UCZENIE MASZYNOWE

KLASYFIKATORY MINIMALNOODLEGŁOŚCIOWE

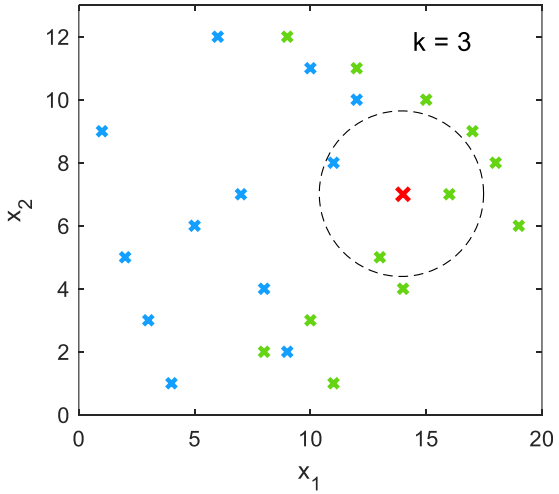
Dr hab. inż. Grzegorz Dudek
Wydział Elektryczny
Politechnika Częstochowska

Projekt finansowany w ramach programu Ministra Nauki i Szkolnictwa Wyższego pod nazwą „Regionalna Inicjatywa Doskonałości" w latach 2019 - 2022 nr projektu 020/RID/2018/19 kwota finansowania 12 000 000 PLN

KLASYFIKATOR *K*-NN – IDEA

Jedną z najczęściej stosowanych metod klasyfikacji jest metoda pamięciowa (nieparametryczna) *k*-najbliższych sąsiadów. Metoda nie wymaga estymacji warunkowych funkcji gęstości.

W metodzie tej przykład zalicza się do tej klasy, do której należy większość z jego *k*-najbliższych sąsiadów. Klasyfikacja odbywa się poprzez wyznaczenie odległości lub podobieństwa pomiędzy klasyfikowanym przykładem **x***, a każdym przykładem ze zbioru trenującego. Klasyfikowanemu przykładowi przypisywana jest klasa reprezentowana przez największą liczbę przykładów spośród *k* najbliższych sąsiadów.

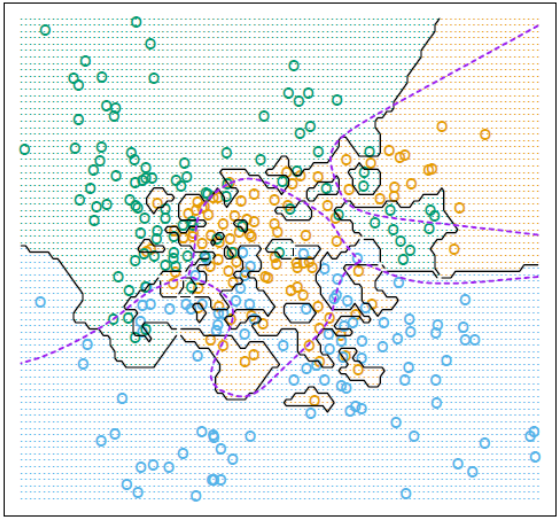


W przypadku gdy kilka klas reprezentowanych jest przez tę samą liczbę sąsiadów, wybierana jest klasa sąsiadów najbliższych klasyfikowanemu obiektowi lub klasa liczniej reprezentowana w zbiorze trenującym.

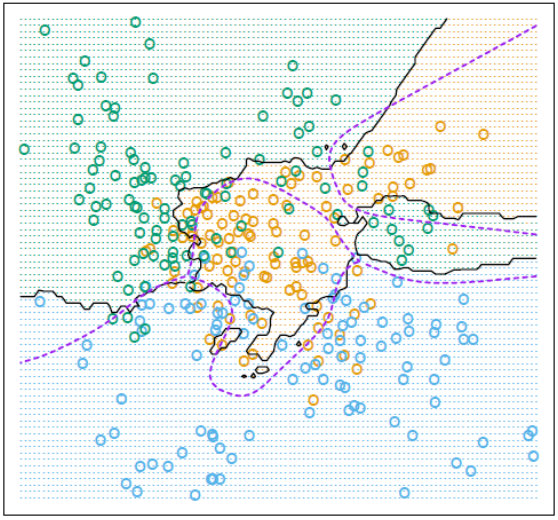
KLASYFIKATOR *K*-NN – IDEA

Parametr *k* należy dobrać eksperymentalnie w celu otrzymania jak najlepszej klasyfikacji dla danego zbioru danych. Dla małych wartości *k* linia/powierzchnia dyskryminacyjna dopasowuje się dokładnie do danych (podatność na szumy). Dla dużych wartości *k* linia/powierzchnia dyskryminacyjna jest gładsza.

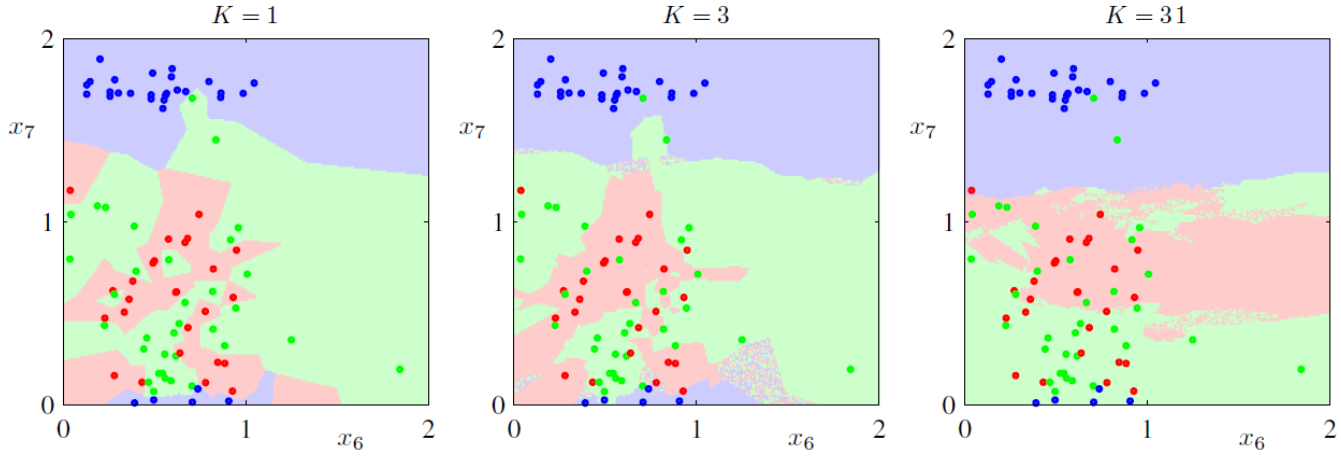
1-Nearest Neighbor



15-Nearest Neighbors

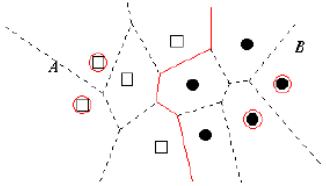


KLASYFIKATOR *K*-NN – IDEA

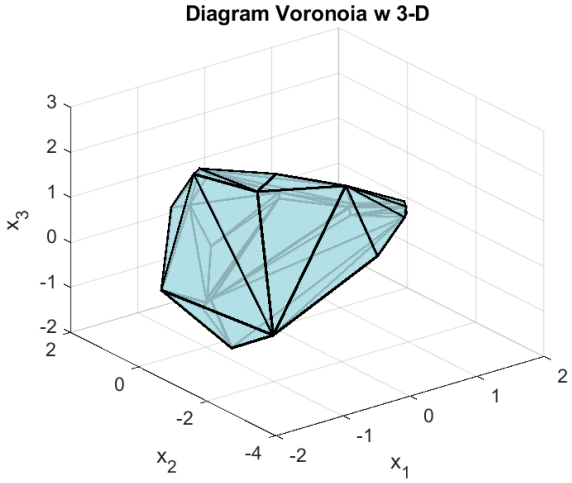


KLASYFIKATOR *K*-NN – DIAGRAM VORONOIA

Granica decyzyjna dla metody 1-NN należy do diagramu Voronoia.



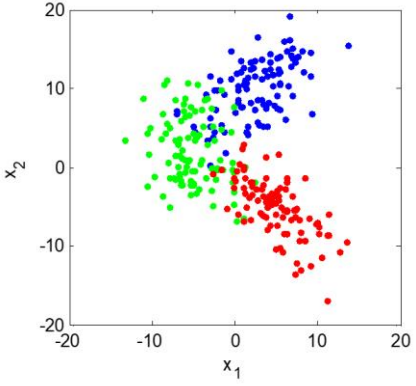
Węzły regionów Voronoia, których krawędzie nie należą do granicy decyzyjnej są nadmiarowe. Można je usunąć ze zbioru uczącego.

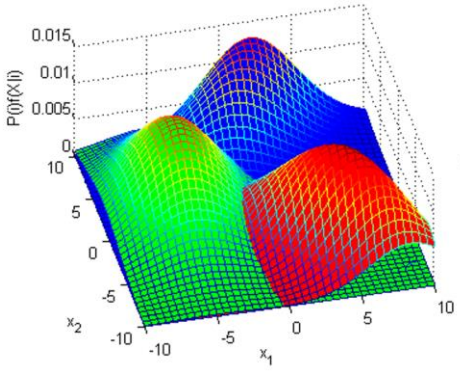


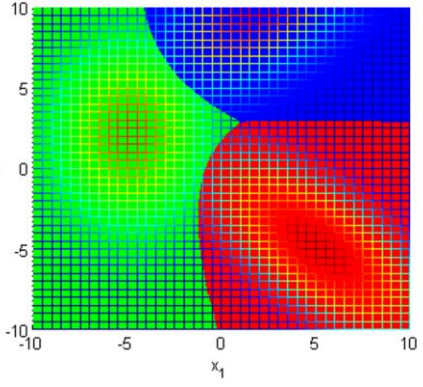
Mówiąc o metodach konstrukcji klasyfikatorów nie sposób pominąć pytania o teoretycznie najlepszy, w sensie minimum prawdopodobieństwa mylnej decyzji, klasyfikator dla danego zestawu cech. Taki klasyfikator można by zbudować znając rozkłady lub gęstości rozkładów prawdopodobieństw dla każdej z klas. W przypadku ciągłych wartości cech, należałoby skorzystać ze znanego wzoru Bayesa:

$$p(j/\boldsymbol{x})=p(j)*f(\boldsymbol{x}/j)/f(\boldsymbol{x}), \qquad (2.1)$$

dla wektorów $\boldsymbol{x}$, których składowe przyjmują wartości ciągłe. W pierwszym ze wzorów $p(j/\boldsymbol{x})$ jest prawdopodobieństwem, że obiekt o cechach $\boldsymbol{x}$ jest z klasy j, $p(j)$ prawdopodobieństwem a priori klasy j, $f(\boldsymbol{x}/j)$ oraz $f(\boldsymbol{x})$ oznaczają gęstości rozkładu prawdopodobieństw wartości wektorów $\boldsymbol{x}$ odpowiednio dla klasy j oraz gęstość rozkładu wektorów $\boldsymbol{x}$ bez względu na klasę.







Klasyfikator działający według podanego wyżej wzoru nazywa się klasyfikatorem Bayesowskim. Przypisuje on punkt $\boldsymbol{x}$ do klasy, do której należy on z największym prawdopodobieństwem, tj. wskazuje taką klasę i, że

$$p(i)*f(\boldsymbol{x}/i)=\max_j p(j)*f(\boldsymbol{x}/j). \qquad (2.2)$$

Relacja (2.2) oznacza, że $p(i/\boldsymbol{x})$ ma wartość maksymalną. Wartości funkcji występujących po prawej stronie wzoru (2.1) można oszacować, biorąc pod uwagę zbiór uczący U i otoczenie punktu $\boldsymbol{x}$ zawierające k punktów oraz korzystając z następujących przybliżeń:

$$p(j)\approx m_{jU}/m_U,\ f(\boldsymbol{x}/j)\approx(k_j/m_{jU})/V,\ f(\boldsymbol{x})\approx(k/m_U)/V, \qquad (2.3)$$

gdzie k jest liczbą dobraną eksperymentalnie, k_j jest liczbą punktów z klasy j wśród k najbliższych „sąsiadów" klasyfikowanego punktu $\boldsymbol{x}$, m_{jU} jest liczbą punktów z klasy j w zbiorze uczącym, m_U liczebością zbioru uczącego, a V objętością najmniejszej hiperkuli zawierającej ww. k najbliższych „sąsiadów". Mówiąc o hiperkuli należy określić funkcję odległości.

Z podstawienia przybliżeń (2.3) do wzoru (2.1) wynika, że:

$$p(j/\boldsymbol{x}) \approx k_j/k, \tag{2.4}$$

co oznacza, że prawdopodobieństwo $p(j/\boldsymbol{x})$ osiągnie największą wartość dla klasy najliczniej reprezentowanej wśród k najbliższych „sąsiadów" klasyfikowanego punktu $\boldsymbol{x}$, czyli dla i spełniającego relację

$$p(i/\boldsymbol{x}) \approx \max_j k_j/k. \tag{2.5}$$

Reguła, która przypisuje klasę i spełniającą warunek (2.5) nazywa się regułą *k najbliższych sąsiadów* (k-NS). Jest ona do dziś jedną z najbardziej popularnych, najczęściej stosowanych i najefektywniejszych metod klasyfikacji.

KLASYFIKATOR *K*-NN – MIARY ODLEGŁOŚCI

Do określania odległości pomiędzy punktami stosuje się:

- odległość euklidesową:

$$d(\mathbf{x}_a, \mathbf{x}_b) = [(\mathbf{x}_a - \mathbf{x}_b)^{\mathsf{T}} (\mathbf{x}_a - \mathbf{x}_b)]^{1/2} = \sqrt{\sum_{t=1}^{T} (x_{a,t} - x_{b,t})^2} \, .$$

- odległość miejską (*city-block*, *Manhattan*):

$$d(\mathbf{x}_a, \mathbf{x}_b) = \sum_{t=1}^{T} | \, x_{a,t} - x_{b,t} \, | .$$

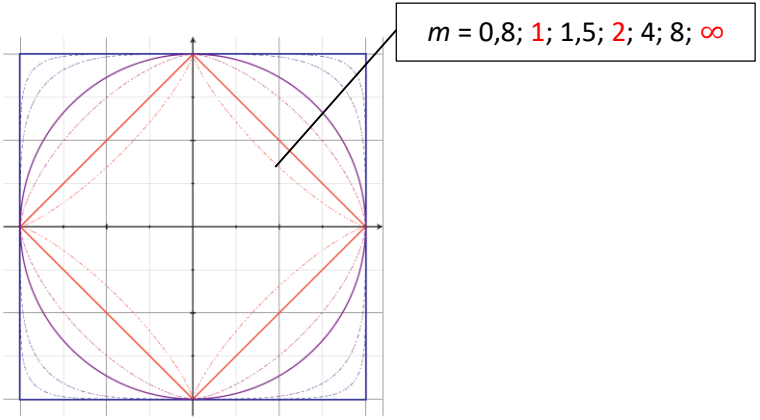
- odległość Czebyszewa:

$$d(\mathbf{x}_a, \mathbf{x}_b) = \max_{1 \le t \le T} | \, x_{a,t} - x_{b,t} \, | .$$

Każda z tych funkcji stanowi szczególny przypadek odległości Minkowskiego:

$$d(\mathbf{x}_a, \mathbf{x}_b) = \left(\sum_{t=1}^{T} | x_{a,t} - x_{b,t} |^m \right)^{1/m} .$$

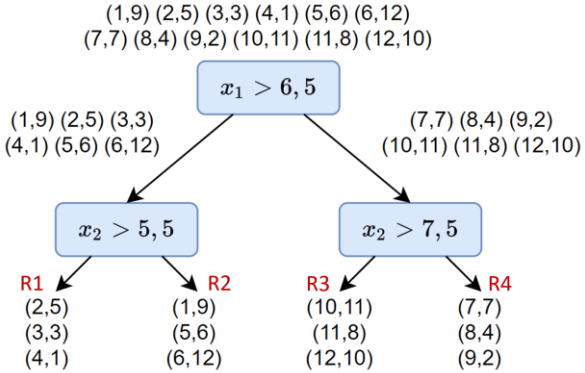
Miary oparte na odległości Minkowskiego nie są niezmiennicze względem skali wartości zmiennych. Zmiana skali powoduje zmianę odległości pomiędzy obserwacjami. Aby temu zapobiec, zaleca się wcześniejszą normalizację obserwacji.

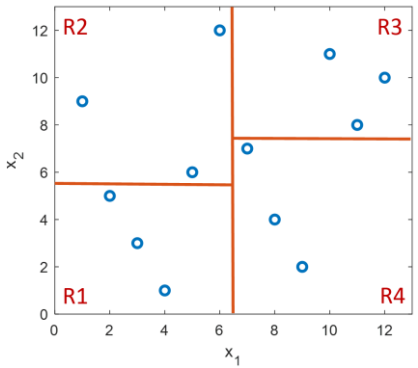


Drzewo kD (*k-dimensional tree, kD-tree*) – struktura danych, będąca wariantem drzew binarnych, używana do podziału przestrzeni.

Algorytm budowy:

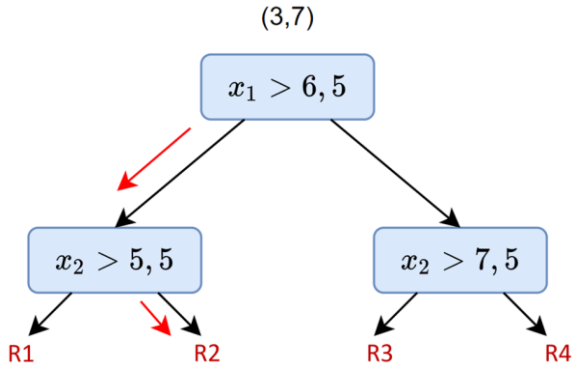
1. Wybierz losowo atrybut
2. Znajdź medianę dla tego atrybutu w zbiorze uczącym
3. Podziel przykłady na dwie części: o wartościach atrybutu większych od mediany i nie większych od mediany
4. Powtórz kroki 1-4
5. Dla każdego przykładu zapamiętaj region, do którego został przydzielony

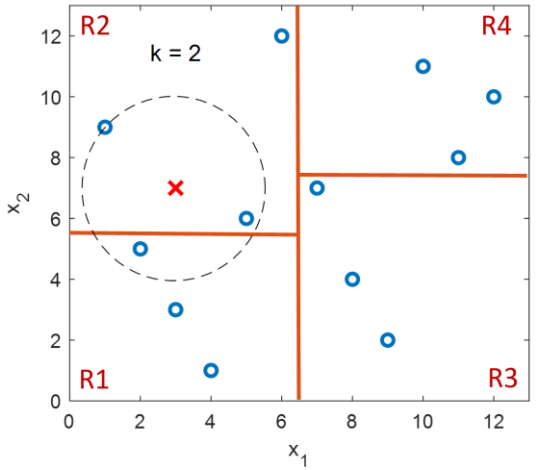




Algorytm klasyfikacji:

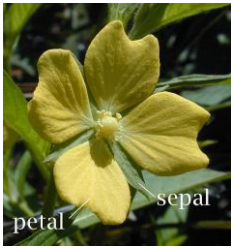
1. Przechodząc po drzewie znajdź region, do którego „wpada” klasyfikowany przykład x*
2. Znajdź *k* najbliższych przykładów w tym regionie dla x*
3. Sprawdź, czy odległość przykładu x* do najdalszego sąsiada jest mniejsza od odległości do granic sąsiednich regionów
4. Jeśli nie, szukaj sąsiadów także w regionach sąsiednich



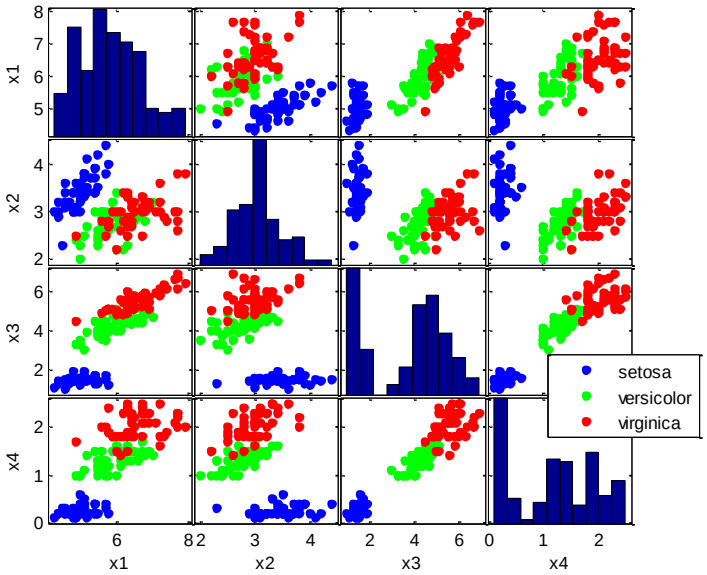


PRZYKŁAD APLIKACYJNY

Klasyfikacja zbioru danych Iris (Fisher's Iris)



Zbiór zawiera po 50 przykładów każdego z trzech gatunków kwiatu: sentosa, virginica oraz versicolor. Przykłady zbudowane są z czterech atrybutów: szerokość i długość płatków sepal i petal

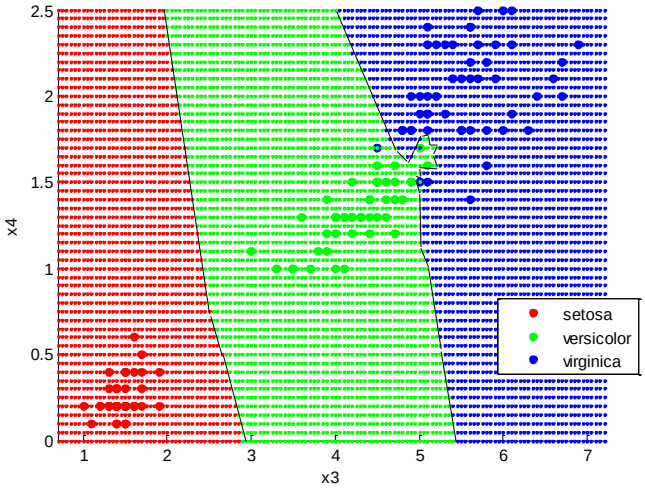


X1 (Sepal length)	X2 (Sepal width)	X3 (Petal length)	X4 (Petal width)	Klasa (Species)
5,1	3,5	1,4	0,2	1 (setosa)
4,9	3	1,4	0,2	1 (setosa)
…	…	…	…	…
6,1	2,8	4	1,3	2 (versicolor)
…	…	…	…	…
6,2	3,4	5,4	2,3	3 (virginica)
5,9	3	5,1	1,8	3 (virginica)

Klasyfikacja zbioru danych Iris (Fisher's Iris)

Macierz błędów klasyfikacji
(oszacowane w procedurze leave-one-out)

	Klasa prawdziwa		
Klasa przewidywana ↓	1	2	3
1	**50**	**0**	**0**
2	**0**	**47**	**3**
3	**0**	**3**	**47**
nierozpoznana	**0**	**0**	**0**



Odsetek poprawnie sklasyfikowanych przykładów: **96,00%**. Gdy usuniemy dwa pierwsze atrybuty przykładów, pozostawiając atrybut x_3 i x_4 wynik klasyfikacji będzie taki sam! Linie decyzyjne dla tego przypadku przedstawia rys. powyżej.

PRZYKŁAD APLIKACYJNY

Matlab: openExample('stats/ClassifyingQueryDataUsingKnnsearchExample')