

3 Klasyfikacja

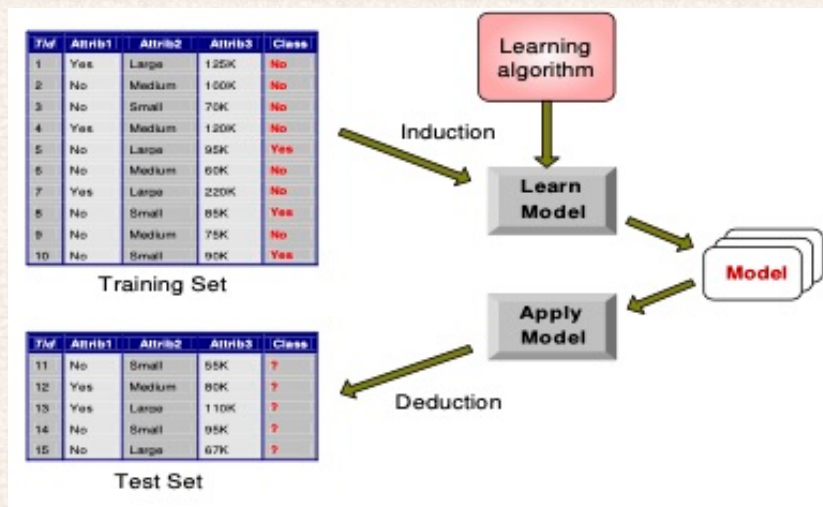
Problem klasyfikacji

- Mając bazę danych $D = \{t_1, t_2, \dots, t_n\}$ i skończony zbiór klas $C = \{C_1, \dots, C_m\}$, problem klasyfikacji polega na zdefiniowaniu odwzorowania $f: D \rightarrow C$, gdzie każdy t_i jest przypisany do jednej klasy.
- W rzeczywistości f dzieli D na klasy równoważności.
- Prognozowanie jest podobnym procesem, ale można je postrzegać jako posiadające nieskończoną liczbę klas.

Definicja klasyfikacji

- Dany jest zbiór rekordów (zbiór uczący)
 - Każdy rekord zawiera zestaw atrybutów, jeden z nich jest klasą.
- Znajdź model dla atrybutu klasy jako funkcji wartości innych atrybutów.
- Cel: wcześniej niewidziane rekordy powinny być przypisane do klasy tak dokładnie, jak to możliwe.
 - Zestaw testowy służy do określenia dokładności modelu. Zwykle dany zestaw danych jest podzielony na zbiory uczące i testowe, przy czym zbiór uczący służy do budowy modelu, a zestaw testowy służy do jego walidacji.

Ilustracja zadania klasyfikującego



Przykłady zadań klasyfikujących

- Przewidywanie komórek nowotworowych jako łagodnych lub złośliwych
- Klasyfikowanie transakcji kartą kredytową jako legalnych lub oszukańczych
- Klasyfikowanie drugorzędowych struktur białek jako alfa-helisy, arkusza beta lub zwoju nie-uporządkowanego
- Kategoryzowanie wiadomości, takich jak finanse, pogoda, rozrywka, sport itp

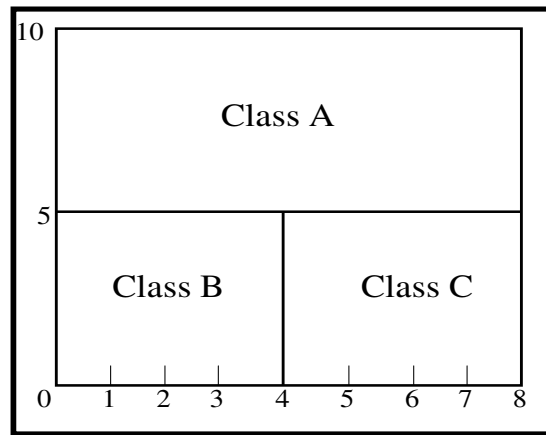
Techniki klasyfikacji

Podejście:

- Stwórz określony model, oceniając dane szkoleniowe (lub korzystając z wiedzy ekspertów dziedzinowych).
- Zastosuj opracowany model do nowych danych. Klasy muszą być wstępnie zdefiniowane.

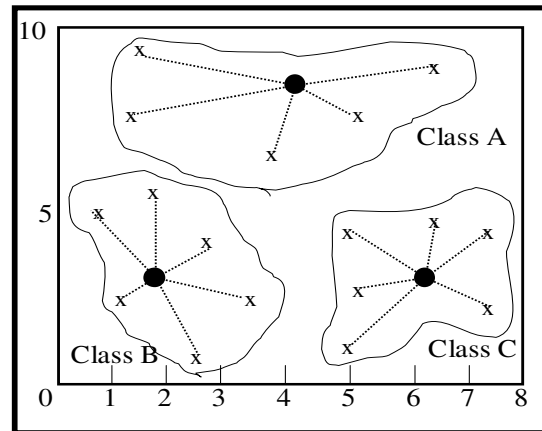
Najpopularniejsze techniki wykorzystują drzewa decyzyjne (DT), sieci neuronowe (NN) lub są oparte na odległościach lub metodach statystycznych.

Definiowanie klas



Partitioning Based

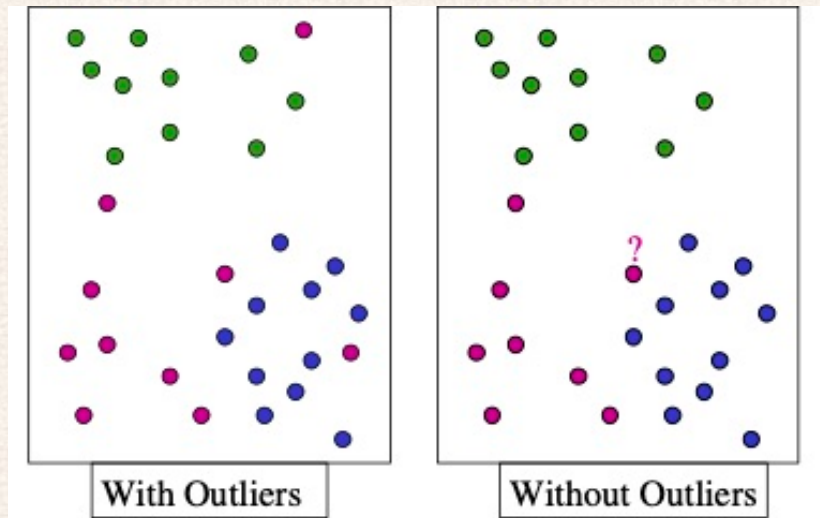
Distance Based



Problemy w klasyfikacji

- Brakujące dane
 - Ignoruj
 - Zastąp wartością przyjętą
- Obsługa wartości odstających w danych szkoleniowych
- Mierzenie wydajności
 - Dokładność klasyfikacji na danych testowych
 - Macierz nieporozumień

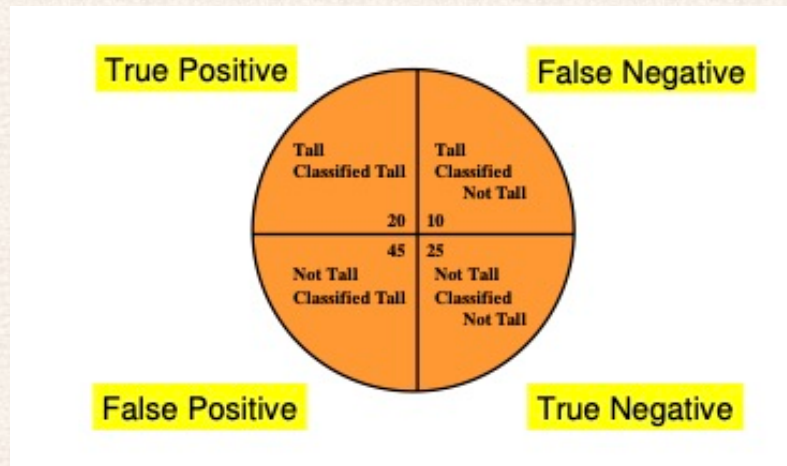
Obsługa wartości odstających



Przykład: wzrost w grupie studentów

Name	Gender	Height	Output1	Output2
Kristina	F	1.60	Short	Medium
Jim	M	2.02	Tall	Medium
Maggie	F	1.90	Medium	Tall
Martha	F	1.88	Medium	Tall
Stephanie	F	1.71	Short	Medium
Bob	M	1.85	Medium	Medium
Kathy	F	1.60	Short	Medium
Dave	M	1.72	Short	Medium
Worth	M	2.12	Tall	Tall
Steven	M	2.10	Tall	Tall
Debbie	F	1.78	Medium	Medium
Todd	M	1.95	Medium	Medium
Kim	F	1.89	Medium	Tall
Amy	F	1.81	Medium	Medium
Wynette	F	1.75	Medium	Medium

Wydajność klasyfikacji



Przykład macierzy nieporozumień

Actual Membership	Assignment		
	Short	Medium	Tall
Short	0	4	0
Medium	0	5	3
Tall	0	1	2

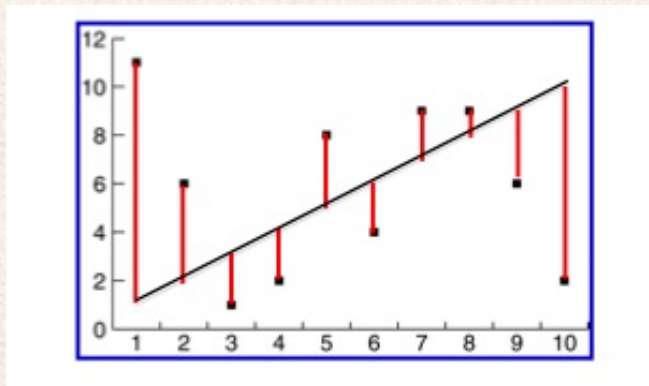
Regresja

- Załóżmy, że dane pasują do predefiniowanej funkcji
- Określ najlepsze wartości współczynników regresji $c_0, c_1, \dots, c_n$.
- Załóżmy błąd: $y = c_0 + c_1x_1 + \dots + c_nx_n + \varepsilon$
- Oszacuj błąd przy użyciu średniokwadratowego błędu treningu:

$$y_i = c_0 + c_1x_{1i} + \varepsilon_i, i = 1, \dots, k$$

$$L = \sum_{i=1}^k \varepsilon_i^2 = \sum_{i=1}^k (y_i - c_0 - c_1x_{1i})^2$$

Słabe dopasowanie regresji liniowej

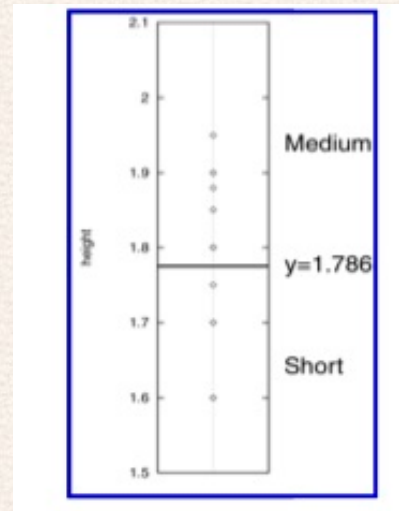


Klasyfikacja przy użyciu regresji

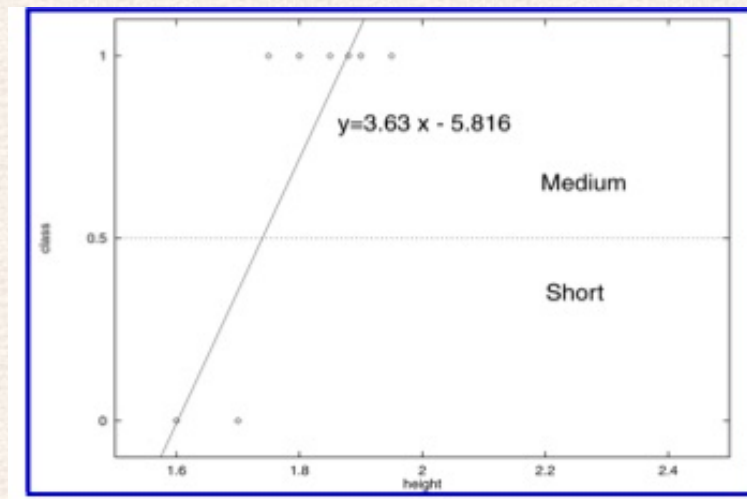
Dzielenie: użyj funkcji regresji, aby podzielić obszar na regiony.

Przewidywanie: użyj funkcji regresji do przewidywania funkcji przynależności do klasy. Dane wejściowe obejmują żądaną klasę.

Dzielenie



Przewidywanie



Klasyfikatory oparte na instancji

Przykłady:

- Rote-learner
 - Zapamiętuje całe dane szkoleniowe i przeprowadza klasyfikację tylko wtedy, gdy atrybuty rekordu dokładnie odpowiadają jednemu z przykładów szkoleniowych
- Najbliższy sąsiad
 - Wykorzystuje k „najbliższych” punktów (najbliższych sąsiadów) do wykonania klasyfikacji

Klasyfikacja za pomocą miar odległości

- Umieść przedmioty w klasie, do której są „najbliższe”.
- Musi określić odległość między przedmiotem a klasą.
- Klasy reprezentowane przez
 - Centroid: wartość centralna.
 - Medoid: punkt reprezentacyjny.
 - Zestaw pojedynczych punktów
- Algorytm: K-najbliższych sąsiadów (KNN)

Klasyfikacja najbliższych sąsiadów

Wymaga trzech rzeczy

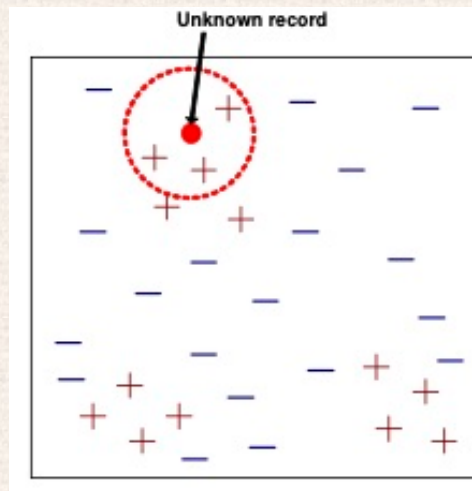
- Zbiór przechowywanych rekordów
- Metryka odległości do obliczenia odległości między rekordami
- wartość k , tj. liczba najbliższych sąsiadów do pobrania

Aby sklasyfikować nieznany rekord:

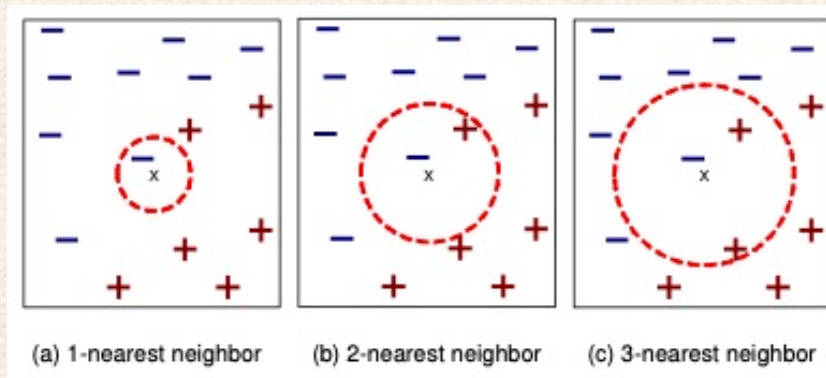
- Oblicz odległość do innych

Rejestry szkoleń

- Zidentyfikuj k najbliższych sąsiadów
- Użyj etykiet klas najbliższych sąsiadów w celu ustalenia etykiety klasy o nieznanym zapisie



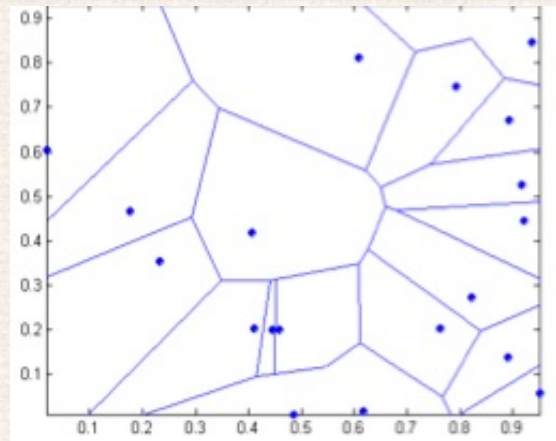
Definicja k -najbliższego sąsiada



k -najbliżsi sąsiedzi rekordu x to punkty danych, które mają k najmniejszą odległość do x

1-najbliższy sąsiad

Diagram Woronoja:



Klasyfikacja najbliższego sąsiada

- Oblicz odległość między dwoma punktami:

- Odległość euklidesowa

$$d(p, q) = \sqrt{\sum (p_i - q_i)^2}$$

- Określ klasę z listy najbliższych sąsiadów

- weź większość głosów przy etykietach klas wśród k – najbliższych sąsiadów

- zważ głos zgodnie z odległością:

- waga, $w = 1 / d^2$

k najbliższych sąsiadów (KNN)

- Zestaw treningowy obejmuje zajęcia.
- Zbadaj k pozycji w pobliżu przedmiotu do sklasyfikowania.
- Nowa pozycja umieszczona w klasie o największej liczbie bliskich pozycji.
- $O(q)$ operacji dla każdej sklasyfikowanej krotki (tutaj q jest rozmiarem zestawu uczącego.)

Algorytm KNN

Input

D // training data
N // neighbors
t // tuple to classify

Output

c // class to which t gets classified

KNN Algorithm

$N = \emptyset$;

for each $d \in D$ **do**

if $|N| < K$ **then** $N = N \cup \{d\}$;

else

u = the item in N with max distance (dissimilarity) from t ;

if $\text{sim}(t,u) < \text{sim}(t,d)$ **then** $N = (N - \{u\}) \cup \{d\}$;

c = the class to which most $n \in N$ are classified ;