypisać nowy obiekt ?



Algorytm *k* najbliższych sąsiadów (algorytm *k*-NN)

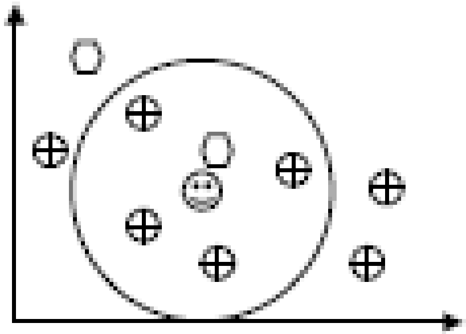
- **Ogólny schemat:**

Krok 1: Poszukaj *k* najbliższych obiektów (sąsiadów) dla x_q

Krok 2: Głosuj wśród *k* najbliższych sąsiadów w celu wyznaczania klasy, do której należy x_q

1-NN, decyzja jest ○

5-NN, decyzja jest ⊕



Zaleta: Bardziej odporny na szumy

- Wyznaczanie odległości obiektów: odległość euklidesowa

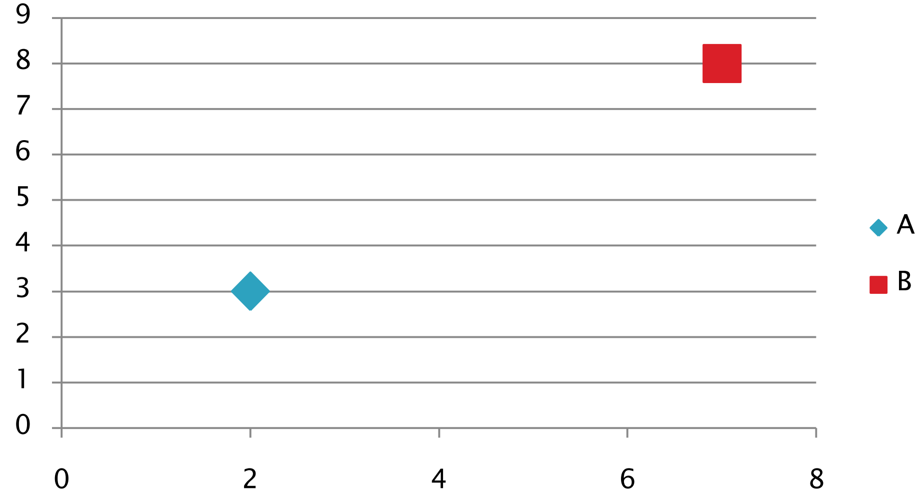
$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ij} - x_{jk})^2}$$

Obiekty są analizowane w ten sposób , że oblicza się odległości bądź podobieństwa między nimi. Istnieją różne miary podobieństwa czy odległości. Powinny być one wybierane konkretnie dla typu danych analizowanych: inne są bowiem miary typowo dla danych binarnych, inne dla danych nominalnych a inne dla danych numerycznych.

gdzie: **x,y** - to wektory wartości cech porównywanych obiektów w przestrzeni p-wymiarowej, gdzie odpowiednio wektory wartości to: oraz .

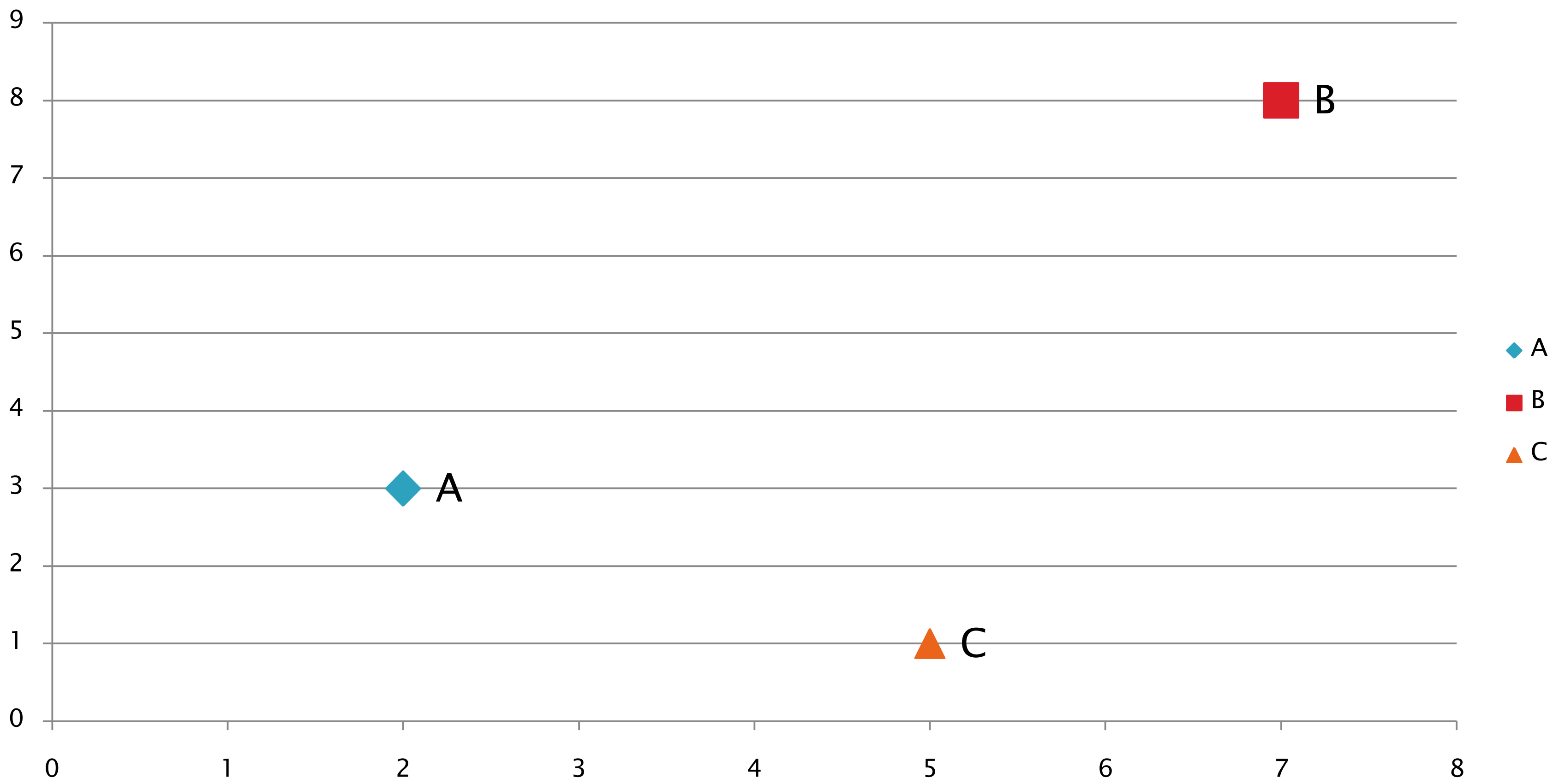
Nazwa	Wzór
odległość euklidesowa	$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$
odległość kątowa	$p(x,y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum_i y_i^2}}$
współczynnik korelacji liniowej Pearsona	$p(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$
Miara Gowera	$p(x,y) = \frac{\sum_{k=1}^{n} s_{ijk} w_{ijk}}{\sum_{k=1}^{n} w_{ijk}}$

Oblicz odległość punktu A o współrzędnych (2,3) do punktu B o współrzędnych (7,8).



D (A,B) = pierwiastek $((7-2)^2 + (8-3)^2)$ = pierwiastek (25 + 25) = pierwiastek (50) = 7.07

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$



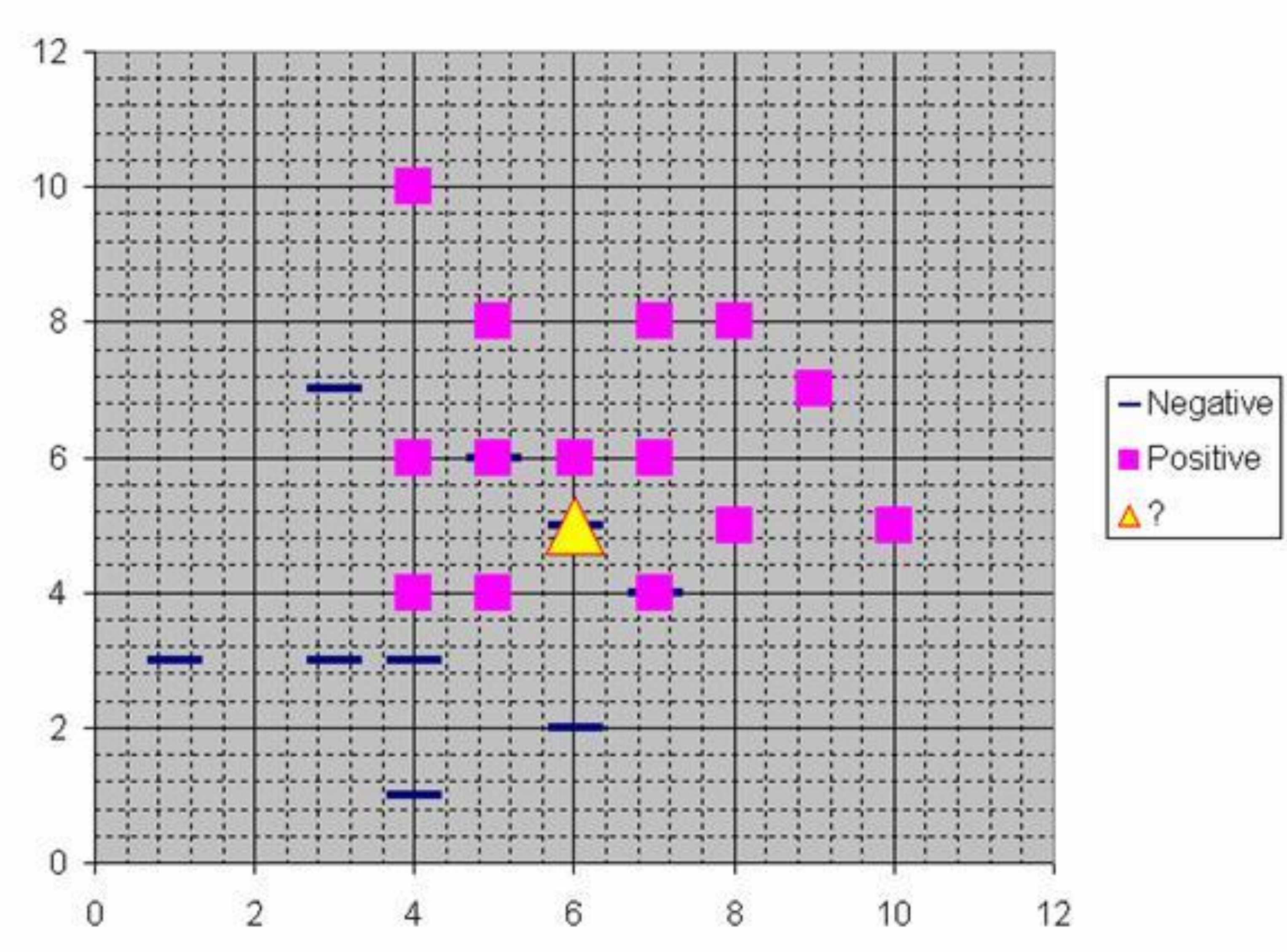
- Mając dane punkty:
- A(2,3), B(7,8) oraz C(5,1) oblicz odległości między punktami:
- D (A,B) = pierwiastek $((7\text{-}2)^2 + (8\text{-}3)^2)$ = pierwiastek (25 + 25) = pierwiastek (50) = 7.07
- D (A,C) = pierwiastek $((5\text{-}2)^2 + (3\text{-}1)^2)$ = pierwiastek (9 + 4) = pierwiastek (13) = 3.60
- D (B,C) = pierwiastek $((7\text{-}5)^2 + (3\text{-}8)^2)$ = pierwiastek (4 + 25) = pierwiastek (29) = 5.38

Przebieg algorytmu:

1. porównanie wartości zmiennych objaśniających dla obserwacji *C* z wartościami tych zmiennych dla każdej obserwacji w zbiorze uczącym.
2. wybór *k* (ustalona z góry liczba) najbliższych do *C* obserwacji ze zbioru uczącego.
3. Uśrednienie wartości zmiennej objaśnianej dla wybranych obserwacji, w wyniku czego uzyskujemy prognozę.

Przez "najbliższą obserwację" mamy na myśli, taką obserwację, której odległość do analizowanej przez nas obserwacji jest możliwie najmniejsza.

Do której klasy przypisać nowy obiekt ?



Algorytm *k* najbliższych sąsiadów (algorytm *k*-NN)

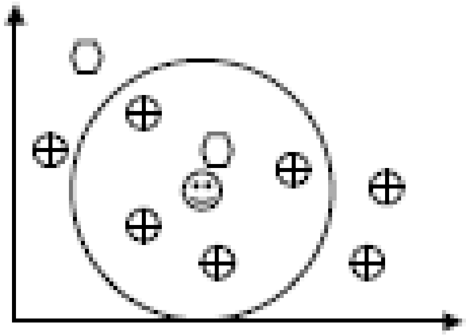
- **Ogólny schemat**:

Krok 1: Poszukaj *k* najbliższych obiektów (sąsiadów) dla x_q

Krok 2: Głosuj wśród *k* najbliższych sąsiadów w celu wyznaczania klasy, do której należy x_q

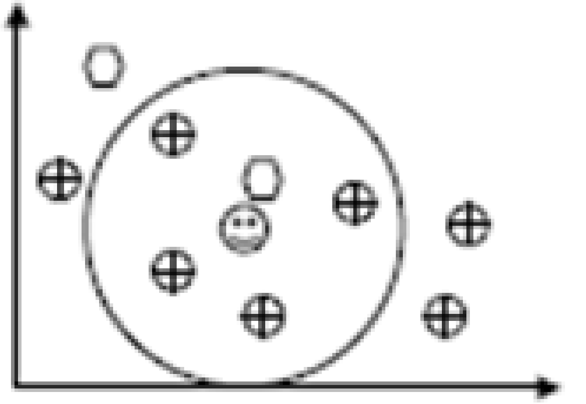
1-NN, decyzja jest ○

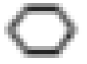
5-NN, decyzja jest ⊕

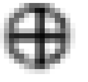


Zaleta: Bardziej odporny na szumy

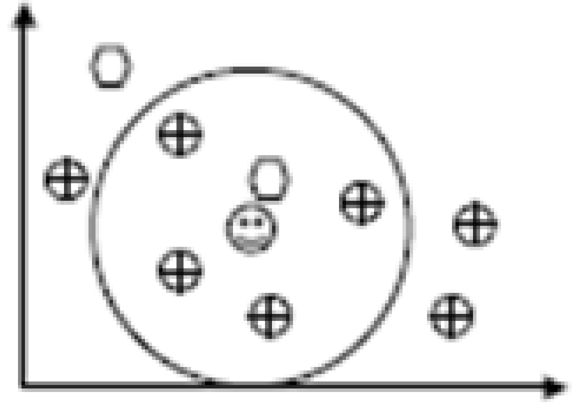
1-NN

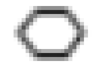


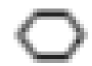
Najbliższy dla naszego obiektu „buźka” jest obiekt 

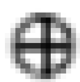
Więc przypiszemy nowemu obiektowi klasę: 

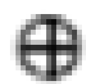
5-NN



Mimo, że najbliższy dla naszego obiektu „bużka” jest obiekt 

Metodą głosowania ustalimy, że skoro mamy wziąć pod uwagę 5 najbliższych sąsiadów tego obiektu, a widać, że 1 z nich ma klasę: 

Zaś 4 pozostałe klasę: 

To przypiszemy nowemu obiektowi klasę: 

	A	B	C
1	**a**	**b**	**klasa**
2	5	5	+
3	7	7	+
4	5	3	+
5	7	3	+
6	3	3	+
7	5	4	+
8	5	2	+
9	3	1	+
10	7	5	+
11	5	1	+
12	8	4	-
13	4	6	-
14	4	6	-
15	10	8	-
16	10	6	-
17	8	5	-
18	7	4	-
19	4	9	-
20	5	5	-
21	4	8	-
22	9	10	-
23	10	7	-
24	6	4	-
25	4	10	-
26	3	6	?

Obiekt klasyfikowany podany jako ostatni : a = 3, b = 6

Teraz obliczmy odległości poszczególnych obiektów od wskazanego. Dla uproszczenia obliczeń posłużymy sie wzorem:

Obliczamy odległość każdego punktu do tego nowo klasyfikowanego

D2 =ZAOKR(PIERWIASTEK(POTĘGA((A2-A26);2)+POTĘGA((B2-B26);2));2)

	a	b	klasa	d
2	5	5	+	2.24
3	7	7	+	4.12
4	5	3	+	3.61
5	7	3	+	5
6	3	3	+	3
7	5	4	+	2.83
8	5	2	+	4.47
9	3	1	+	5
10	7	5	+	4.12
11	5	1	+	5.39
12	8	4	-	5.39
13	4	6	-	1
14	4	6	-	1
15	10	8	-	7.28
16	10	6	-	7
17	8	5	-	5.1
18	7	4	-	4.47
19	4	9	-	3.16
20	5	5	-	2.24
21	4	8	-	2.24
22	9	10	-	7.21
23	10	7	-	7.07
24	6	4	-	3.61
25	4	10	-	4.12
26	3	6	?	?

Funkcja MIN.K Excela

Zwraca k-tą najmniejszą wartość ze zbioru danych. Funkcji tej należy używać do uzyskiwania wartości znajdujących się w określonej względnej pozycji w zbiorze danych.

+ Pokaż wszystko

Składnia

MIN.K(**tablica**;**k**)

Tablica to tablica lub zakres danych numerycznych, dla których należy określić k-tą najmniejszą wartość.

K to pozycja (od najniższej) w tablicy lub w zakresie danych, którą ma zwrócić funkcja.

Spostrzeżenia

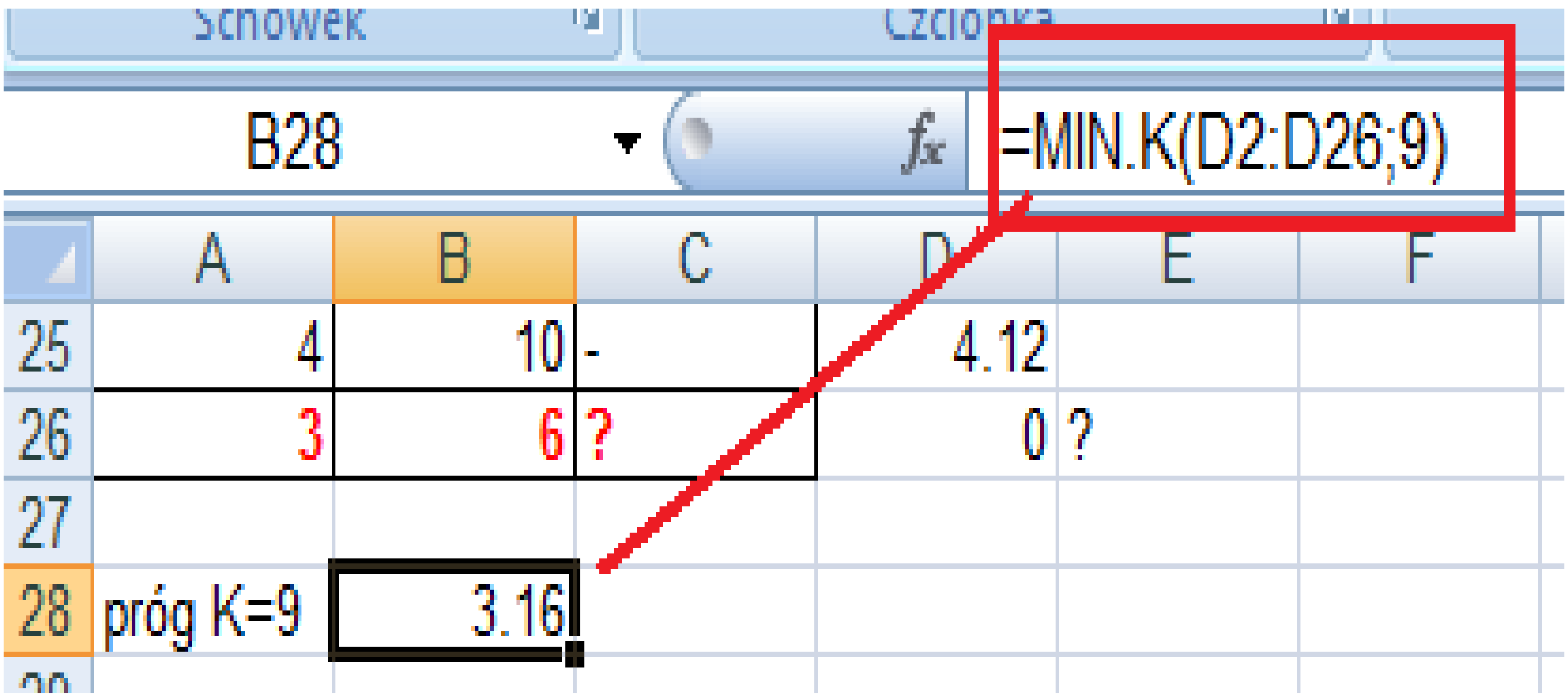
- Jeśli tablica jest pusta, funkcja MIN.K zwraca wartość błędu #LICZBA!.
- Jeśli k ≤ 0 lub jeśli wartość k jest większa niż liczba punktów danych, funkcja MIN.K zwraca wartość błędu #LICZBA!.
- Jeśli n jest liczbą punktów danych w tablicy, funkcja MIN.K(tablica;1) równa jest najmniejszej wartości, a funkcja MIN.K(tablica;n) równa jest największej wartości.

Schowek | Czcionka

B28 | =MIN.K(D2:D25;9)

	A	B	C	D	E	F
1	**a**	**b**	**klasa**	**d**		
2	5	5	+	2,24	+	
3	7	7	+	4,12		
4	5	3	+	3,61	+	
5	7	3	+	5		
6	3	3	+	3	+	
7	5	4	+	2,83	+	
8	5	2	+	4,47		
9	3	1	+	5		
10	7	5	+	4,12		
11	5	1	+	5,39		
12	8	4	-	5,39		
13	4	6	-	1	-	
14	4	6	-	1	-	
15	10	8	-	7,28		
16	10	6	-	7		
17	8	5	-	5,1		
18	7	4	-	4,47		
19	4	9	-	3,16	-	
20	5	5	-	2,24	-	
21	4	8	[illegible]	2,24	-	
22	9	10	-	7,21		
23	10	[illegible]	-	7,07		
24	6	4	-	3,61		
25	4	10	-	4,12		
26	3	6	-	0		
27						
28	próg K=9	3,61				
29	ile "+"	4				
30	ile "-"	5				

Znajdujemy więc k najbliższych sąsiadów. Załóżmy, że szukamy 9 najbliższych sąsiadów. Wyróżnimy ich kolorem zielonym.
Sprawdzamy, które z tych 9 najbliższych sąsiadów są z klasy „+" a które z klasy „-" ?
By to zrobić musimy znaleźć k najbliższych sąsiadów (funkcja Excela o nazwie MIN.K)



Zliczamy + i – jeśli są sąsiadami naszego nowego obiektu „26”

E2 =JEŻELI(D2<=B28;C2;)

	A	B	C	D	E
1	a	b	klasa	d	
2	5	5	+	2,24	+
3	7	7	+	4,12	
4	5	3	+	3,61	+
5	7	3	+	5	
6	3	3	+	3	+
7	5	4	+	2,83	+
8	5	2	+	4,47	
9	3	1	+	5	
10	7	5	+	4,12	
11	5	1	+	5,39	
12	8	4	-	5,39	
13	4	6	-	1	-
14	4	6	-	1	-
15	10	8	-	7,28	
16	10	6	-	7	
17	8	5	-	5,1	
18	7	4	-	4,47	
19	4	9	-	3,16	-
20	5	5	-	2,24	-
21	4	8	-	2,24	-
22	9	10	-	7,21	
23	10	7	-	7,07	
24	6	4	-	3,61	
25	4	10	-	4,12	
26	3	6	-	0	
27					
28	próg K=9	3,61			
29	ile "+"	4			
30	ile "-"	5			

Ostatecznie klasyfikujemy obiekt nowy do tej klasy, która jest bardziej liczna

C26 =JEŻELI(B29>B30;"+";"-")

	A	B	C	D	E
1	a	b	klasa	d	
2	5	5	+	2,24	+
3	7	7	+	4,12	
4	5	3	+	3,61	+
5	7	3	+	5	
6	3	3	+	3	+
7	5	4	+	2,83	+
8	5	2	+	4,47	
9	3	1	+	5	
10	7	5	+	4,12	
11	5	1	+	5,39	
12	8	4	-	5,39	
13	4	6	-	1	-
14	4	6	-	1	-
15	10	8	-	7,28	
16	10	6	-	7	
17	8	5	-	5,1	
18	7	4	-	4,47	
19	4	9	-	3,16	-
20	5	5	-	2,24	-
21	4	8	-	2,24	-
22	9	10	-	7,21	
23	10	7	-	7,07	
24	6	4	-	3,61	
25	4	10	-	4,12	
26	3	6	-	0	
27					
28	próg K=9	3,61			
29	ile "+"	4			
30	ile "-"	5			
31					

A co gdy mamy wiele zmiennych ?

Wyobraźmy sobie, że nie mamy 2 zmiennych opisujących każdy obiekt, ale tych zmiennych jest np. 5: {v1,v2,v3,v4,v5} i że obiekty opisane tymi zmiennymi to 3 punkty: A, B i C:

	V1	V2	V3	V4	V5
A	0.7	0.8	0.4	0.5	0.2
B	0.6	0.8	0.5	0.4	0.2
C	0.8	0.9	0.7	0.8	0.9

Policzmy teraz odległość między punktami:

D (A,B) = pierwiastek $((0.7-0.6)^2 + (0.8-0.8)^2 + (0.4-0.3)^2 + (0.5-0.4)^2 + (0.2-0.2)^2)$ = pierwiastek $(0.01 + 0.01 + 0.01)$ = pierwiastek $(0.03) = 0.17$

D (A,C) = pierwiastek $((0.7-0.8)^2 + (0.8-0.9)^2 + (0.4-0.7)^2 + (0.5-0.8)^2 + (0.2-0.9)^2)$ = pierwiastek $(0.01 + 0.01 + 0.09 + 0.09 + 0.49)$ = pierwiastek $(0.69) = 0.83$

D (B,C) = pierwiastek $((0.6-0.8)^2 + (0.8-0.9)^2 + (0.5-0.7)^2 + (0.4-0.8)^2 + (0.2-0.9)^2)$ = pierwiastek $(0.04 + 0.01 + 0.04+0.16 + 0.49)$ = pierwiastek $(0.74) = 0.86$

Szukamy najmniejszej odległości, bo jeśli te dwa punkty są najbliżej siebie, dla których mamy najmniejszą odległości ! A więc najmniejsza odległość jest między punktami A i B !

K-NN

Schemat algorytmu:

- Poszukaj obiektu najbliższego w stosunku do obiektu klasyfikowanego.
- Określenie klasy decyzyjnej na podstawie obiektu najbliższego.

Cechy algorytmu:

- Bardziej odporny na szumy - w poprzednim algorytmie obiekt najbliższy klasyfikowanemu może być zniekształcony - tak samo zostanie zaklasyfikowany nowy obiekt.
- Konieczność ustalenia liczby najbliższych sąsiadów.
- Wyznaczenie miary podobieństwa wśród obiektów (wiele miar podobieństwa).
- Dobór parametru k - liczby sąsiadów:
- Jeśli k jest małe, algorytm nie jest odporny na szumy – jakość klasyfikacji jest niska. Jeśli k jest duże, czas działania algorytmu rośnie - większa złożoność obliczeniowa. Należy wybrać k, które daje najwyższą wartość klasyfikacji.