UCZENIE MASZYNOWE

KLASYFIKATORY BAYESA

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

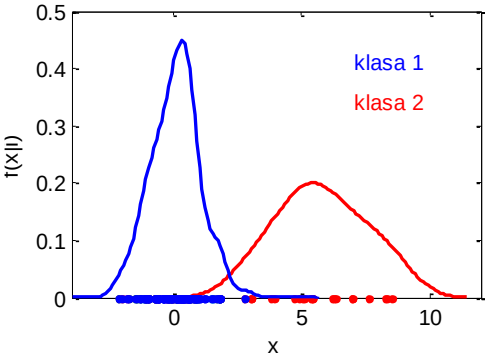
TWIERDZENIE BAYESA

Wiedza pozyskiwana przez metody probabilistyczne ma postać konstruowanych na podstawie danych trenujących oszacowań prawdopodobieństw wystąpienia w tych danych określonych wartości atrybutów i klas.

Reguła Bayesa bazuje na założeniu, że znane są:

$P(i)$ – prawdopodobieństwo *a priori*, że przykład należy do klasy i oraz

$f(\mathbf{x}|i)$ – gęstość rozkładu prawdopodobieństwa przykładów w klasie i.



Reguła Bayesa pozwala wyznaczyć prawdopodobieństwo *a posteriori* (po zaobserwowaniu **x**), że przykład należy do klasy i (i = 1, 2, ..., K):

$$P(i \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid i)P(i)}{\sum_{l=1}^{K} f(\mathbf{x} \mid l)P(l)}$$

FUNKCJE DYSKRYMINACYJNE DLA REGUŁY BAYESA

Twierdzenie Bayesa pozwala wybrać klasę (hipotezę) najbardziej prawdopodobną w świetle zaobserwowanych danych: spośród wszystkich klas (hipotez) wybieramy tę, która ma największą wartość $P(i \mid \mathbf{x})$. Prawdopodobieństwa *a posteriori* mogą pełnić rolę funkcji dyskryminacyjnych:

$$g_i(\mathbf{x}) = P(i \mid \mathbf{x})$$

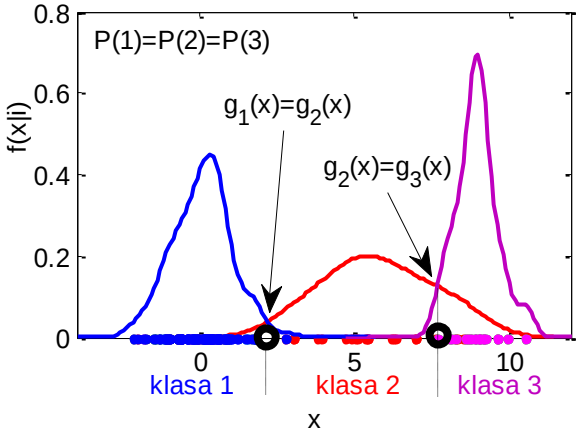
Ponieważ mianownik wzoru Bayesa jest jednakowy dla wszystkich klas możemy zapisać:

$$g_i(\mathbf{x}) = f(\mathbf{x} \mid i)P(i)$$

Przykład **x** zliczamy do klasy *i*, jeśli zachodzi:

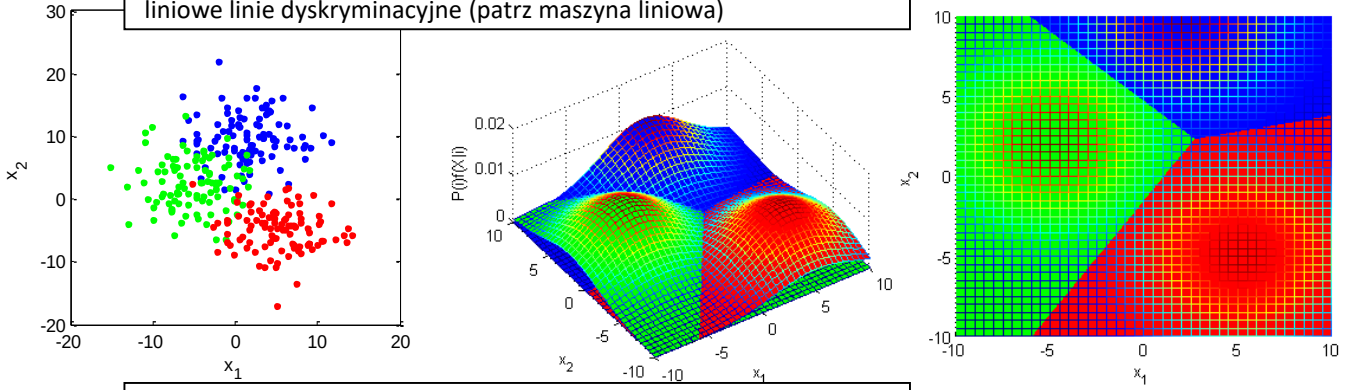
$$g_i(\mathbf{x}) > g_j(\mathbf{x}), \quad i, j = 1,2,...,K, i \neq j \text{, czyli}$$

$$f(\mathbf{x} \mid i)P(i) > f(\mathbf{x} \mid j)P(j), \quad i, j = 1,2,...,K, i \neq j$$

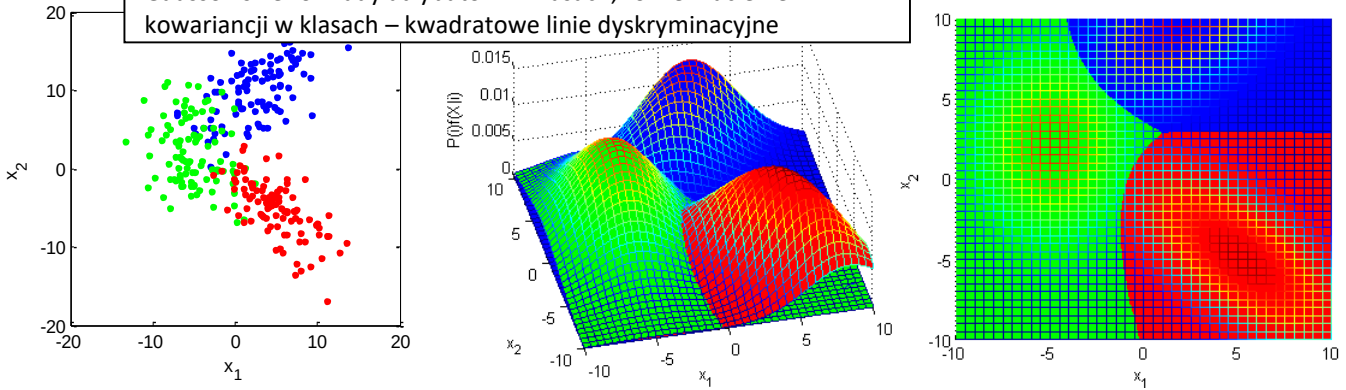


FUNKCJE DYSKRYMINACYJNE DLA REGUŁY BAYESA

Gaussowskie rozkłady atrybutów w klasach, jednakowe wariancje – liniowe linie dyskryminacyjne (patrz maszyna liniowa)



Gaussowskie rozkłady atrybutów w klasach, różne macierze kowariancji w klasach – kwadratowe linie dyskryminacyjne



FUNKCJE DYSKRYMINACYJNE DLA REGUŁY BAYESA

Jeśli atrybuty w klasach mają rozkłady normalne (Gaussa):

$$f(\mathbf{x}\,|\,i) = \frac{1}{(2\pi)^{n/2}\,|\,\Sigma_i\,|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \Sigma_i^{-1}(\mathbf{x}-\mathbf{m}_i)\right)$$

gdzie $\mathbf{m}_i$ – średnia przykładów w klasie *i*, Σ_i, $|\Sigma_i|$ – macierz kowariancji w klasie *i* i jej wyznacznik

Funkcje dyskryminacyjne możemy wyrazić jako $g_i(\mathbf{x}) = \ln\big(f(\mathbf{x}\,|\,i)P(i)\big) = \ln f(\mathbf{x}\,|\,i) + \ln P(i)$. Wtedy ogólny wzór na funkcję dyskryminacyjną reguły Bayesa dla klas gaussowskich:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \Sigma_i^{-1}(\mathbf{x}-\mathbf{m}_i) - \frac{1}{2}\ln|\,\Sigma_i\,| + \ln P(i) \qquad (*)$$

Z powyższego wzoru wynika, że w ogólnym przypadku funkcje te są kwadratowe, co oznacza, że powierzchnie dyskryminacyjne są stopnia drugiego (hiperelipsoidy, hiperparaboloidy).

FUNKCJE DYSKRYMINACYJNE DLA REGUŁY BAYESA

Gdy atrybuty w klasach są statystycznie niezależne i posiadają tę samą wariancję (tzn. $\Sigma_i = \sigma^2 \mathbf{I}$) otrzymujemy liniowe powierzchnie dyskryminacyjne. Gdy dodatkowo przyjmiemy, że *P*(i) dla poszczególnych klas są równe otrzymamy*:

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{m}_i)^T(\mathbf{x}-\mathbf{m}_i) = -\frac{1}{2\sigma^2}\sqrt{\sum_{k=1}^{n}(x_k - m_{i,k})^2}$$

Wartość funkcji dyskryminacyjnej jest w tym przypadku zanegowaną odległością **x** od $\mathbf{m}_i$ (maszyna liniowa).

Powyżej założyliśmy, że dane (wektory liczb rzeczywistych **x**) opisane są wielowymiarowymi rozkładami ciągłymi $f(\mathbf{x}|i)$. Jeśli dane są dyskretne zamiast gęstości stosujemy rozkłady prawdopodobieństw $P(\mathbf{x}|i)$. Wtedy funkcje dyskryminacyjne mają postać:

$$g_i(\mathbf{x}) = P(\mathbf{x}|i)P(i)$$

* Patrz Stąpor K.: Automatyczna klasyfikacja obiektów. Exit, Warszawa 2005, str. 60.

FUNKCJE DYSKRYMINACYJNE DLA REGUŁY BAYESA – PRZYKŁAD

Dane są dwie klasy opisane średnimi: $\mathbf{m}_1 = [-3, -3]^T$ i $\mathbf{m}_2 = [3, 3]^T$ oraz macierzami kowariancji:

$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ i $\Sigma_2 = \begin{bmatrix} 9 & 0 \\ 0 & 9 \end{bmatrix}$. Prawdopodobieństwa wystąpienia obiektów z tych klas wynoszą:

P(1) = P(2) = 0.5. Wyznacz funkcje dyskryminacyjne i powierzchnie decyzyjne klasyfikatora Bayesa.

Korzystamy ze wzoru (*) pomijając $\ln P(i)$.

Obliczamy $\Sigma^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\Sigma^{-1} = \begin{bmatrix} 1/9 & 0 \\ 0 & 1/9 \end{bmatrix}$, $|\Sigma_1| = 1$, $|\Sigma_2| = 81$, stąd

$$g_1(\mathbf{x}) = -\frac{1}{2}\begin{bmatrix} x_1+3 \\ x_2+3 \end{bmatrix}^T \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1+3 \\ x_2+3 \end{bmatrix} - \frac{1}{2}\ln 1 = -\frac{1}{2}[x_1+3, x_2+3] \cdot \begin{bmatrix} x_1+3 \\ x_2+3 \end{bmatrix} - 0 =$$

$$= -\frac{1}{2}\left((x_1+3)\cdot(x_1+3) + (x_2+3)\cdot(x_2+3)\right) = -\frac{1}{2}\left((x_1^2+6x_1+9) + (x_2^2+6x_2+9)\right) = -\frac{1}{2}(x_1^2+x_2^2+6x_1+6x_2+18)$$

$$g_2(\mathbf{x}) = -\frac{1}{2}\begin{bmatrix} x_1-3 \\ x_2-3 \end{bmatrix}^T \cdot \begin{bmatrix} 9 & 0 \\ 0 & 9 \end{bmatrix} \cdot \begin{bmatrix} x_1-3 \\ x_2-3 \end{bmatrix} - \frac{1}{2}\ln 81 = -\frac{1}{18}(x_1^2+x_2^2-6x_1-6x_2+18) - 2\ln 3$$

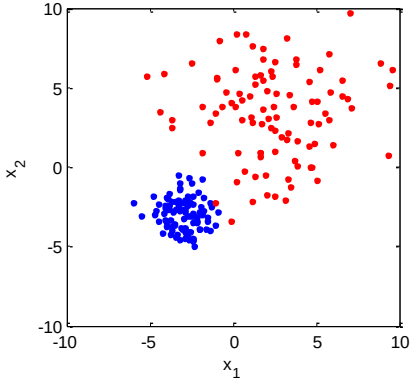
FUNKCJE DYSKRYMINACYJNE DLA REGUŁY BAYESA – PRZYKŁAD

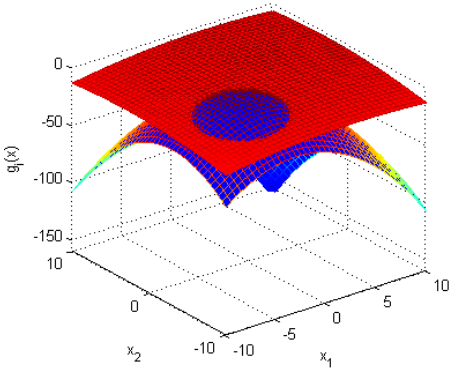
Powierzchnia decyzyjna: $g_1(\mathbf{x}) - g_2(\mathbf{x}) = 0$

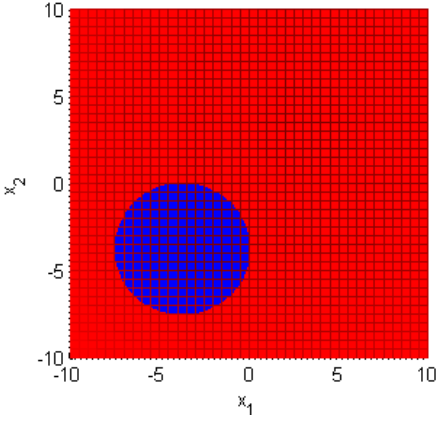
$$-\frac{1}{2}(x_1^2 + x_2^2 + 6x_1 + 6x_2 + 18) + \frac{1}{18}(x_1^2 + x_2^2 - 6x_1 - 6x_2 + 18) + 2\ln 3 = 0 \quad /\cdot(-18)$$

$$8x_1^2 + 8x_2^2 + 60x_1 + 60x_2 + 144 - 36\ln 3 = 0 \quad /:4$$

$$2x_1^2 + 2x_2^2 + 15x_1 + 15x_2 + 36 - 9\ln 3 = 0 \quad \text{- równanie okręgu}$$







EMPIRYCZNY KLASYFIKATOR BAYESA

Klasyfikator Bayesa wymaga znajomości prawdopodobieństwa a priori $P(i)$ poszczególnych klas oraz warunkowych gęstości $f(\mathbf{x}|i)$. W praktyce wielkości te są nieznane i estymuje się je na podstawie zbioru trenującego. W ten sposób powstaje empiryczny klasyfikator Bayesa.

Estymacja $P(i)$ – udział poszczególnych klas w zbiorze uczącym:

$$P(i) = \frac{N_i}{N} \quad i = 1,2,...,K$$

gdzie: N – liczba przykładów w zbiorze uczącym, N_i – liczba przykładów z klasy i w zbiorze uczącym

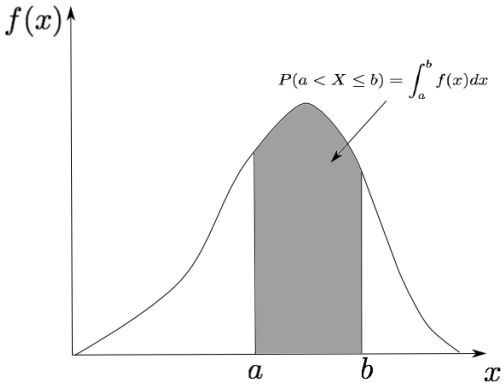
Estymacja $f(\mathbf{x}|i)$:

- podejście parametryczne: znana jest postać funkcyjna $f(\mathbf{x}|i)$, nieznane są parametry tej funkcji. Parametry estymuje się na podstawie zbioru uczącego
- podejście nieparametryczne: brak założeń do postaci $f(\mathbf{x}|i)$. Funkcje te szacuje się metodami nieparametrycznymi na zbiorze uczącym

FUNKCJA GĘSTOŚCI PRAWDOPODOBIEŃSTWA

Funkcja gęstości prawdopodobieństwa *f*(*x*) (dla rzeczywistych zmiennych losowych) – funkcja, która pozwala wyznaczyć prawdopodobieństwo tego, że zmienna losowa *X* przyjmuje wartości z przedziału (a, b]:

$$P(a < X \leq b) = \int_a^b f(x)dx\,.$$



Im zmienna losowa częściej przyjmuje wartości z otoczenia liczby *x*, tym wartość *f*(*x*) jest większa.

Funkcja *f*(*x*) spełnia warunki:

$$f(x) \geq 0 \quad \text{dla każdego } x$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

PARAMETRYCZNA ESTYMACJA FUNKCJI GĘSTOŚCI

Parametryczna procedura estymacji $f(x)$ polega na arbitralnym wyborze typu rozkładu[†] (np. normalny, jednostajny, trójkątny, beta) i ustaleniu wartości jego parametrów.

W przypadku wielowymiarowego rozkładu normalnego szukamy wartości dwóch parametrów dla każdej klasy (niezależnie): wektora średnich oraz macierzy kowariancji:

$$\mathbf{m}_i = \frac{1}{N_i}\sum_{l=1}^{N_i}\mathbf{x}_l^i, \quad i = 1,2,...,K \text{ (numer klasy)}$$

$$\mathbf{\Sigma}_i = \frac{1}{N_i}\sum_{l=1}^{N_i}(\mathbf{x}_l^i - \mathbf{m}_i)(\mathbf{x}_l^i - \mathbf{m}_i)^T, \quad i = 1,2,...,K$$

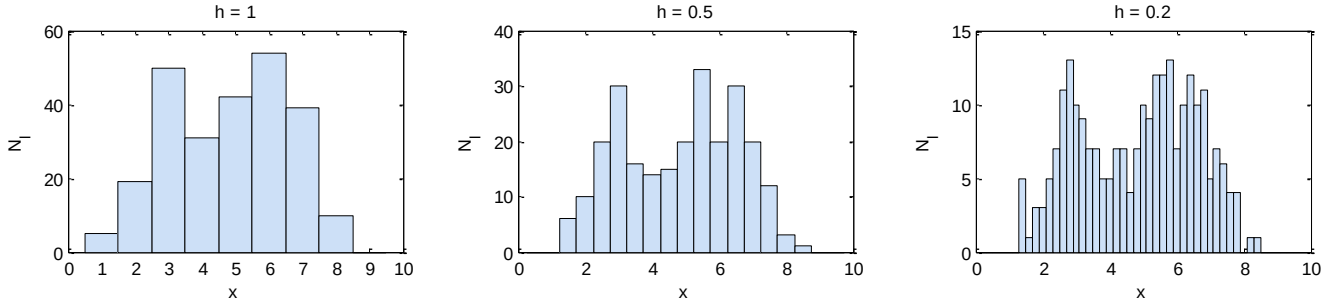
gdzie $\mathbf{x}_l^i$ są przykładami ze zbioru trenującego należącymi do klasy *i*.

Dane występujące w praktyce mogą wyrażać specyficzne rozkłady, do których trudno dopasować rozkłady standardowe. Wtedy stosuje się nieparametryczne metody estymacji gęstości.

[†] Standardowe typy rozkładów możesz podejrzeć w Matlabie uruchamiając narzędzie disttool.

NIEPARAMETRYCZNA ESTYMACJA FUNKCJI GĘSTOŚCI – HISTOGRAM

Najprostszą formą nieparametrycznej estymacji gęstości jest histogram. Aby go utworzyć zakres zmiennej *x* dzielimy na równe przedziały o długości *h*. Następnie zliczamy ile przykładów uczących leży w poszczególnych przedziałach (N_l). Słupek reprezentujący *l*-ty przedział ma wysokość N_l.



Estymator funkcji gęstości otrzymany na podstawie histogramu ma postać:

$$f(x) = \frac{N_l}{Nh}$$

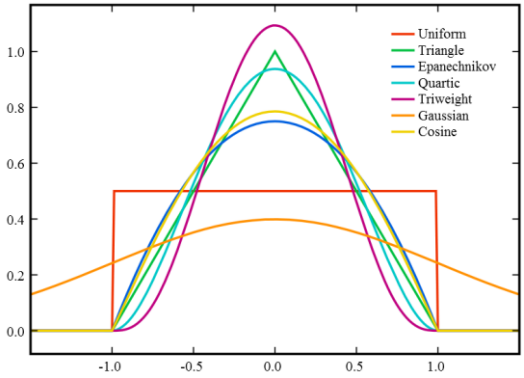
gdzie *l* jest przedziałem do którego należy *x*

Np. gęstość w punkcie x = 5.2 oszacowana na podstawie powyższych trzech histogramów wynosi: 42/(250·1)=0,168, 20/(250·0,5)=0,16 oraz 9/(250·0,2)=0,18

NIEPARAMETRYCZNA ESTYMACJA FUNKCJI GĘSTOŚCI – ESTYMATORY JĄDROWE

Estymacja gęstości za pomocą estymatorów jądrowych polega na „rozpięciu” nad każdym przykładem uczącym x_k pewnej funkcji K, zwanej jądrem, która spełnia warunki:

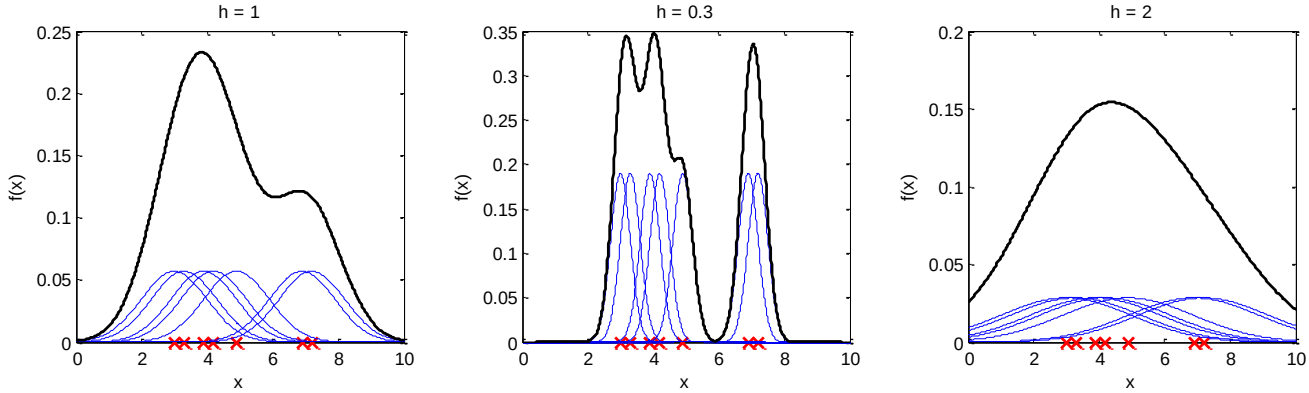
- jest nieujemna,
- osiąga maksimum w zerze,
- jest symetryczna względem zera oraz
- spełnia warunek $\int_{-\infty}^{\infty} K(x)dx = 1$.



Suma funkcji K rozpiętych nad wszystkimi przykładami uczącymi x_k podzielona przez liczbę przykładów i parametr szerokości jądra h daje jądrowy estymator gęstości:

$$f(x) = \frac{1}{Nh} \sum_{k=1}^{N} K\left(\frac{x - x_k}{h} \right)$$

NIEPARAMETRYCZNA ESTYMACJA FUNKCJI GĘSTOŚCI – ESTYMATORY JĄDROWE



Estymator jądrowy jest mało wrażliwy na postać funkcji *K*. Często używa się jądra normalnego:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \quad \rightarrow \quad K\left(\frac{x - x_k}{h}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - x_k)^2}{2h^2}\right)$$

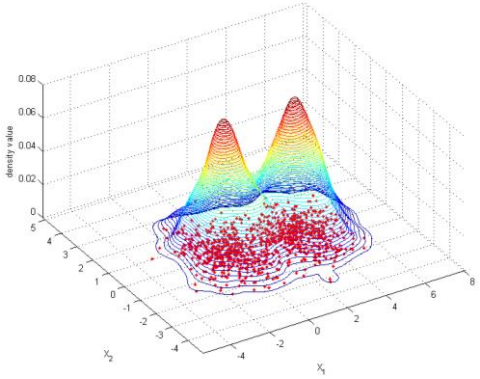
Parametr *h* zwany jest też parametrem wygładzania. Przy dużych wartościach *h* otrzymujemy większe wygładzenie funkcji gęstości. Małe wartości *h* uwydatniają szczegóły zbioru trenującego. W praktyce wartość *h* dobiera się zależnie od liczby przykładów uczących *N* i ich wymiaru *n*.

NIEPARAMETRYCZNA ESTYMACJA FUNKCJI GĘSTOŚCI – ESTYMATORY JĄDROWE

Jeśli *x* jest zmienną wektorową jądrowy estymator gęstości najczęściej przybiera postać sumy jąder produktowych (iloczynów jednowymiarowych jąder dla poszczególnych współrzędnych):

$$f(\mathbf{x}) = \frac{1}{Nh_1 \ldots h_n} \sum_{k=1}^{N} \prod_{l=1}^{n} K\left(\frac{x_l - x_{k,l}}{h_l} \right)$$

gdzie z *l*-tym wymiarem skojarzony jest parametr wygładzania h_l.



Po zastosowaniu jąder normalnych otrzymujemy:

$$f(\mathbf{x}) = \frac{1}{Nh_1 \ldots h_n (2\pi)^{n/2}} \sum_{k=1}^{N} \exp\left(-\sum_{l=1}^{n} \frac{(x_l - x_{k,l})^2}{2h_l^2} \right)$$

W takim przypadku parametry wygładzania można wyznaczyć ze wzoru:

$$h_l = \sigma_l \left(\frac{4}{(n+2)N} \right)^{\frac{1}{n+4}} \approx \sigma_l N^{-\frac{1}{n+4}}$$

gdzie σ_l jest odchyleniem standardowym x_l estymowanym z próby

PODEJŚCIE NIEPARAMETRYCZNE – NAIWNY KLASYFIKATOR BAYESA

Funkcje gęstości określamy dla każdej klasy *i*. Jeśli założyć warunkową niezależność wartości atrybutów przy ustalonej klasie, funkcję gęstości można zapisać:

$$f(\mathbf{x} \mid i) = \prod_{j=1}^{n} f(x_j \mid i)$$

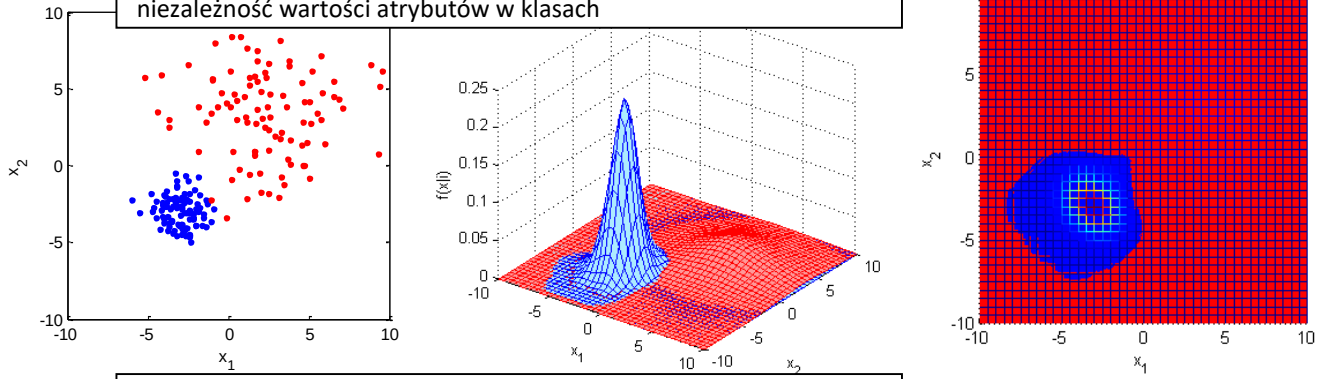
co oznacza estymację $f(\mathbf{x}|i)$ na podstawie funkcji gęstości pojedynczych atrybutów (tzn. dla każdego wymiaru/atrybutu oddzielnie) w obrębie danej klasy. Takie podejście zastosowano powyżej w produktowych estymatorach jądrowych. Możemy zatem zapisać:

$$f(\mathbf{x} \mid i) = \frac{1}{N_i h_1^i \ldots h_n^i} \sum_{k=1}^{N_i} \prod_{l=1}^{n} K\left(\frac{x_l^i - x_{k,l}^i}{h_l^i} \right)$$

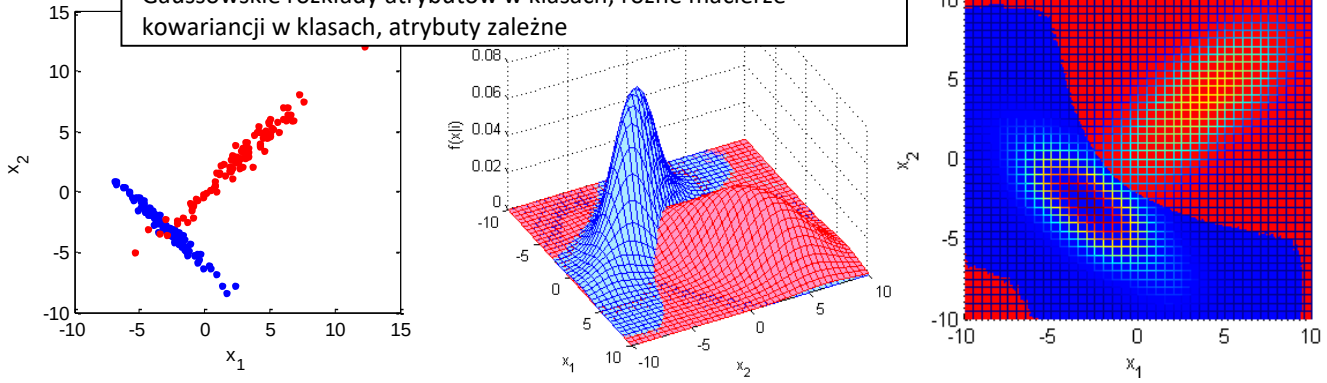
Klasyfikator wykorzystujący takie podejście (zapisanie wielowymiarowej funkcji gęstości jako iloczynów jednowymiarowych funkcji gęstości) nazywa się naiwnym klasyfikatorem Bayesa.

NAIWNY KLASYFIKATOR BAYESA

Gaussowskie rozkłady atrybutów w klasach, warunkowa niezależność wartości atrybutów w klasach

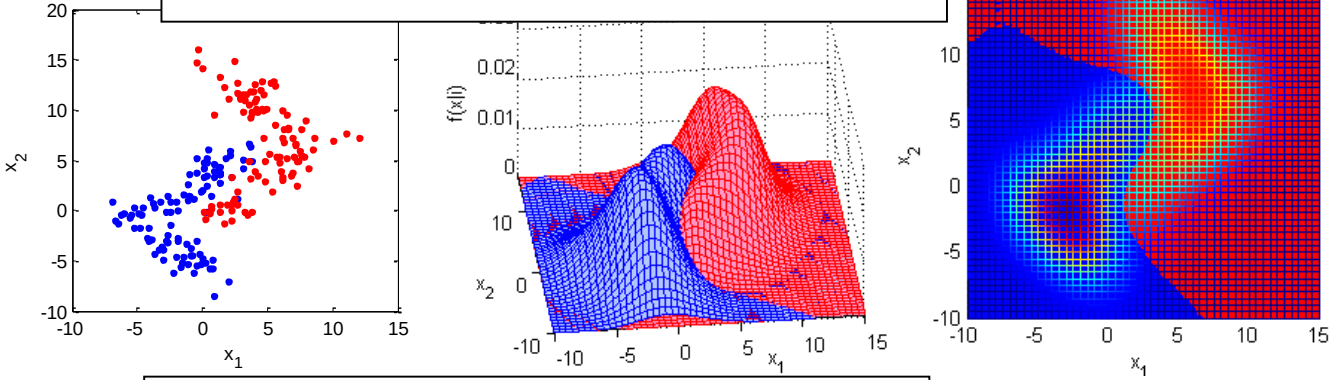


Gaussowskie rozkłady atrybutów w klasach, różne macierze kowariancji w klasach, atrybuty zależne

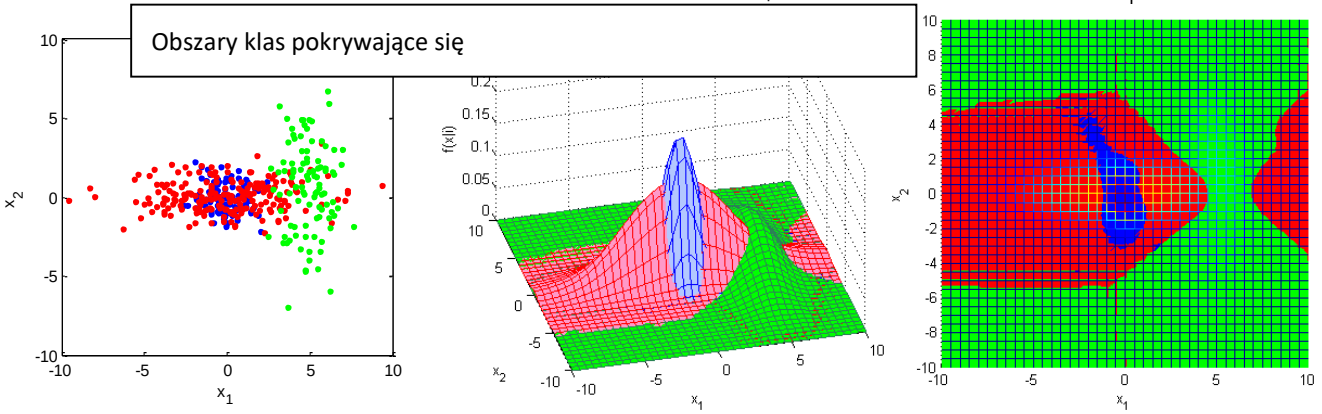


NAIWNY KLASYFIKATOR BAYESA

Niegaussowskie rozkłady atrybutów w klasach



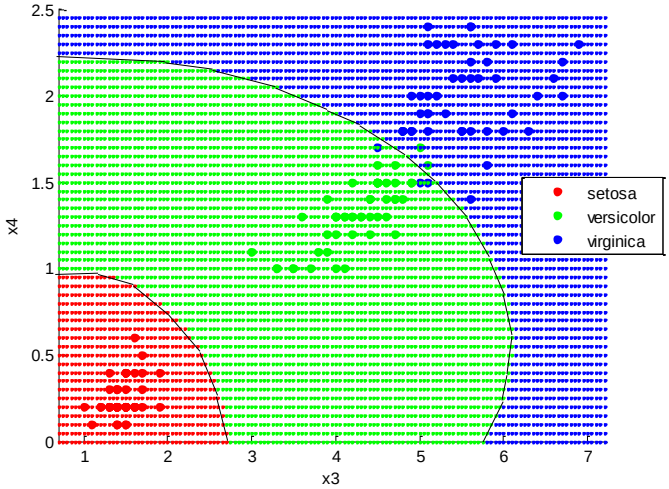
Obszary klas pokrywające się



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**	**46**	**3**
3	**0**	**4**	**47**
nierozpoznana	**0**	**0**	**0**



Odsetek poprawnie sklasyfikowanych przykładów: **95,33%**. Gdy usuniemy dwa pierwsze atrybuty przykładów, pozostawiając atrybut x_3 i x_4 wynik klasyfikacji będzie lepszy: **96,00%**! Linie decyzyjne dla tego przypadku przedstawia rys. powyżej.

OPTYMALNOŚĆ REGUŁY BAYESA

W powyższych rozważaniach przyjęto, że z każdą złą decyzją klasyfikatora (błędną klasyfikacją przykładu z klasy *j* do klasy *i*) związana była jednakowa strata, niezależna od *i* i *j*. Wartość straty odzwierciedla ujemne konsekwencje błędnej decyzji.

Załóżmy teraz, że rozważamy przydzielenie przykładu **x** do klasy *i*. Jeśli jego prawdziwą klasą jest *j* ponosimy stratę $L_{i,j}$. Ponieważ prawdopodobieństwo zdarzenia, że klasą przykładu jest *j* wynosi $P(j|\mathbf{x})$, wobec tego średnia wartość straty jaką ponosimy w przypadku przydzielenia przykładu do klasy *i* wynosi:

$$R(i|\mathbf{x}) = \sum_{j=1}^{K} L_{i,j} P(j|\mathbf{x}) \qquad (**)$$

$R(i|\mathbf{x})$ nazywa się ryzykiem warunkowym.

Naszym celem jest teraz wybranie takiej klasy *i*, dla której osiągamy minimum $R(i|\mathbf{x})$. Klasyfikator działający według tej reguły nazywany jest optymalnym klasyfikatorem bayesowskim. Żaden klasyfikator nie może mieć ryzyka całkowitego mniejszego niż ryzyko całkowite klasyfikatora bayesowskiego.

Jeśli przyjmiemy prostą funkcje straty:

$$L_{i,j} = \begin{cases} 0, & \text{gdy } i = j \\ 1, & \text{gdy } i \neq j \end{cases}$$

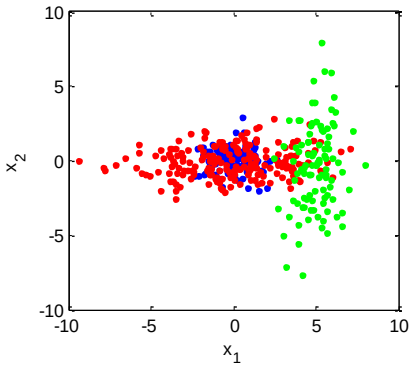
otrzymamy:

$$R(i \mid \mathbf{x}) = \sum_{j=1,\, j \neq i}^{K} P(j \mid \mathbf{x}) = 1 - P(i \mid \mathbf{x})$$

Minimum tego ryzyka osiągane jest dla maksimum prawdopodobieństwa *a posteriori* $P(i \mid \mathbf{x})$ i reguła klasyfikacji sprowadza się do tej rozważanej wcześniej (patrz slajd 3).

KLASYFIKATOR BAYESA – PRZYKŁAD

Lekarz ma zdiagnozować pacjenta na podstawie dwóch symptomów chorobowych x_1 i x_2. Zbiór trenujący obejmujący ludzi chorych na chorobę A (klasa 1 - niebieska), chorych na chorobę B (klas 2 - czerwona) i zdrowych (klasa 3 - zielona) pokazano na rysunku.



Choroba A wymaga wykonania skomplikowanej i kosztownej, ale ratującej życie operacji, natomiast choroba B – prostego leczenia farmakologicznego.

- Błędne zaklasyfikowanie osoby zdrowej do klasy choroby A wiąże się z dużą stratą: $L_{1,3} = 0{,}9$ (zbędna operacja) .
- Błędne zaklasyfikowanie osoby zdrowej do klasy choroby B wiąże się z małą stratą: $L_{2,3} = 0{,}1$ (nieszkodliwe leczenie).
- Błędne zaklasyfikowanie osoby chorej na A do klasy choroby B wiąże się z dużą stratą: $L_{2,1} = 1{,}0$ (brak ratującej życie operacji).

KLASYFIKATOR BAYESA – PRZYKŁAD

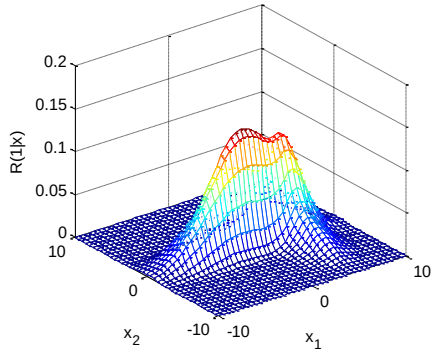
- Błędne zaklasyfikowanie osoby chorej na B do klasy choroby A wiąże się z dużą stratą: $L_{1,2} = 0{,}9$ (zbędna operacja).
- Błędne zaklasyfikowanie osoby chorej na B do klasy osób zdrowych wiąże się z niewielką stratą: $L_{3,2} = 0{,}3$ (brak leczenia).
- Błędne zaklasyfikowanie osoby chorej na A do klasy osób zdrowych wiąże się z dużą stratą: $L_{3,1} = 1{,}0$ (brak ratującej życie operacji).
- Zaklasyfikowanie poprawne nie przynosi straty: $L_{i,i} = 0$

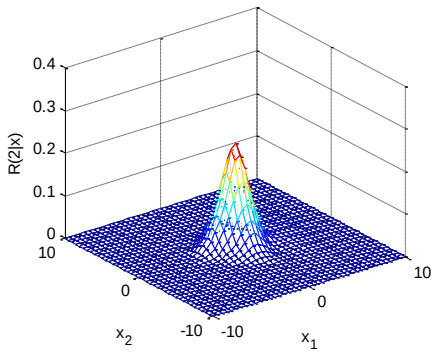
Straty możemy przedstawić w macierzy strat:

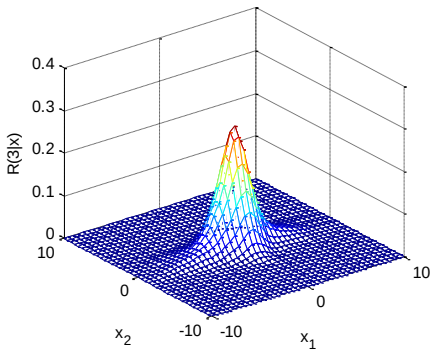
$$\mathbf{L} = \begin{bmatrix} L_{1,1} & L_{1,2} & L_{1,3} \\ L_{2,1} & L_{2,2} & L_{2,3} \\ L_{3,1} & L_{3,2} & L_{3,3} \end{bmatrix} = \begin{bmatrix} 0 & 0{,}9 & 0{,}9 \\ 1 & 0 & 0{,}1 \\ 1 & 0{,}3 & 0 \end{bmatrix}$$

KLASYFIKATOR BAYESA – PRZYKŁAD

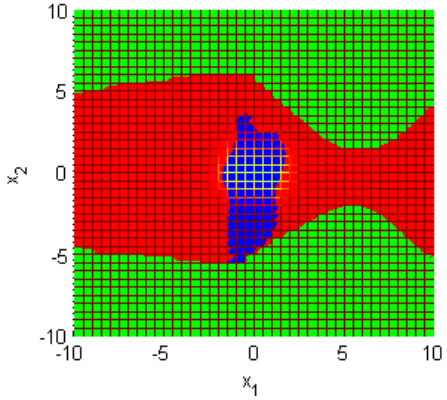
Ryzyka warunkowe wyznaczone dla poszczególnych klas ze wzoru (**) pokazano na rysunkach:







Obszary decyzyjne:



KLASYFIKATOR BAYESA – PRZYKŁAD

Przypuśćmy, że rozwój medycyny pozwolił chorobę A leczyć farmakologicznie zamiast operacyjnie oraz, że leczenie farmakologiczne choroby B przynosi bardzo groźne skutki uboczne. Powoduje to zmianę wartości strat: $L_{1,2} = 0{,}3$, $L_{1,3} = 0{,}2$, $L_{2,3} = 0{,}9$, i $L_{2,1} = 0{,}9$. Macierz strat:

$$\mathbf{L} == \begin{bmatrix} 0 & 0{,}3 & 0{,}2 \\ 0{,}9 & 0 & 0{,}9 \\ 1 & 0{,}3 & 0 \end{bmatrix}$$

Taka zmiana wartości strat wpływa na ryzyka warunkowe i linie decyzyjne separujące klasy: