SYSTEMY UCZĄCE SIĘ

WYKŁAD 11. KLASYFIKATORY MINIMALNOODLEGŁOŚCIOWE

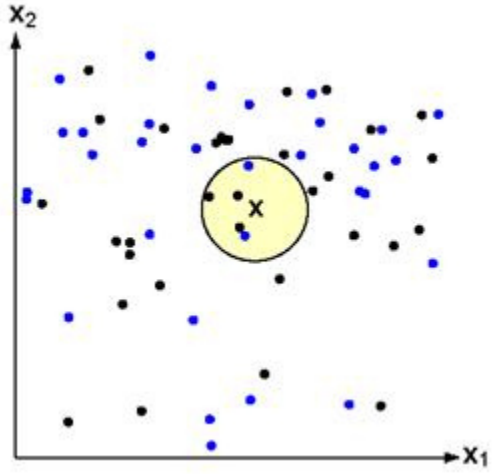
Dr hab. inż. Grzegorz Dudek
Wydział Elektryczny
Politechnika Częstochowska

Częstochowa 2014

KLASYFIKATOR *K*-NAJBLIŻSZYCH SĄSIADÓW

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 sie 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 obiektów spośród *k* najbliższych sąsiadów.

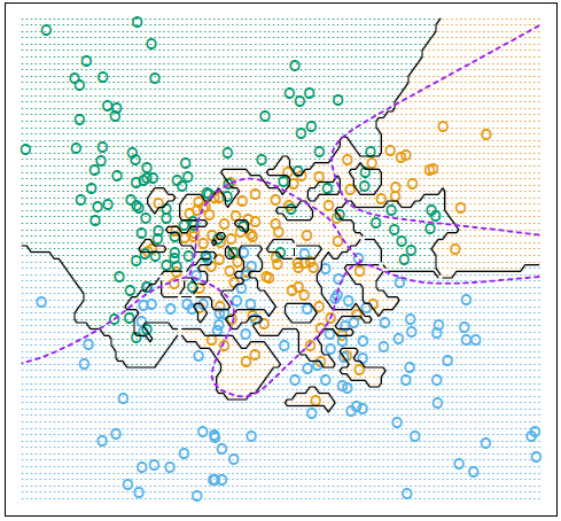


W przypadku gdy kilka klas reprezentowanych jest przez tą samą liczbę sąsiadów, wybierana jest klasa sąsiadów najbliższych klasyfikowanemu obiektowi lub klasa liczniej reprezentowana w zbiorze trenującym.

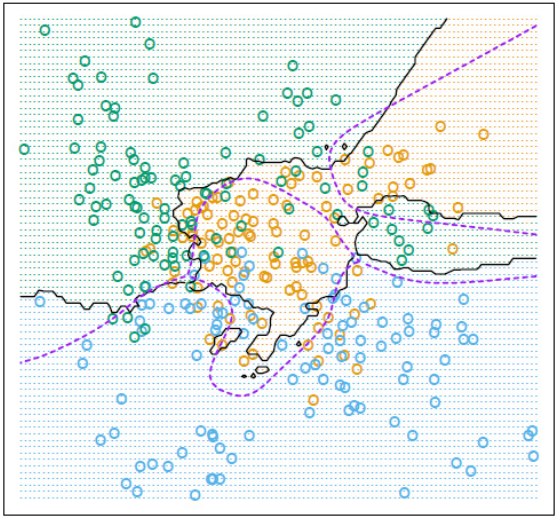
KLASYFIKATOR *K*-NAJBLIŻSZYCH SĄSIADÓW

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.

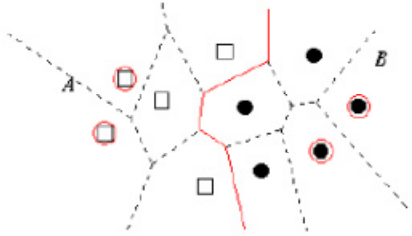
1-Nearest Neighbor



15-Nearest Neighbors



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.

KLASYFIKATOR *K*-NAJBLIŻSZYCH SĄSIADÓW

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. Dla wygody dalszych rozważań przez funkcję odległości będzie rozumiana jedna z dwu metryk: Euklidesowa lub miejska. Jeżeli będzie to istotne, to zostanie wyraźnie zaznaczone, która z tych metryk brana jest pod uwagę. 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

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

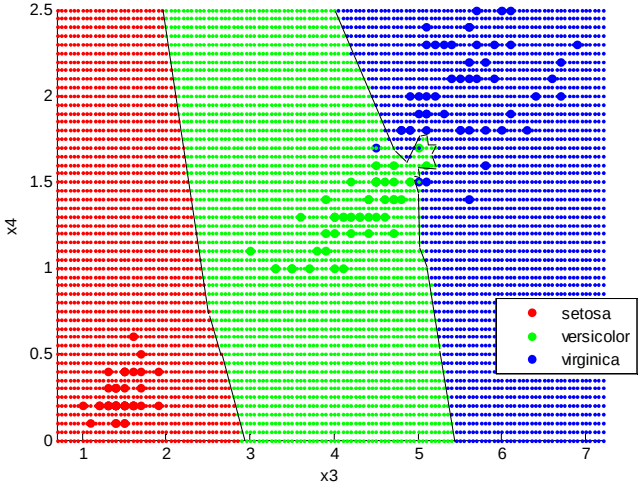
Reguła, która przypisuje klasę i spełniającą warunek (2.5) nazywa się regułą *k najbliższych sąsiadów* (k-NS). Jest ona do dziś jedną z najbardziej popularnych, najczęściej stosowanych i najefektywniejszych metod klasyfikacji.

KLASYFIKATOR *K*-NAJBLIŻSZYCH SĄSIADÓW – PRZYKŁAD

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.

* Porównaj z rozwiązaniem za pomocą drzew decyzyjnych - wykład 3, slajdy 14-16 i naiwnego klasyfikatora Bayesa - wykład 7 slajd 19.