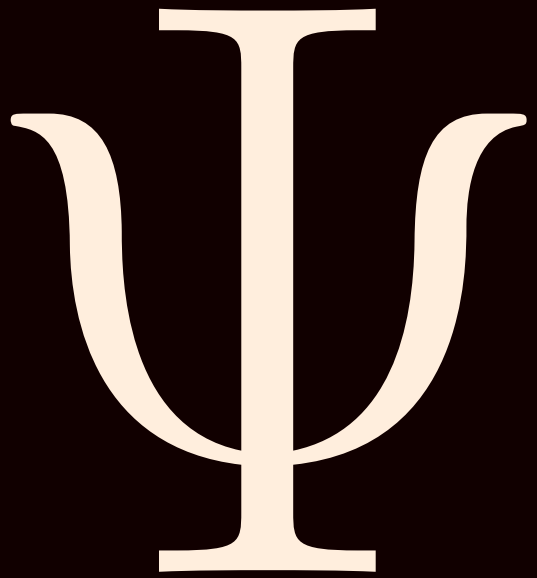
Metody Sztucznej Inteligencji

Wykład 2: **Zadanie klasyfikacji**

dr inż. Paweł Ksieniewicz
Katedra Systemów i Sieci Komputerowych

13 marca 2019



Uczenie nadzorowane

Uczenie nadzorowane

↬ Regresja:

Uczenie nadzorowane

- Regresja:
  - predykcja **liczby rzeczywistej** powiązanej z zadanym wzorcem (wektorem cech).

Uczenie nadzorowane

- Regresja:
  - predykcja **liczby rzeczywistej** powiązanej z zadanym wzorcem (wektorem cech).
- Klasyfikacja:

Uczenie nadzorowane

- Regresja:
  - predykcja **liczby rzeczywistej** powiązanej z zadanym wzorcem (wektorem cech).
- Klasyfikacja:
  - predykcja **liczby całkowitej** (etykiety) powiązanej z zadanym wzorcem.

Przykład

ZWIERZĘ	*jajorodne?*	*ma łuski?*	*jadowite?*	*zmiennoceplne?*	*ile nóg?*	*czy gad?*
kobra	1	1	1	1	0	1
grzechotnik	1	1	1	1	0	1
boa dusiciel	0	1	0	1	0	1
kurczak	1	1	0	1	2	0
gupik	0	1	0	0	0	0
Andrzej	0	0	0	0	2	0
zebra	0	0	0	0	4	0
pyton	1	1	0	1	0	1
aligator	1	1	0	1	4	1

	kobra	*grzechotnik*	*boa dusiciel*	*kura*	*gupik*	*Andrzej*	*zebra*	*pyton*	*aligator*
kobra	—	0.00_{1}	3.14_{4}	2.70_{3}	3.79_{6}	4.65_{7}	5.11_{8}	2.41_{2}	3.43_{5}
grzechotnik	0.00_{1}	—	3.14_{4}	2.70_{3}	3.79_{6}	4.65_{7}	5.11_{8}	2.41_{2}	3.43_{5}
boa dusiciel	3.14_{4}	3.14_{5}	—	2.36_{3}	2.12_{2}	3.43_{7}	4.04_{8}	2.01_{1}	3.17_{6}
kurczak	2.70_{4}	2.70_{5}	2.36_{3}	—	3.17_{6}	3.79_{7}	3.98_{8}	1.22_{1}	1.22_{2}
gupik	3.79_{6}	3.79_{7}	2.12_{1}	3.17_{4}	—	2.70_{2}	3.43_{5}	2.92_{3}	3.81_{8}
Andrzej	4.65_{7}	4.65_{8}	3.43_{3}	3.79_{4}	2.70_{2}	—	1.22_{1}	3.98_{5}	3.98_{6}
zebra	5.11_{7}	5.11_{8}	4.04_{5}	3.98_{4}	3.43_{2}	1.22_{1}	—	4.51_{6}	3.79_{3}
pyton	2.41_{3}	2.41_{4}	2.01_{2}	1.22_{1}	2.92_{6}	3.98_{7}	4.51_{8}	—	2.45_{5}
aligator	3.43_{4}	3.43_{5}	3.17_{3}	1.22_{1}	3.81_{7}	3.98_{8}	3.79_{6}	2.45_{2}	—
czy gad?	1	1	1	0	0	0	0	1	1

Podejście minimalno-odległościowe

Podejście minimalno-odległościowe

- Najbardziej banalnym rozwiązaniem jest **najbliższy sąsiad**.

Podejście minimalno-odległościowe

- Najbardziej banalnym rozwiązaniem jest **najbliższy sąsiad**.
- Jest to tak zwane **uczenie leniwe**.

Podejście minimalno-odległościowe

- Najbardziej banalnym rozwiązaniem jest **najbliższy sąsiad**.
- Jest to tak zwane **uczenie leniwe**.
- Uczenie polega wprost na zapamiętaniu wszystkich przypadków (wzorców, obiektów) ze zbioru uczącego.

Podejście minimalno-odległościowe

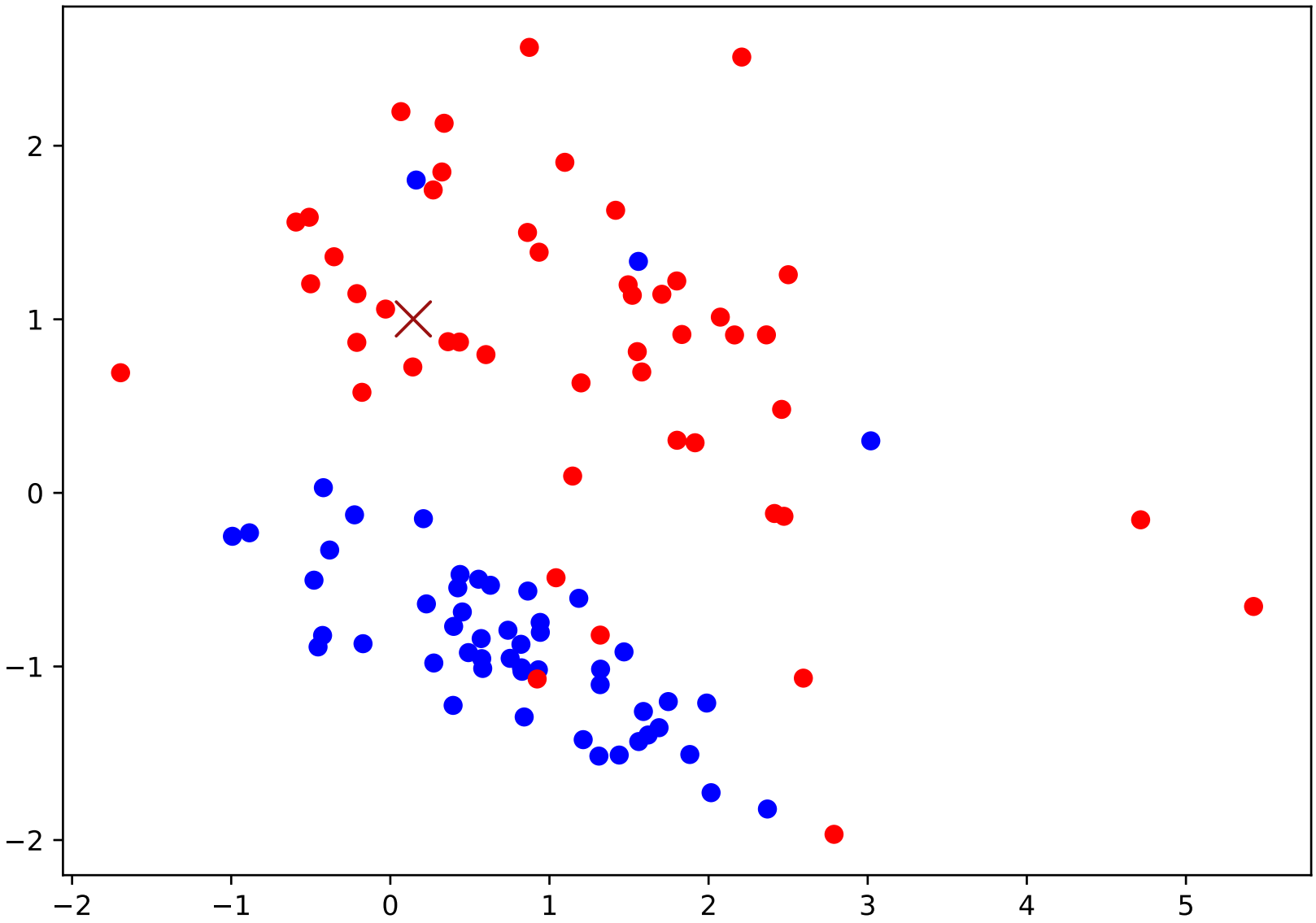
- Najbardziej banalnym rozwiązaniem jest **najbliższy sąsiad**.
- Jest to tak zwane **uczenie leniwe**.
- Uczenie polega wprost na zapamiętaniu wszystkich przypadków (wzorców, obiektów) ze zbioru uczącego.
- Predykcja składa się z całych dwóch kroków:

Podejście minimalno-odległościowe

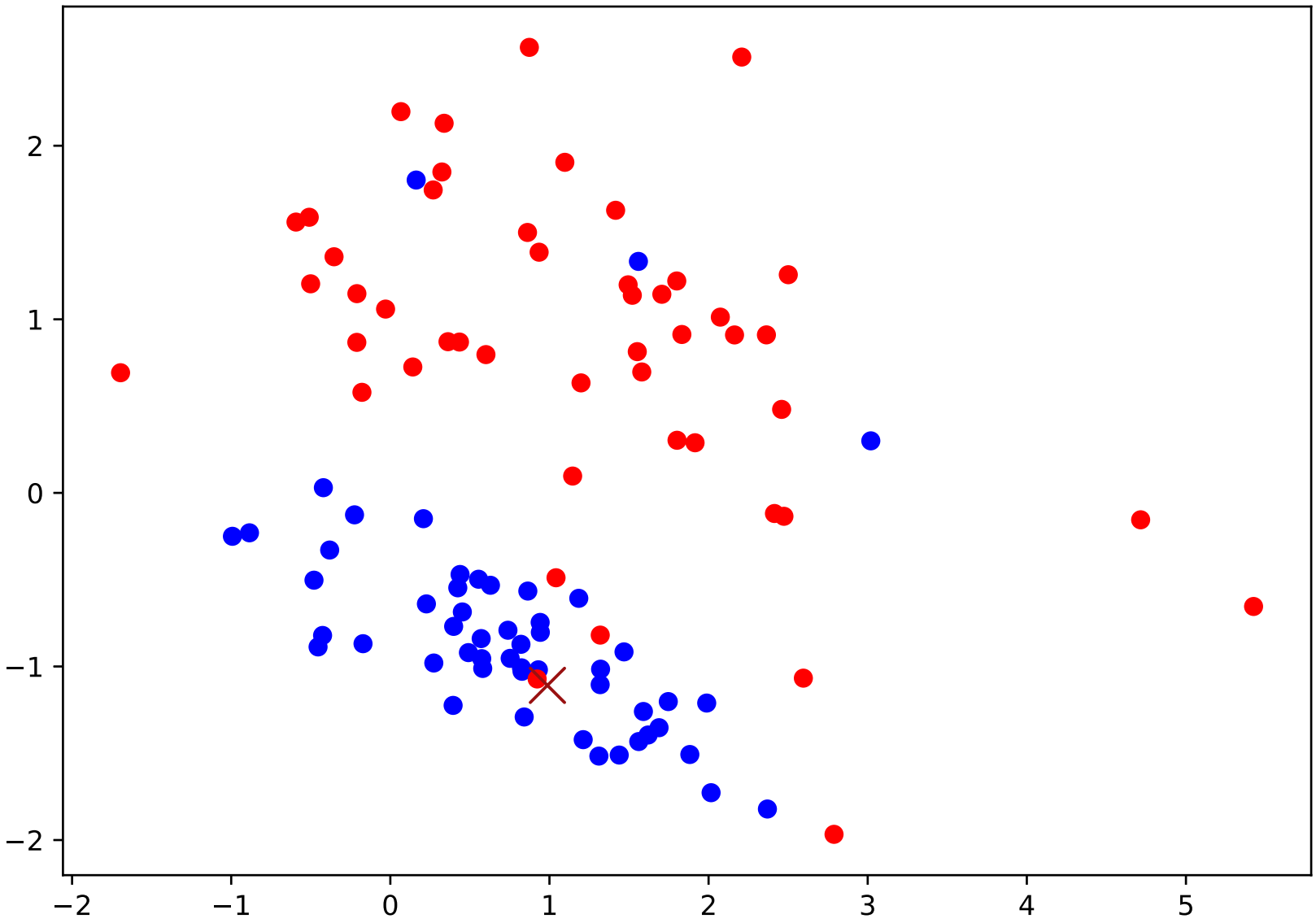
- Najbardziej banalnym rozwiązaniem jest **najbliższy sąsiad**.
- Jest to tak zwane **uczenie leniwe**.
- Uczenie polega wprost na zapamiętaniu wszystkich przypadków (wzorców, obiektów) ze zbioru uczącego.
- Predykcja składa się z całych dwóch kroków:
  - znajdź najbliższego sąsiada,

Podejście minimalno-odległościowe

- Najbardziej banalnym rozwiązaniem jest **najbliższy sąsiad**.
- Jest to tak zwane **uczenie leniwe**.
- Uczenie polega wprost na zapamiętaniu wszystkich przypadków (wzorców, obiektów) ze zbioru uczącego.
- Predykcja składa się z całych dwóch kroków:
  - znajdź najbliższego sąsiada,
  - zwróć etykietę odnalezionego najbliższego sąsiada.


2
1
0
−1
−2
−2
−1
0
1
2
3
4
5

	kobra	*grzechotnik*	*boa dusiciel*	*kura*	*gupik*	*Andrzej*	*zebra*	*pyton*	*aligator*
kobra	—	0.00_1	3.14_4	2.70_3	3.79_6	4.65_7	5.11_8	2.41_2	3.43_5
grzechotnik	0.00_1	—	3.14_4	2.70_3	3.79_6	4.65_7	5.11_8	2.41_2	3.43_5
boa dusiciel	3.14_4	3.14_5	—	2.36_3	2.12_2	3.43_7	4.04_8	2.01_1	3.17_6
kurczak	2.70_4	2.70_5	2.36_3	—	3.17_6	3.79_7	3.98_8	1.22_1	1.22_2
gupik	3.79_6	3.79_7	2.12_1	3.17_4	—	2.70_2	3.43_5	2.92_3	3.81_8
Andrzej	4.65_7	4.65_8	3.43_3	3.79_4	2.70_2	—	1.22_1	3.98_5	3.98_6
zebra	5.11_7	5.11_8	4.04_5	3.98_4	3.43_2	1.22_1			
pyton	2.41_3	2.41_4	2.01_2	1.22_1	2.92_6	3.98_7			
aligator	3.43_4	3.43_5	3.17_3	1.22_1	3.81_7	3.98_8			
czy gad?	1	1	1	0	0	0	0	1	1


2
1
0
−1
−2
−2
−1
0
1
2
3
4
5

K najbliższych sąsiadów — kNN

K najbliższych sąsiadów — kNN

↬ Jest to nadal **uczenie leniwe**.

K najbliższych sąsiadów — kNN

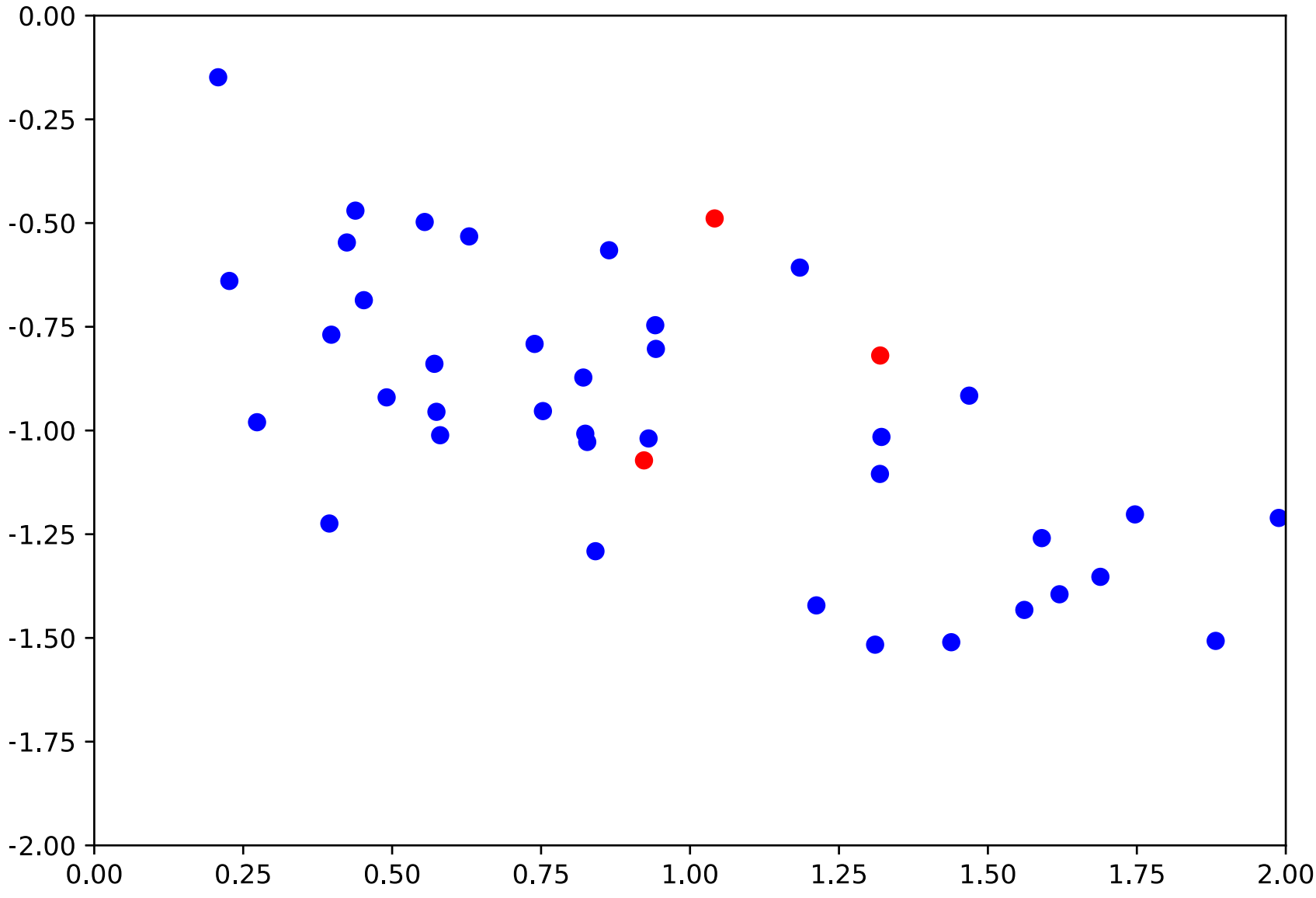
- Jest to nadal **uczenie leniwe**.
- Predykcja nadal składa się z całych dwóch kroków:

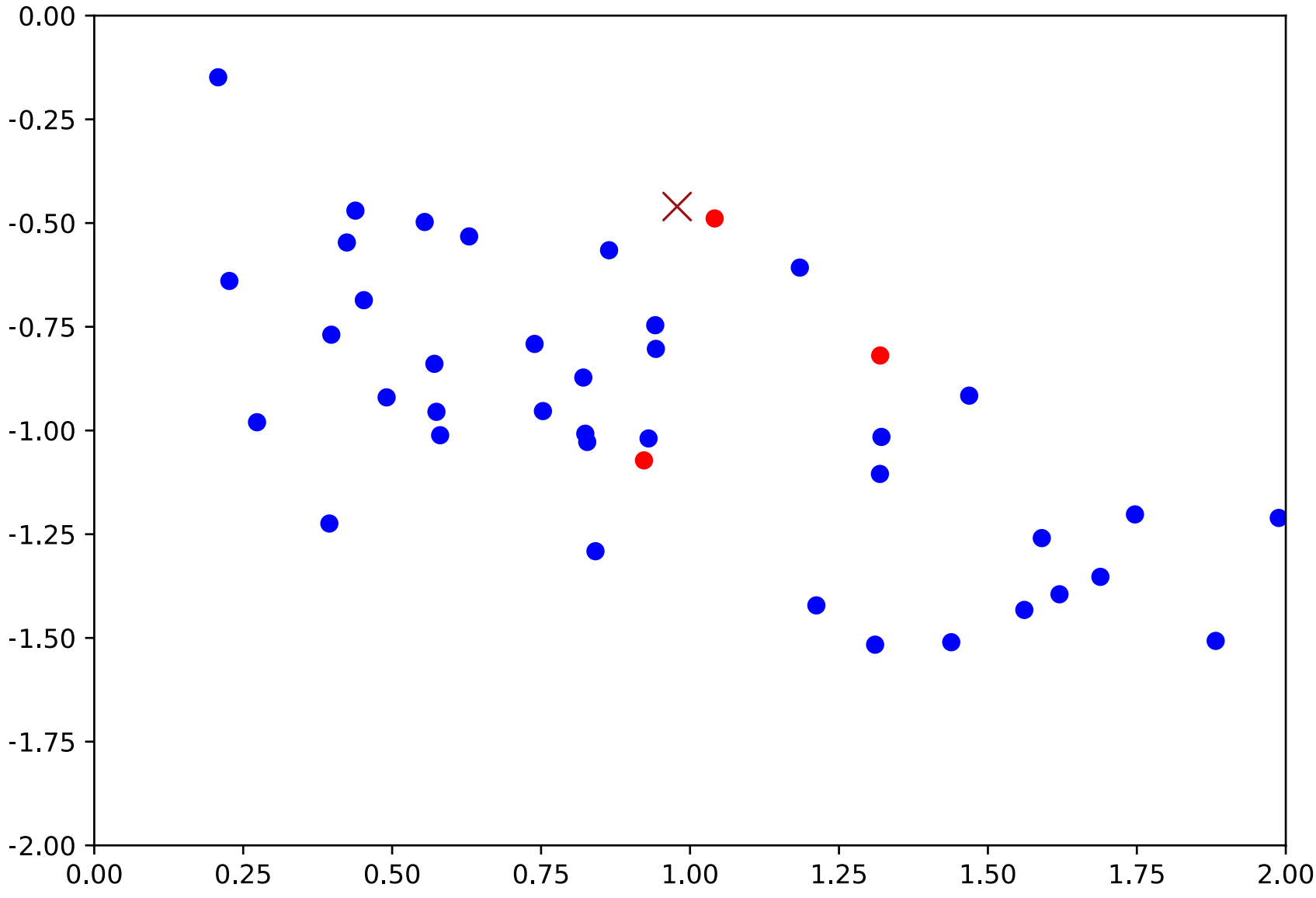
K najbliższych sąsiadów — kNN

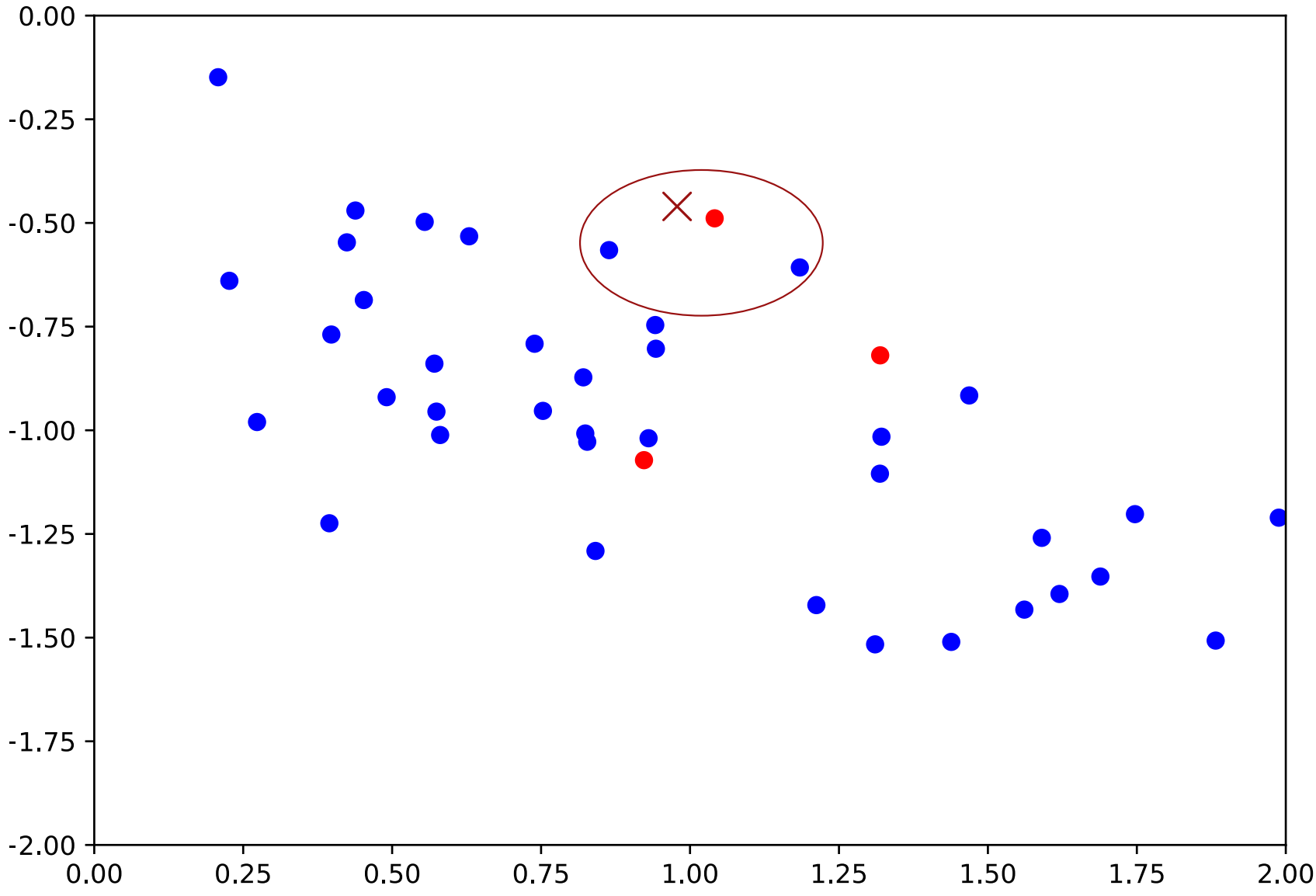
- Jest to nadal **uczenie leniwe**.
- Predykcja nadal składa się z całych dwóch kroków:
  - znajdź k najbliższych sąsiadów,

K najbliższych sąsiadów — kNN

- Jest to nadal **uczenie leniwe**.
- Predykcja nadal składa się z całych dwóch kroków:
  - znajdź k najbliższych sąsiadów,
  - zwróć najczęściej pojawiającą się etykietę wśród odnalezionych najbliższych sąsiadów.


0.00
-0.25
-0.50
-0.75
-1.00
-1.25
-1.50
-1.75
-2.00
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00


0.00
-0.25
-0.50
-0.75
-1.00
-1.25
-1.50
-1.75
-2.00
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00


0.00
-0.25
-0.50
-0.75
-1.00
-1.25
-1.50
-1.75
-2.00
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

	kobra	*grzechotnik*	*boa dusiciel*	*kura*	*gupik*	*Andrzej*	*zebra*	*pyton*	*aligator*
kobra	—	0.00_1	3.14_4	2.70_3	3.79_6	4.65_7	5.11_8	2.41_2	3.43_5
grzechotnik	0.00_1	—	3.14_4	2.70_3	3.79_6	4.65_7	5.11_8	2.41_2	3.43_5
boa dusiciel	3.14_4	3.14_5	—	2.36_3	2.12_2	3.43_7	4.04_8	2.01_1	3.17_6
kurczak	2.70_4	2.70_5	2.36_3	—	3.17_6	3.79_7	3.98_8	1.22_1	1.22_2
gupik	3.79_6	3.79_7	2.12_1	3.17_4	—	2.70_2	3.43_5	2.92_3	3.81_8
Andrzej	4.65_7	4.65_8	3.43_3	3.79_4	2.70_2	—	1.22_1	3.98_5	3.98_6
zebra	5.11_7	5.11_8	4.04_5	3.98_4	3.43_2	1.22_1			
pyton	2.41_3	2.41_4	2.01_2	1.22_1	2.92_6	3.98_7			
aligator	3.43_4	3.43_5	3.17_3	1.22_1	3.81_7	3.98_8			
czy gad?	1	1	1	0	0	0	0	1	1

Ile powinno wynosić k?

Ile powinno wynosić k?

- Powinno być **nieparzyste**, dla ewidentnego wyniku głosowania.

Ile powinno wynosić k?

- Powinno być **nieparzyste**, dla ewidentnego wyniku głosowania.
- Im **mniejsze** k, tym mniejsza próbka, a więc większy wpływ obserwacji odstających.

Ile powinno wynosić k?

- Powinno być **nieparzyste**, dla ewidentnego wyniku głosowania.
- Im **mniejsze** k, tym mniejsza próbka, a więc większy wpływ obserwacji odstających.
- Im **większe** k, tym bardziej upraszczamy klasyfikację do wyboru **klasy dominującej według prawdopodobieństwa a priori**.

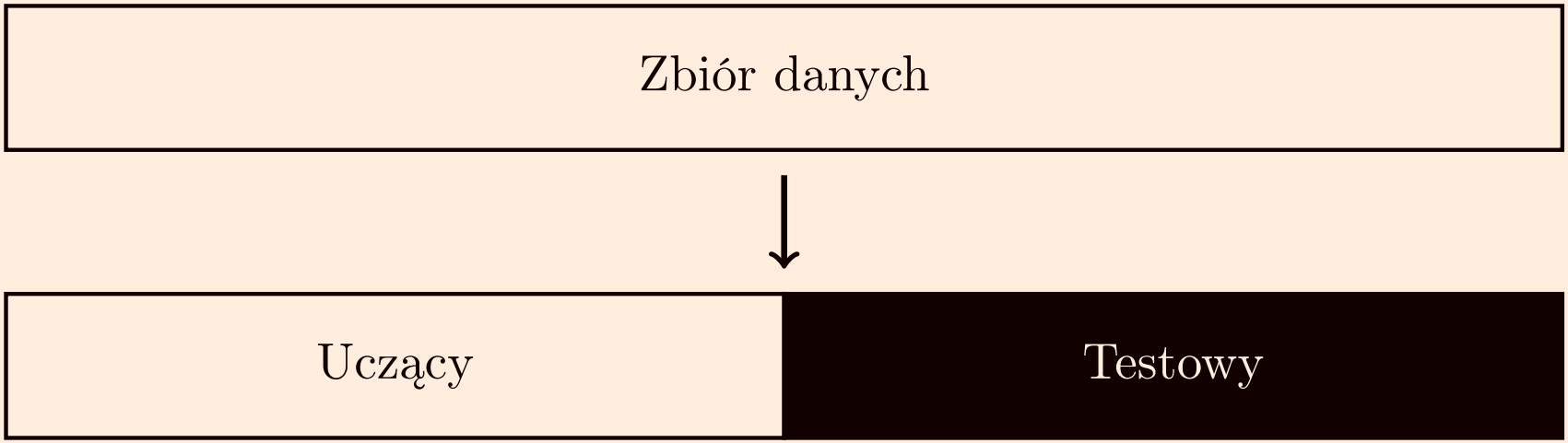
Ile powinno wynosić k?

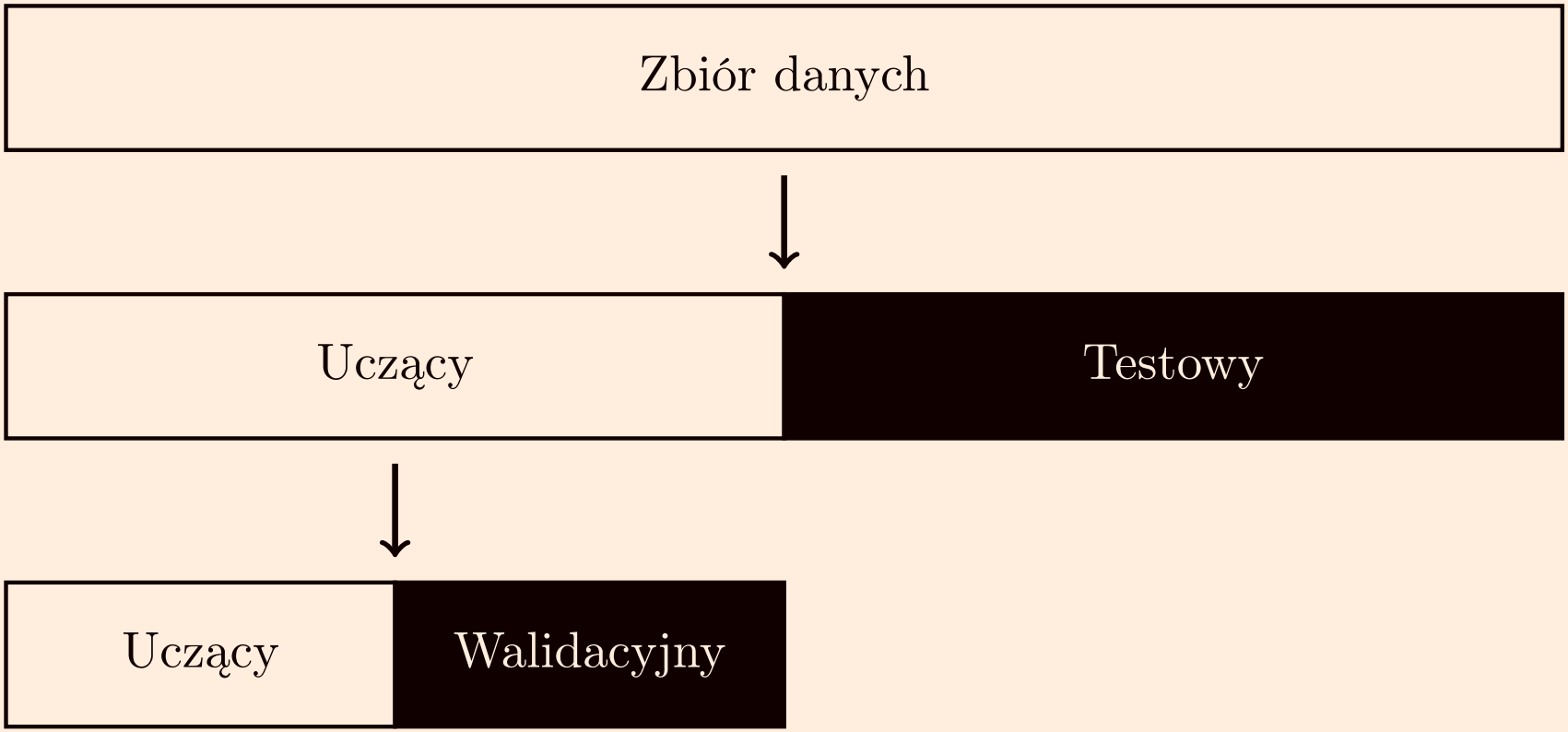
- Powinno być **nieparzyste**, dla ewidentnego wyniku głosowania.
- Im **mniejsze** k, tym mniejsza próbka, a więc większy wpływ obserwacji odstających.
- Im **większe** k, tym bardziej upraszczamy klasyfikację do wyboru **klasy dominującej według prawdopodobieństwa a priori**.
- W większości pakietów **domyślnie wynosi 5**.

Ile powinno wynosić k?

- Powinno być **nieparzyste**, dla ewidentnego wyniku głosowania.
- Im **mniejsze** k, tym mniejsza próbka, a więc większy wpływ obserwacji odstających.
- Im **większe** k, tym bardziej upraszczamy klasyfikację do wyboru **klasy dominującej według prawdopodobieństwa a priori**.
- W większości pakietów **domyślnie wynosi 5**.
- Najlepszy byłby dobór eksperymentalny przez **zbiór walidujący**.

Zbiór danych


Zbiór danych
Uczący
Testowy


Zbiór danych
Uczący
Testowy
Uczący
Walidacyjny

Wady i zalety kNN

Wady i zalety kNN

Uczenie jest nieprzyzwoicie wręcz szybkie. *(bo go nie ma)*

Wady i zalety kNN

- 👍 Uczenie jest nieprzyzwoicie wręcz szybkie. *(bo go nie ma)*
- 👍 Nie wymaga żadnej matematyki do wyjaśnienia. *(o ile nie skupiamy się na definicji odległości)*

Wady i zalety kNN

- 👍 Uczenie jest nieprzyzwoicie wręcz szybkie. *(bo go nie ma)*
- 👍 Nie wymaga żadnej matematyki do wyjaśnienia. *(o ile nie skupiamy się na definicji odległości)*
- 👎 Jest szalenie nieefektywny pamięciowo. *(musi zapisać w pamięci wszystko)*

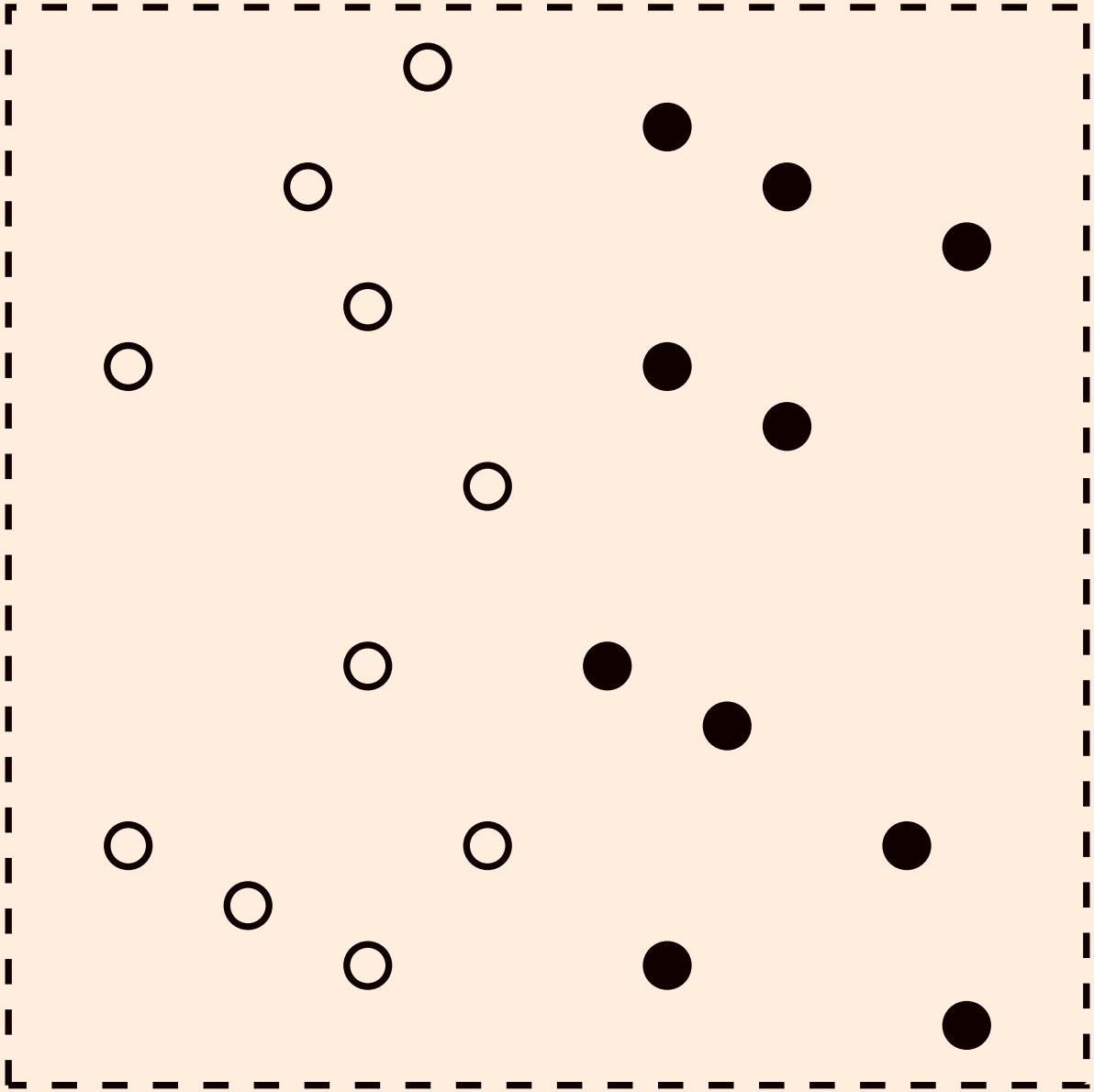
Wady i zalety kNN

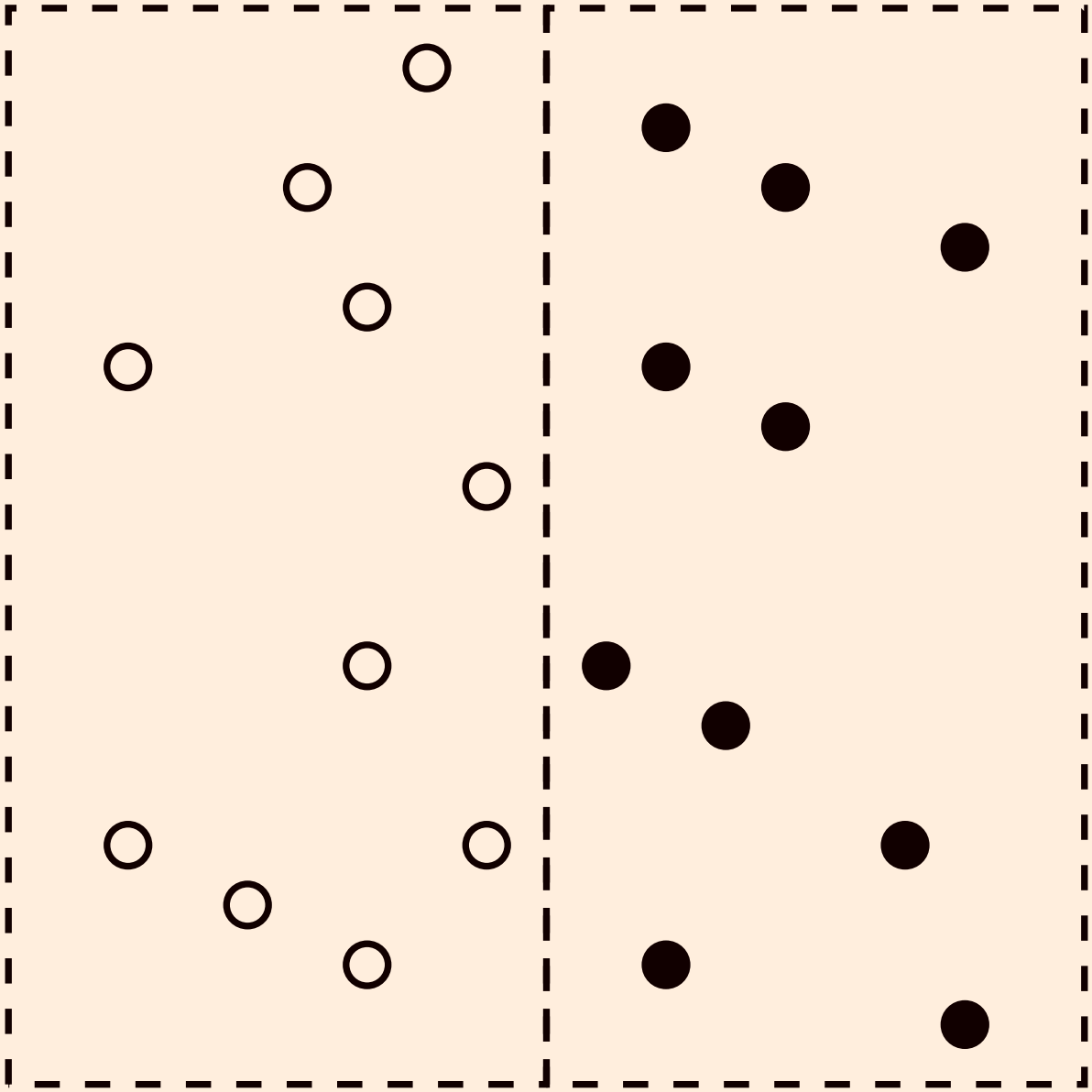
- 👍 Uczenie jest nieprzyzwoicie wręcz szybkie. *(bo go nie ma)*
- 👍 Nie wymaga żadnej matematyki do wyjaśnienia. *(o ile nie skupiamy się na definicji odległości)*
- 👎 Jest szalenie nieefektywny pamięciowo. *(musi zapisać w pamięci wszystko)*
- 👎 Jest przygnębiająco nieefektywny obliczeniowo. *(im większe k i im większa liczba wzorców, tym dłużej szuka)*

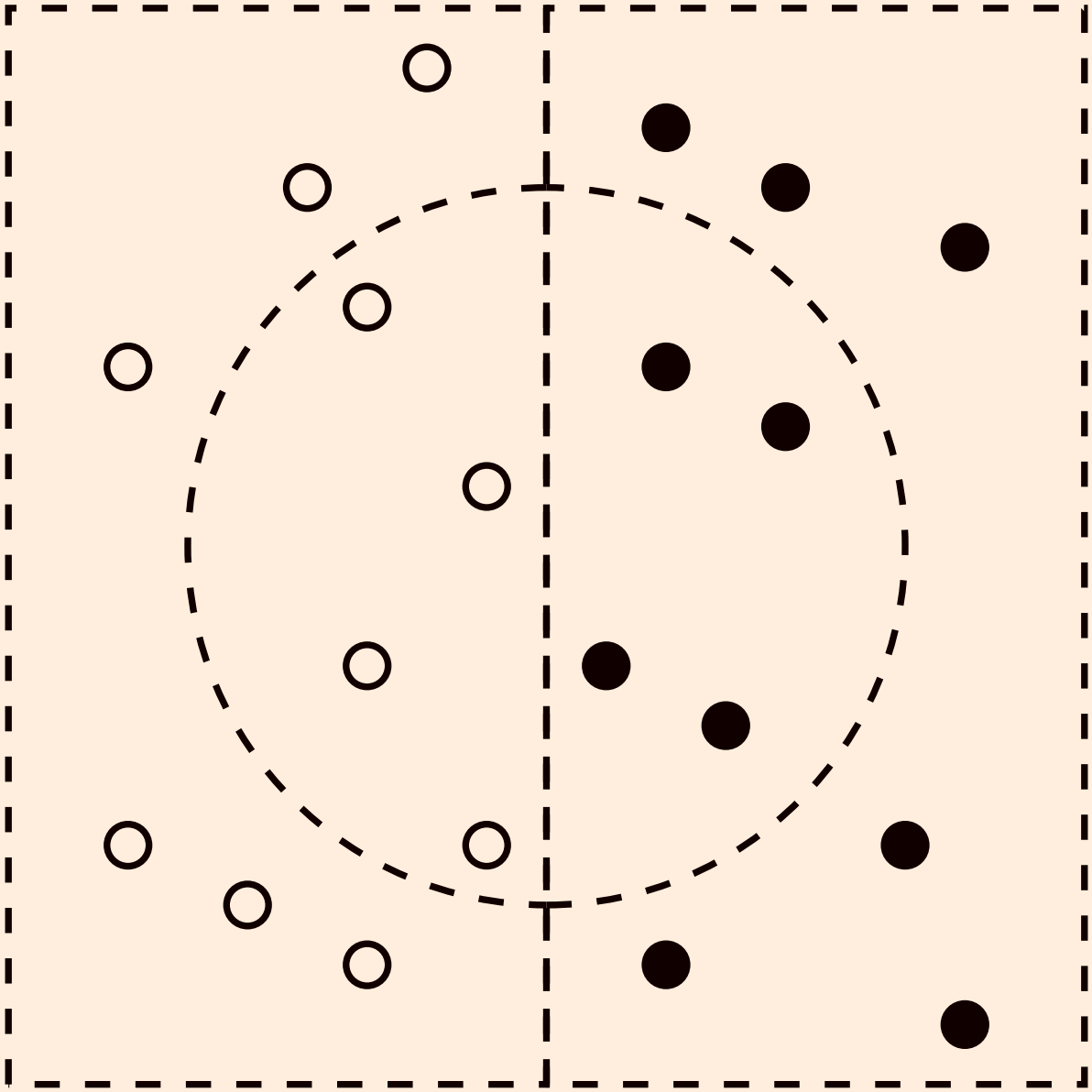
Wady i zalety kNN

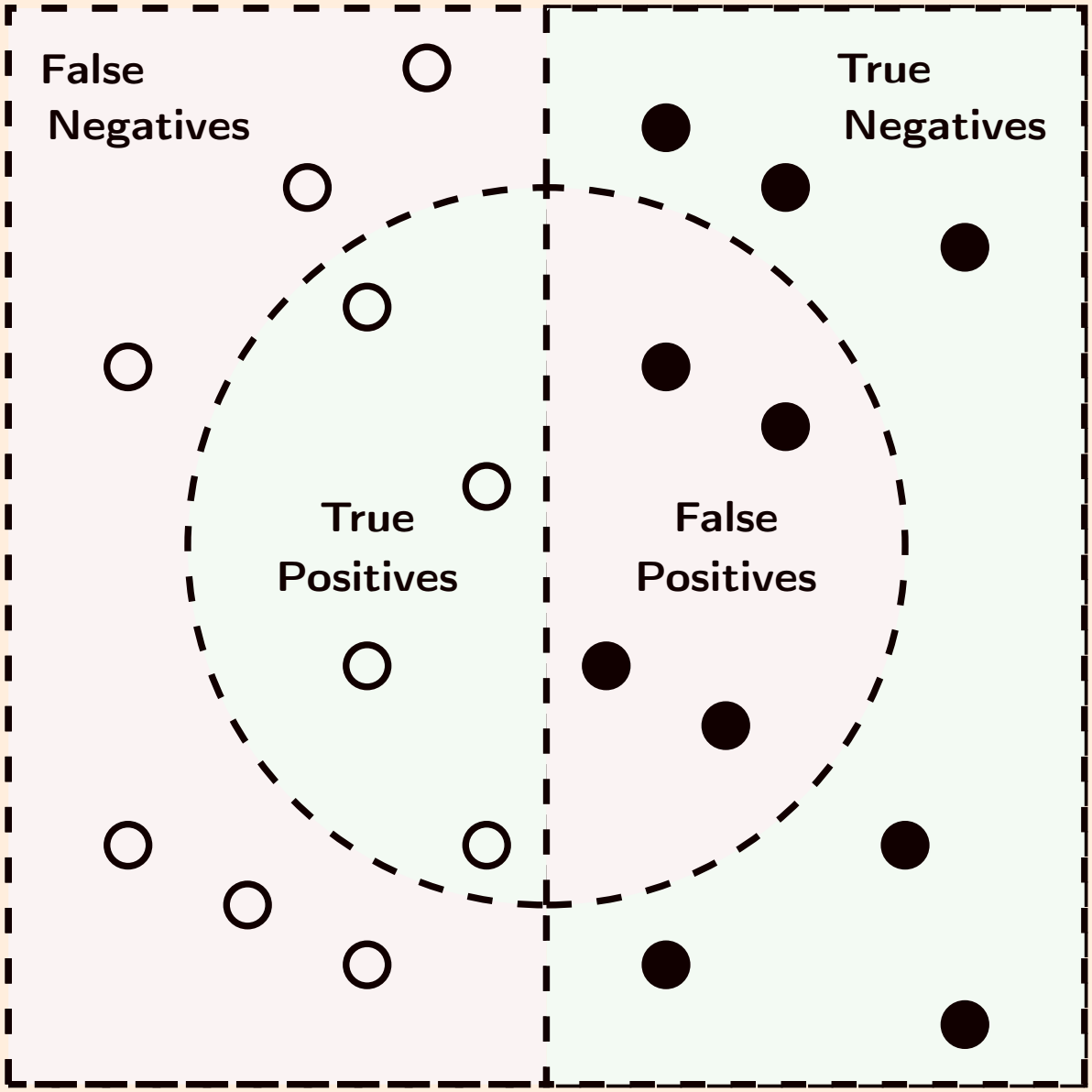
- 👍 Uczenie jest nieprzyzwoicie wręcz szybkie. *(bo go nie ma)*
- 👍 Nie wymaga żadnej matematyki do wyjaśnienia. *(o ile nie skupiamy się na definicji odległości)*
- 👎 Jest szalenie nieefektywny pamięciowo. *(musi zapisać w pamięci wszystko)*
- 👎 Jest przygnębiająco nieefektywny obliczeniowo. *(im większe k i im większa liczba wzorców, tym dłużej szuka)*
- 👎 NIE MÓWI NAM NIC O PROBLEMIE. *(nie budujemy w nim żadnego modelu)*

Miary oceny jakości








False
Negatives
True
Negatives
True
Positives
False
Positives

Macierz pomyłek dla problemu binarnego

	Wynik predykcji		
Etykieta	Prawda	Fałsz	Suma
Prawda	TP	FN	P
Fałsz	FP	TN	N
Suma	P'	N'	

Błąd i dokładność

$$ERR = \frac{FN + FP}{TP + FN + FP + TN}$$

$$ACC = 1 - ERR$$

Czy dokładność to dobra miara?

↬ Nie znaczy przesadnie wiele, jeśli mamy do czynienia z **problemem niezbalansowanym**.

Czy dokładność to dobra miara?

- Nie znaczy przesadnie wiele, jeśli mamy do czynienia z **problemem niezbalansowanym**.
- **Najczęściej mamy do czynienia z problemem niezbalansowanym.**

Czy dokładność to dobra miara?

- Nie znaczy przesadnie wiele, jeśli mamy do czynienia z **problemem niezbalansowanym**.
- **Najczęściej mamy do czynienia z problemem niezbalansowanym.**

Jeśli klasyfikujemy bardzo rzadką chorobę (załóżmy, że występuje u jednego człowieka na milion) i ograniczymy nasz klasyfikator do informowania każdego pacjenta o tym że jest zdrowy, niezależnie od wyników badań, otrzymamy dokładność powyżej 99%.

Czy dokładność to dobra miara?

- Nie znaczy przesadnie wiele, jeśli mamy do czynienia z **problemem niezbalansowanym**.
- **Najczęściej mamy do czynienia z problemem niezbalansowanym.**

Jeśli klasyfikujemy bardzo rzadką chorobę (załóżmy, że występuje u jednego człowieka na milion) i ograniczymy nasz klasyfikator do informowania każdego pacjenta o tym że jest zdrowy, niezależnie od wyników badań, otrzymamy dokładność powyżej 99%.

Wszyscy chorzy umrą w niewiedzy i mękach.

Precision i recall

Precision i recall

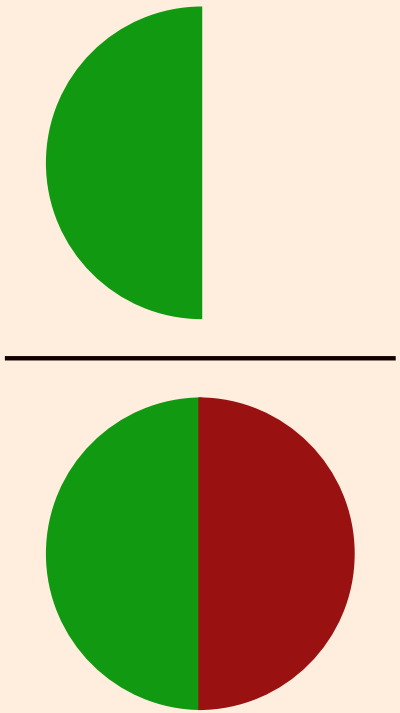
Precision

Ile z **ocenionych pozytywnie wzorców** zostało ocenionych słusznie.

Precision i recall

Precision

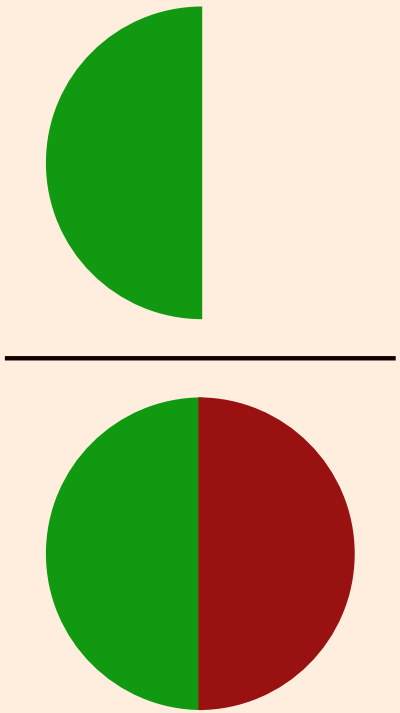
Ile z **ocenionych pozytywnie wzorców** zostało ocenionych słusznie.



Precision i recall

Precision

Ile z **ocenionych pozytywnie wzorców** zostało ocenionych słusznie.



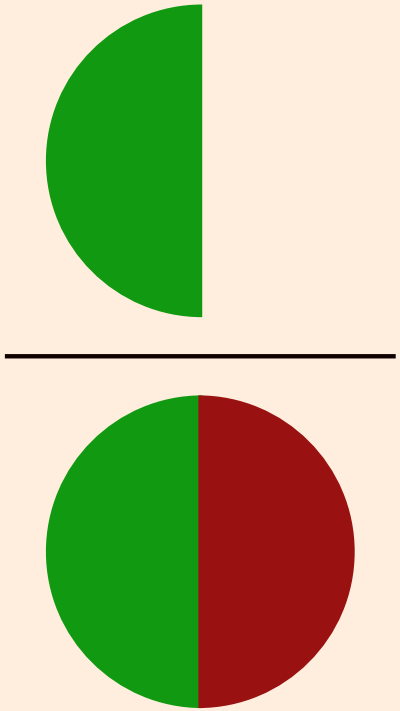
Recall

Ile z **pozytywnych wzorców** zostało ocenionych słusznie.

Precision i recall

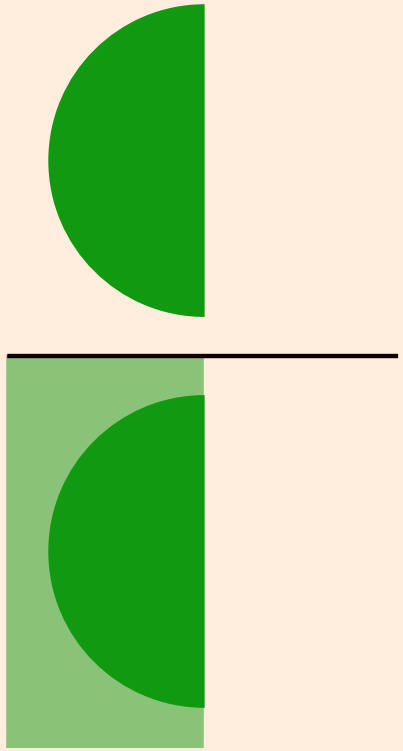
Precision

Ile z **ocenionych pozytywnie wzorców** zostało ocenionych słusznie.



Recall

Ile z **pozytywnych wzorców** zostało ocenionych słusznie.



F-score

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$