Eksploracja danych

# Wykład 4
# Klasyfikacja najbliższych sąsiadów

dr hab.inż. Marcin Iwanowski

## Klasyfikacja najbliższych sąsiadów

- Podstawowe informacje
- Klasyfikatory minimalnoodległościowe
- Klasyfikator k najbliższych sąsiadów (k-NN)
- Warianty klasyfikatora k-NN
- Klasyfikator NP – najbliższego prototypu
- Warianty klasyfikatora NP

## Klasyfikacja najbliższych sąsiadów

- Podstawowe informacje
- Klasyfikatory minimalnoodległościowe
- Klasyfikator k najbliższych sąsiadów (k-NN)
- Warianty klasyfikatora k-NN
- Klasyfikator NP – najbliższego prototypu
- Warianty klasyfikatora NP



## Zadanie klasyfikacji

- Klasyfikacja - określanie przynależności obiektów do określonych klas (kategorii).
- Obiekty w procesie klasyfikacji są opisane atrybutami opisującymi i pojedynczym atrybutem decyzyjnym
- Obiekt o nieznanej przynależności w wyniku klasyfikacji otrzymuje etykietę kategorii – wcześniej nieznaną wartość atrybutu decyzyjnego
- Atrybut decyzyjny jest zawsze typu kategorycznego
- Atrybuty opisujące mogą być dowolnych typów

## Klasyfikacja – przykład 1

- System informatyczny wspomagający doradcę bankowego w zakresie oceny zdolności kredytowej potencjalnego kredytobiorcy
- Na podstawie danych klienta określana jest kategoria klientów do której jest zaliczany – w zależności od otrzymanej kategorii dobierana jest oferta (parametry) kredytu
- Atrybuty opisujące – niezbędne do oceny kredytobiorcy np. wiek, poziom dochodów, liczba osób na utrzymaniu itd.
- Atrybut decyzyjny – kategoria klientów



## Klasyfikacja – przykład 2

- Wykrywanie oszustw
- Na przykład - nadużyć płatności kartami kredytowymi
- Atrybuty opisujące to parametry płatności (częstotliwość, miejsca, kwoty itp.)
- Wynik klasyfikacji to kategoria płatności:
  - Prawidłowa (brak akcji)
  - Podejrzana (sugestia telefonu do klienta w celu weryfikacji)
  - Nieuczciwa (blokada)

## Klasyfikacja – przykład 3

- Diagnostyka medyczna
- Określanie choroby na podstawie objawów
- Atrybuty opisujące to objawy
- Atrybut decyzyjny (wynik klasyfikacji) to jednostka chorobowa



## Klasyfikator

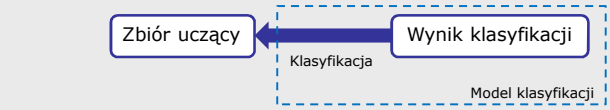
- Algorytm realizujący zadanie klasyfikacji
- Klasyfikator jest charakteryzowany przez model i jego parametry
- Model klasyfikatora – ustalany przez projektanta w zależności od specyfiki zadania klasyfikacji
- Parametry modelu – ustalane w procesie uczenia klasyfikatora
- Klasyfikacja jest uczeniem nadzorowanym – wymaga znajomości danych referencyjnych – zbioru uczącego

# Zbiór uczący

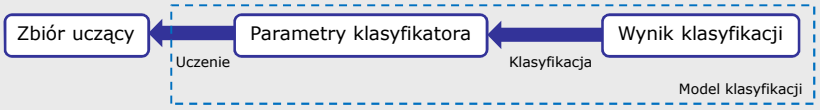
- Klasyfikator wymaga istnienia zbioru uczącego – macierzy danych zawierającej obiekty opisane atrybutami opisującymi oraz atrybutem decyzyjnego (obiekty o znanej przynależności do klas)
- Jest to zbiór referencyjny – zawiera dane (obiekty) sklasyfikowane
- Zbiór ten jest w trakcie klasyfikacji wykorzystywany
  - Bezpośrednio – klasyfikatory minimalnoodległościowe
  - Pośrednio – klasyfikatory wykorzystujące funkcję dyskryminacji i inne metody klasyfikacji



# Uczenie i jego brak




*lazy learning*




*eager learning*

# Zbiór uczący - przykład

Obiekty:

*Kombinacje warunków pogodowych*

Atrybuty opisujące:

*Warunki pogodowe*

Atrybut decyzyjny:

*Decyzja o grze*

| Nr | Pogoda | Temperatura | Wilgotność | Wiatr | Golf |
|---|---|---|---|---|---|
| 1 | Słońce | Wysoka | Wysoka | Nie | nie |
| 2 | Słońce | Wysoka | Wysoka | Tak | nie |
| 3 | Zachmurzenie | Wysoka | Wysoka | Nie | tak |
| 4 | Deszcz | Średnia | Wysoka | Nie | tak |
| 5 | Deszcz | Niska | Normalna | Nie | tak |
| 6 | Deszcz | Niska | Normalna | Tak | nie |
| 7 | Zachmurzenie | Niska | Normalna | Tak | tak |
| 8 | Słońce | Średnia | Wysoka | Nie | nie |
| 9 | Słońce | Niska | Normalna | Nie | tak |
| 10 | Deszcz | Średnia | Normalna | Nie | tak |
| 11 | Słońce | Średnia | Normalna | Tak | tak |
| 12 | Zachmurzenie | Średnia | Wysoka | Tak | tak |
| 13 | Zachmurzenie | Wysoka | Normalna | Nie | tak |
| 14 | Deszcz | Średnia | Wysoka | Tak | nie |

a. decyzyjny

Zadanie klasyfikacji:

*Jest słonecznie, temperatura jest średnia, wilgotność wysoka i nie ma wiatru*
*Czy mam grać w golfa ?*

Wynik klasyfikacji:

*Przy takich warunkach masz nie grać !*



# Proces uczenia

- Niezbędny w klasyfikatorach pośrednio wykorzystujących zbiór uczący („eager learning”)
- Na podstawie analizy danych ze zbioru uczącego są wyznaczane parametry klasyfikatora
- Po nauczeniu klasyfikator może dokonywać klasyfikacji – przypisywania obiektów o nieznanej wartości atrybutu decyzyjnego do odpowiednich klas
- W jej trakcie nie jest analizowany zbiór uczący, lecz wykorzystywane parametry wyznaczone w procesie uczenia

## Zbiór uczący i zbiór testowy

- Do weryfikacji przyjętego modelu i wyznaczonych parametrów klasyfikacji wykorzystuje się zbiór testowy
- Jest to macierz danych podobna do macierzy zbioru uczącego (tj. z atrybutami opisującymi i decyzyjnym), ale zawierająca inne obiekty
- Klasyfikując obiekty z tego zbioru możemy porównać otrzymany wynik klasyfikacji z wartością atrybutu decyzyjnego
- Im dla większej liczby obiektów wartości te się pokrywają tym lepszy mamy klasyfikator
- Często oryginalny zbiór (macierz) danych dzieli się w ustalonym stosunku na dwie części np.:
  - 70% to zbiór uczący
  - 30% to zbiór testowy



## Klasyfikacja najbliższych sąsiadów

- Podstawowe informacje
- Klasyfikatory minimalnoodległościowe
- Klasyfikator k najbliższych sąsiadów (k-NN)
- Warianty klasyfikatora k-NN
- Klasyfikator NP – najbliższego prototypu
- Warianty klasyfikatora NP

## Klasyfikatory minimalnoodległościowe

- Inaczej – klasyfikatory najbliższych sąsiadów
- Jest wyznaczana odległość między rozpoznawanym obiektem a obiektami ze zbioru uczącego (ściślej – odległości między wektorami atrybutów)
- Kluczowe znaczenie – sposób wyznaczania odległości między obiektami



## Odległość między obiektami

- Kluczowe zagadnienie w tego typu klasyfikacji
- Możliwe różne rodzaje odległości
- Atrybuty ilościowe:
  - Euklidesowa
  - Miejska
  - Mahalanobisa
- Niewspółmierność atrybutów:
  - Odległość ważona
  - Normalizacja min-max
  - Standaryzacja

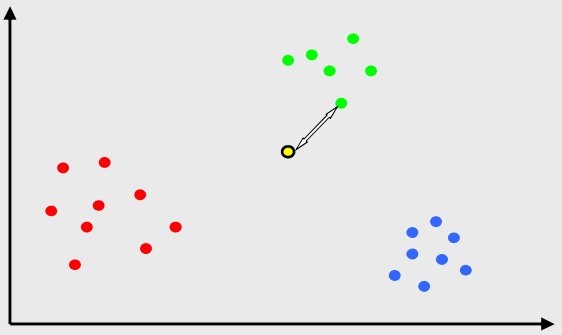
## Z uczeniem czy bez ?

- Klasyfikatory bez fazy uczenia:
  - W procesie klasyfikacji brany jest pod uwagę cały zbiór uczący
  - Uczenie tzw. leniwe (ang. „lazy learning”)
  - k-NN w swoich różnych wersjach
- Klasyfikatory z fazą uczenia:
  - Przed właściwą klasyfikacją następuje faza uczenia
  - Wyznaczane są pewne parametry modelu klasyfikacji
  - Klasyfikacja właściwa bazuje na modelu i jego parametrach (zbiór uczący na tym etapie nie jest potrzebny)
  - Klasyfikator NP w swoich różnych wersjach



## Klasyfikator najbliższego sąsiada

- Najprostszy klasyfikator minimalnoodległosciowy
- Podczas klasyfikacji rozważane są odległości miedzy klasyfikowanym obiektem a obiektami należącymi do zbioru uczącego
- Rozpoznawany obiekt jest klasyfikowany do klasy charakteryzującej najbliższy obiekt ze zbioru uczącego



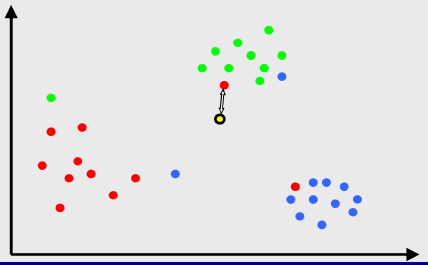
## Wady i zalety

- Zalety:
  - Prostota
  - Brak fazy uczenia
- Wady:
  - Konieczność analizowania całego zbioru uczącego podczas klasyfikacji
  - Czasochłonność !
  - Wrażliwość na błędy w zbiorze uczącym



## Wrażliwość na błędy w zb.uczącym

- Obiekty należące do różnych klas tworzą często naturalne zgrupowania w przestrzeniu wartości atrybutów.
- Błędy w danych uczących objawiają się jako pojedyncze obiekty należące do jednej klasy znajdujące się w obszarze zawierającym zgrupowanie obiektów innej klasy.
- Jeśli rozpoznawany obiekt znajduje się w obszarze zgrupowania pewnej klasy, ale niefortunnie w jego bezpośredniej bliskości znajduje się obiekt błędnie sklasyfikowany w zbiorze uczącym, wówczas rozpoznawany obiekt może zostać zaliczony do niewłaściwej klasy.



## Klasyfikacja najbliższych sąsiadów

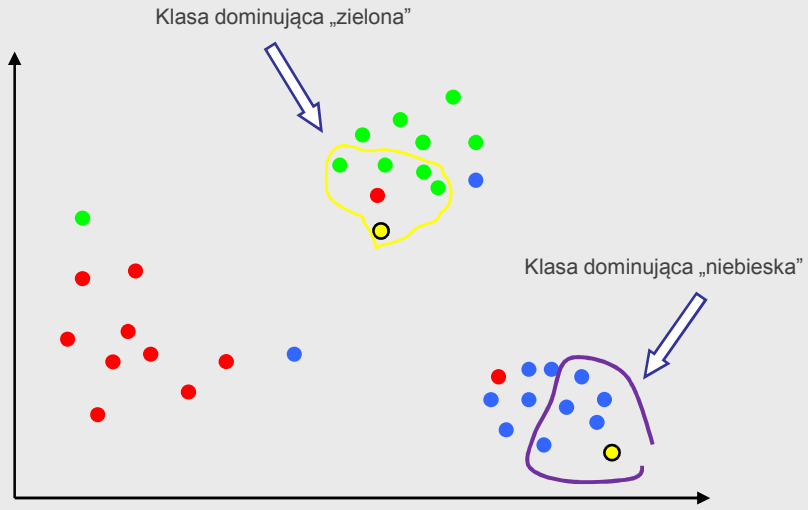
- Podstawowe informacje
- Klasyfikatory minimalnoodległościowe
- Klasyfikator k najbliższych sąsiadów (k-NN)
- Warianty klasyfikatora k-NN
- Klasyfikator NP – najbliższego prototypu
- Warianty klasyfikatora NP



## Klasyfikator k-najbliższych sąsiadów

- Zamiast szukania jednego najbliższego sąsiada, wyszukiwanych jest ustalona liczba *k* najbliższych sąsiadów
- Jako wynik klasyfikacji wybierana jest klasa dominująca
- Zaleta – mniejsze ryzyko błędnej klasyfikacji związanej z istnieniem nietypowego wektora cech
- Wada – wysoki koszt obliczeniowy, konieczne sortowanie odległości
- Poziom ufności: stosunek liczby sąsiadów klasy dominującej do całkowitej liczby uwzględnianych sąsiadów *k*
- ang.: *k*-nearest neighbors (*k*-NN)

# Klasyfikator k-najbl.sąsiadów

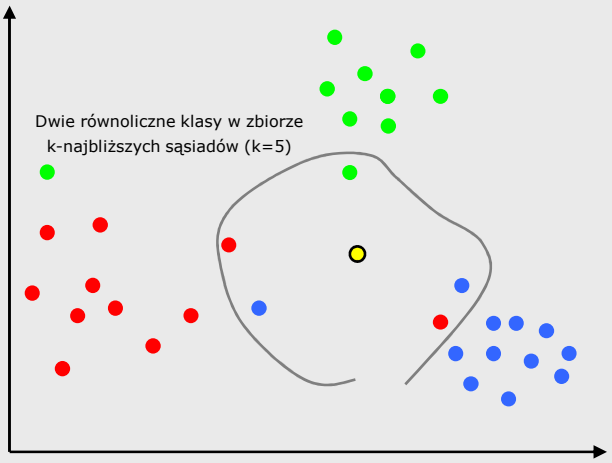





# Dobór wielkości k

- k-nieparzyste (ułatwia głosowanie)
- Często dobieramy eksperymentalnie (uruchamiamy dla różnych k i oceniamy jakość klasyfikacji)
- Im mniej obiektów należących do poszczególnych klas, tym k-mniejsze
- W szczególności, gdy mamy po jednym przedstawicielu każdej klasy, wtedy k=1 (najprostszy klas.minimalnoodległościowy)

# Problem remisu




Możliwe rozwiązania:

- Losowanie klasy
- Wynik nierozstrzygnięty
- Rozważenie wariantu z większym k



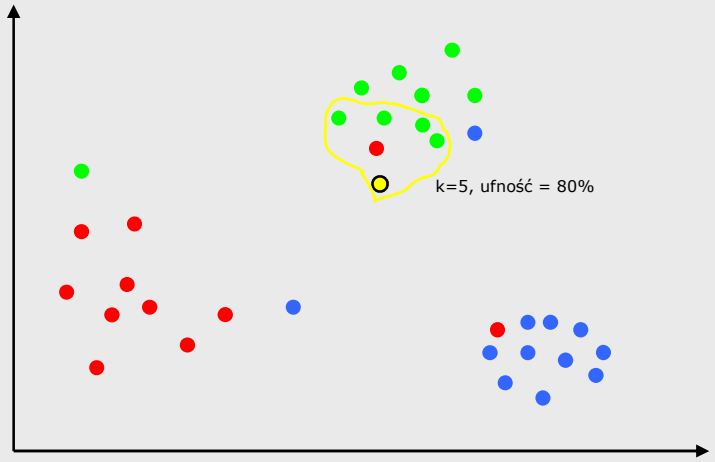
# Poziom ufności (1)

- Poziom ufności określa wiarygodność wyniku klasyfikacji
- Stosunek liczby sąsiadów w klasie dominującej ($l$) do $k$

$$ufność = \frac{l}{k}$$

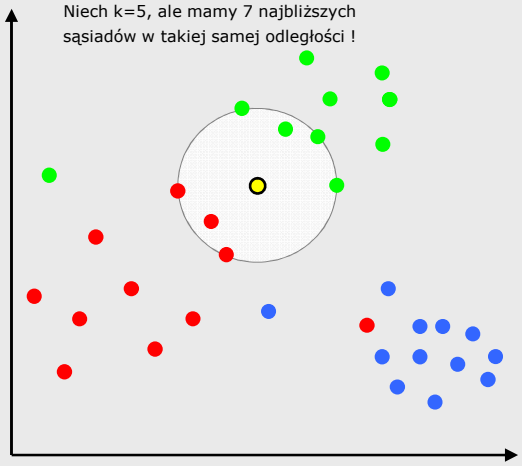
- Jeśli chcemy by decyzja była „mocna” wprowadzamy próg na $l$ (alternatywnie próg na ufność) -> klasyfikator $(k,l)$-NN
- Dla liczby sąsiadów (ufności) poniżej progu – wynik nierozstrzygnięty

# Poziom ufności (2)






# Problem obiektów równoodległych




- Większa niż k liczba obiektów najbliższych jest równoodległa od klasyfikowanego obiektu
- Może się zdarzyć w szczególnie przypadku atrybutów o wartościach dyskretnych
- Rozwiązanie: dopuszczamy w takich przypadkach większą od k liczbę najbliższych sąsiadów (wszystkie te, które są odległe co najwyżej o tyle samo co k-ty obiekt)

## Klasyfikacja najbliższych sąsiadów

- Podstawowe informacje
- Klasyfikatory minimalnoodległościowe
- Klasyfikator k najbliższych sąsiadów (k-NN)
- Warianty klasyfikatora k-NN
- Klasyfikator NP – najbliższego prototypu
- Warianty klasyfikatora NP



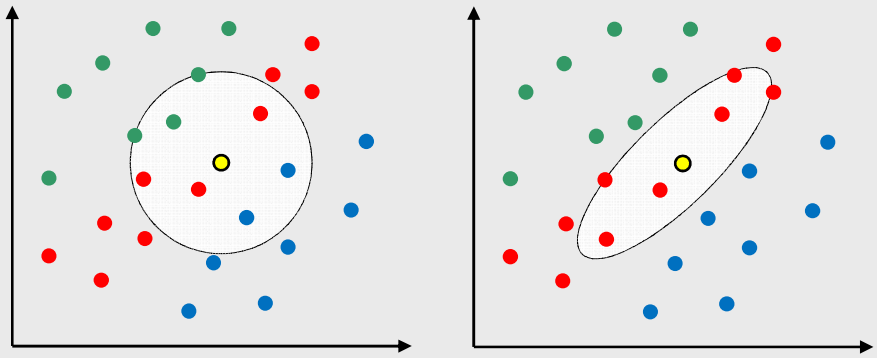
## Zastosowanie innych metryk odległości

- Odległość Euklidesowa nie zawsze pozwala na właściwe oszacowanie (nie)podobieństwa obiektów – np. gdy obszar na którym znajdują się obiekty w przestrzeni atrybutów ma formę wydłużoną
- Czy istnieją metryki pozwalające na uwzględnienie takich sytuacji ? Tak – metryka Mahalanobisa:

$$d_M(\mathbf{a},\mathbf{b}) = \sqrt{(\mathbf{a}-\mathbf{b})C^{-1}(\mathbf{a}-\mathbf{b})^T} \qquad d_E(\mathbf{a},\mathbf{b}) = \sqrt{(\mathbf{a}-\mathbf{b})(\mathbf{a}-\mathbf{b})^T}$$

- Wymagana jest znajomość macierzy kowariancji dla każdej z klas (ich wyliczenie – to już uczenie !)

## Metryka Euklidesowa a Mahalanobisa

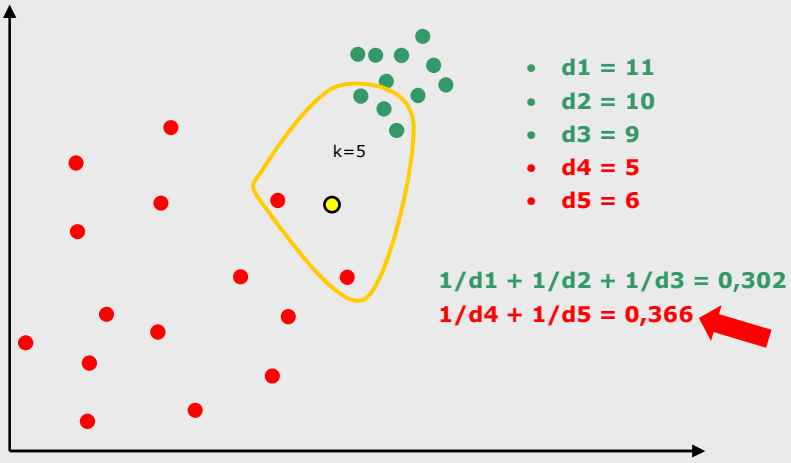




## Ważony klasyfikator k-najbl.sąsiadów

- Klasyczny klasyfikator k-najbliższych sąsiadów nie zawsze klasyfikuje poprawnie bo nie uwzględnia odległości do tych sąsiadów
- Rozwiązanie – ważony klasyfikator k-najbliższych sąsiadów
- Przy wyznaczaniu klasy dla rozpoznawanego obiektu uwzględnia się odległości do poszczególnych najbliższych sąsiadów
- Każdy spośród sąsiadów otrzymuje „wagę” swojego głosu – odwrotność odległości
- Wybierana jest klasa o największej sumie głosów ważonych

# Ważony klasyfikator k-NN






# Rozmyty klasyfikator *k*-NN (1)

- Wariant klasyfikatora ważonego k-NN
- Wynik jest wskazaniem na więcej niż jedną klasę
- Określenie przynależności do klasy – wartość funkcji przynależności
- Rozmyta wersja klasyfikator *k*-NN – rozmyty wynik klasyfikacji – zbiór funkcji przynależności do poszczególnych klas
- Na stopień przynależności oprócz ilości sąsiadów danej klasy wpływa odwrotność odległości od najbliższych sąsiadów

## Rozmyty klasyfikator *k*-NN (2)

- Wartość funkcji przynależności do i-tej klasy:

$$u_i(x) = \frac{\sum_{j=1}^{K} u_{ij} \frac{1}{\left\|x - x_j\right\|^{\frac{2}{m-1}}}}{\sum_{j=1}^{K} \frac{1}{\left\|x - x_j\right\|^{\frac{2}{m-1}}}}$$

- parametr *m* reguluje siłę oddziaływania tych odległości podczas obliczaniu wpływu każdego sąsiada na funkcję przynależności.
- wraz ze wzrostem *m* wpływ odległości jest coraz mniejszy - wówczas silniej decydująca jest liczba sąsiadów danej klasy.



## Klasyfikacja najbliższych sąsiadów

- Podstawowe informacje
- Klasyfikatory minimalnoodległościowe
- Klasyfikator k najbliższych sąsiadów (k-NN)
- Warianty klasyfikatora k-NN
- Klasyfikator NP – najbliższego prototypu
- Warianty klasyfikatora NP

## Klasyfikator najbliższego prototypu (NP)

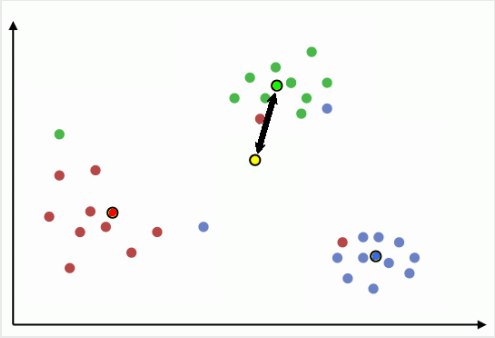
- W klasyfikatorach k-NN liczymy odległości do wszystkich obiektów w zbiorze uczącym
- Przy dużej liczbie obiektów w klasach może być to nieefektywne obliczeniowo, w przypadku klas dobrze określonych w przestrzeni atrybutów – całkowicie zbędne
- Możemy przecież (licznych) przedstawicieli klasy zastąpić specjalnie wyliczonym jej prototypem
- Prototyp (wzorzec) klasy jest „wirtualnym" obiektem kumulującym własności klasy
- Zamiast więc szukać najbliższego obiektu możemy szukać najbliższego prototypu



## NP a uczenie

- Mając prototyp nie trzeba już korzystać ze zbioru uczącego
- Ale wyznaczenie prototypu wymaga jego analizy – faza uczenia !
- W fazie uczenia określamy parametry charakteryzujące klasy
- Parametry te powinny mieć różne wartości dla poszczególnych klas (pozwalać na rozróżnienie klas)
- Miary tendencji centralnej – np. średnia arytmetyczna, alternatywnie mediana

# Klasyfikator NP - ilustracja



- Wzorzec $i$-tej klasy $w_i$ – „wirtualny” (tj. nieistniejący w zbiorze uczącym) obiekt o wartościach atrybutów równych średnim wartościom atrybutów obiektów należących do $i$-tej klasy
- Szukamy najbliższego wzorca w sensie ustalonej metryki
- Klasyfikator 1-NN dla prototypów



# NP a funkcja dyskryminacji

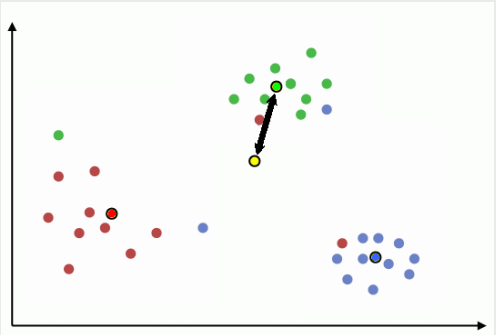
- Klasyfikator NP możemy rozważać wykorzystując pojęcie funkcji dyskryminacji
- Argumentami funkcji dyskryminacji są wartości atrybutów
- Postać funkcji dyskryminacji jest ustalona (zależy od przyjętego modelu klasyfikacji), zaś parametry są wyznaczane niezależnie dla każdej klasy w procesie uczenia
- Postaci funkcji dyskryminacji jest więc tyle, ile klas Postacie różnią się parametrami.
- Parametry są wyznaczane w procesie uczenia

# Funkcja dyskryminacji

- Nieco bardziej formalnie ($n$ – liczba klas):
  - w wyniku uczenia powstaje więc $n$ postaci funkcji dyskryminacji $g_i(\mathrm{p})$, gdzie $0 < i \leq n$ jest indeksem klasy.
  - argumentem tych funkcji jest wektor wartości atrybutów opisujących klasyfikowany obiekt $\mathrm{p}=[p_1,p_2,\ldots,p_m]$, przy czym $p_i$ jest wartością $i$-tego atrybutu.
  - właściwa klasyfikacja wektora atrybutów p obiektu o nieznanej przynależności do klasy polega na wyznaczeniu wartości $g_i(\mathrm{p})$ dla wszystkich $i$, a następnie znalezieniu klasy $j$ dla której wartość ta jest najmniejsza:
$$\forall j \neq i \quad g_j(\mathbf{p}) < g_i(\mathbf{p}).$$
  - alternatywnie, dla niektórych form funkcji dyskryminacji stosuje się zasadę, że wynik klasyfikacji jest determinowany klasą, dla której wartość funkcji dyskryminacji jest największa:
$$\forall j \neq i \quad g_j(\mathbf{p}) > g_i(\mathbf{p}).$$



# Minimalnoodległościowa metoda wzorców



- Funkcja dyskryminacji ($w_i$ – wzorzec $i$-tej klasy):
$$g_i(\mathbf{p}) = d(\mathbf{p}, \mathbf{w}_i)$$
- Dla odległości Euklidesowej możemy przyjąć:
$$g_i(\mathbf{p}) = d_E^{\,2}(\mathbf{p}, \mathbf{w}_i) = (\mathbf{p} - \mathbf{w}_i) \cdot (\mathbf{p} - \mathbf{w}_i)^T$$
- Szukamy takiego klasy j dla której: $\forall j \neq i \quad g_j(\mathbf{p}) < g_i(\mathbf{p}).$

## Klasyfikacja z funkcją dyskryminacji

- Dwie fazy pracy:
  - Uczenie klasyfikatora - wyznaczenia funkcji dyskryminacji na podstawie zbioru uczącego
  - Klasyfikacja właściwa
- Klasyfikacja właściwa obiektu polega na
  1. wyznaczeniu wartości wszystkich funkcji dyskryminacji dla wartości atrybutów tego obiektu
  2. znalezieniu najmniejszej lub największej (zależnie od przyjętego rodzaju klasyfikatora) wartości
  3. przypisaniu klasyfikowanemu obiektowi klasy dla której wartość funkcji jest najmniejsza/największa



## Klasyfikacja najbliższych sąsiadów

- Podstawowe informacje
- Klasyfikatory minimalnoodległościowe
- Klasyfikator k najbliższych sąsiadów (k-NN)
- Warianty klasyfikatora k-NN
- Klasyfikator NP – najbliższego prototypu
- Warianty klasyfikatora NP

# Klasyfikator Mahalanobisa

- Wariant klasyfikatora najbliższych prototypów
- Odległość do wzorca jest odległością Mahalanobisa
- Funkcja dyskryminacji:

$$g_i(\mathbf{p}) = (\mathbf{p} - \mathbf{w_i}) \cdot C_i^{-1} \cdot (\mathbf{p} - \mathbf{w_i})^T$$

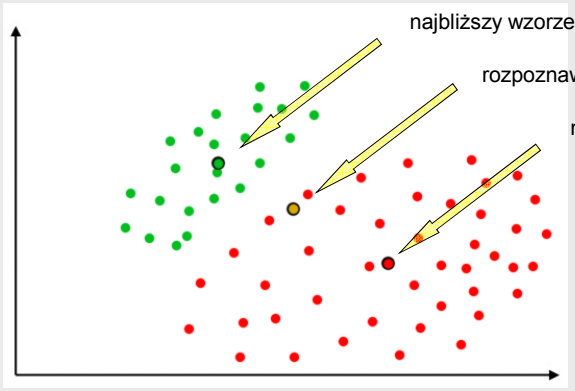
- Macierz kowariancji C – opisuje liniowe zależności między atrybutami obiektów w danej klasie
- Szukamy klasy j dla której:

$$\forall j \neq i \quad g_j(\mathbf{p}) < g_i(\mathbf{p}).$$

- Dobry dla atrybutów o pewnym stopniu skorelowania
- Uczenie – wyznaczenie wzorców klas oraz macierzy kowariancji dla każdej klasy



# Klasyfikator Mahalanobisa



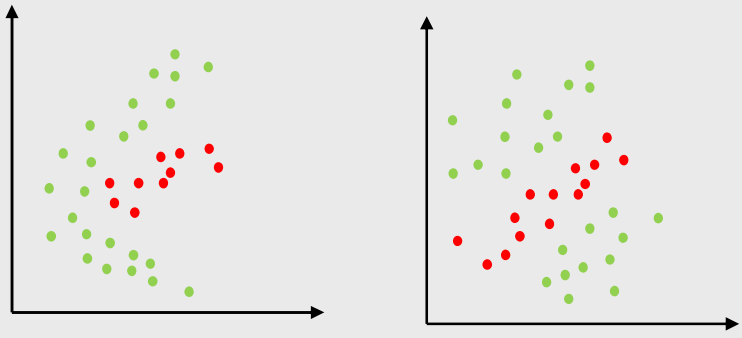

- Klasy o różnym rozproszeniu
- Klasyczna odległość od wzorce powoduje błędną klasyfikację
- Zastosowanie odległości Mahalanobisa rozwiązuje problem

## Zastosowanie k-NN w NP

- Wykorzystanie większej (niż 1) liczby prototypów dla każdej klasy
- Można wówczas zastosować k-NN zamiast 1-NN
- Dodatkowe zadanie w procesie uczenia: znalezienie większej liczby prototypów klas
- Zastosowanie: nietypowe dystrybucje obiektów w przestrzeni atrybutów
- Spojrzenie od drugiej strony: stosujemy k-NN ale na zmodyfikowanym (zredukowanym) zbiorze uczącym



## Kiedy jeden prototyp nie wystarcza ?



- Miara tendencji centralnej obiektów w danej klasie nie pozawala na jednoznaczną identyfikację obiektów tej klasy i odróżnienie ich od obiektów w pozostałych klasach
- Np. prototyp jednej klasy znajduje się na obszarze innej

## Wyznaczanie prototypów klas

- Kilka możliwości:
  - Wykorzystanie dodatkowej wiedzy o klasach (jeśli możliwe dodajemy nowy atrybut).
  - Redukujemy liczbę obiektów w klasach przez usunięcie obiektów niewpływających na klasyfikację (algorytm CNN = condensed NN)
  - Metody grupowania danych (uczenie nienadzorowane), faktyczne wyznaczenie prototypów (obiektów „wirtualnych") -> późniejsze wykłady.
- Dodatkowy nakład obliczeniowy w fazie uczenia
- Korzyść na etapie klasyfikacji właściwej (mniej odległości do wyznaczenia)



## Przykładowe zastosowanie

- Schemat zastosowania do rozpoznawania znaków alfanumerycznych