WYKŁAD 2: Rozpoznawanie obrazów

Teoretyczne wprowadzenie do zadania klasyfikacji.

Optymalny klasyfikator bayesowski. Naiwny klasyfikator Bayesa

Dr inż. Urszula Libal

Teoretyczne wprowadzenie do zadania klasyfikacji

SPIS TREŚCI:

- ❑ Pełna informacja probabilistyczna
- ❑ Funkcja strat
- ❑ Ryzyko
- ❑ Optymalny klasyfikator bayesowski
- ❑ Naiwny klasyfikator Bayesa

1. Pełna informacja probabilistyczna

Rozpoznawany obraz pochodzi z pewnej klasy ze zbioru wszystkich klas

$$\mathcal{M} = \{1, 2, \ldots, M\}\,. \tag{1}$$

Znane są prawdopodobieństwa *a priori* klas oraz funkcje gęstości prawdopodobieństwa w klasach:

klasa 1	klasa 2	$\cdots$	klasa M
p_1	p_2	$\cdots$	p_M
$f_1(x)$	$f_2(x)$	$\cdots$	$f_M(x)$

2. Zero-jedynkowa funkcja strat

Zero-jedynkowa funkcja strat $\mathscr{L}^{0-1}$ przyjmuje dwie wartości

$$\mathscr{L}^{0-1}(C,J) \quad = \quad \begin{cases} 0, & \text{gdy} \quad C = J, \\ 1 & \text{gdy} \quad C \neq J, \end{cases} \tag{2}$$

gdzie C to klasa, do której zaklasyfikowany został obraz pochodzący z klasy J.

3. Ryzyko dla 0-1 funkcji strat

Ryzyko algorytmu Ψ dla dowolnej funkcji strat $\mathscr{L}$ definiujemy następująco

$$\begin{aligned}
\mathscr{R}[\Psi] &= \mathbb{E}\{\mathscr{L}(\Psi(X(\omega)), J(\omega))\} && (3)\\
&= \int_{\mathscr{X}} \sum_{j\in\mathscr{M}} \mathscr{L}(i, j)\, p_j f_j(x)\, dx && (4)\\
&= \sum_{j\in\mathscr{M}} p_j \sum_{i\in\mathscr{M}} \mathscr{L}(i, j) \int_{\mathscr{D}_{\mathscr{X}}^{(i)}} f_j(x)\, dx. && (5)
\end{aligned}$$

4. Problem klasyfikacji dla dwóch klas

Obszary decyzyjne to

$$\mathscr{D}_{\mathscr{X}}^{(1)} = \{x \in \mathscr{X} : \Psi(x) = 1\}, \tag{6}$$

$$\mathscr{D}_{\mathscr{X}}^{(2)} = \{x \in \mathscr{X} : \Psi(x) = 2\}, \tag{7}$$

takie, że

$$\mathscr{D}_{\mathscr{X}}^{(1)} \cup \mathscr{D}_{\mathscr{X}}^{(2)} = \mathscr{X}, \tag{8}$$

$$\mathscr{D}_{\mathscr{X}}^{(1)} \cap \mathscr{D}_{\mathscr{X}}^{(2)} = \emptyset. \tag{9}$$

W przypadku szczególnym, dla dwóch klas $\mathscr{M} = \{1, 2\}$ oraz dla 0-1 funkcji strat, ryzyko wynosi

$$\begin{aligned} \mathscr{R}[\Psi] &= p_1 \mathscr{L}^{0-1}(1, 1) \int_{\mathscr{D}_{\mathscr{X}}^{(1)}} f_1(x)\, dx + p_1 \mathscr{L}^{0-1}(2, 1) \int_{\mathscr{D}_{\mathscr{X}}^{(2)}} f_1(x)\, dx && (10) \\ &+ p_2 \mathscr{L}^{0-1}(1, 2) \int_{\mathscr{D}_{\mathscr{X}}^{(1)}} f_2(x)\, dx + p_2 \mathscr{L}^{0-1}(2, 2) \int_{\mathscr{D}_{\mathscr{X}}^{(2)}} f_2(x)\, dx. && (11) \end{aligned}$$

Po uproszczeniu

$$\mathscr{R}[\Psi] = p_1 \int_{\mathscr{D}_{\mathscr{X}}^{(2)}} f_1(x)\, dx + p_2 \int_{\mathscr{D}_{\mathscr{X}}^{(1)}} f_2(x)\, dx. \qquad (12)$$

Funkcja gęstości prawdopodobieństwa cech pochodzących z obu klas $\{1, 2\}$ to mieszanina gęstości prawdopodobieństw cech w klasach

$$f(x) \quad = \quad p_1 f_1(x) + p_2 f_2(x). \tag{13}$$

Przekształcamy wzór (12)

$$\begin{aligned} \mathscr{R}[\Psi] &= p_1 \int_{\mathscr{X} \setminus \mathscr{D}_{\mathscr{X}}{}^{(1)}} f_1(x)\, dx + p_2 \int_{\mathscr{X} \setminus \mathscr{D}_{\mathscr{X}}{}^{(2)}} f_2(x)\, dx && (14) \\ &= \int_{\mathscr{X}} f(x) - p_1 \int_{\mathscr{D}_{\mathscr{X}}{}^{(1)}} f_1(x)\, dx - p_2 \int_{\mathscr{D}_{\mathscr{X}}{}^{(2)}} f_2(x)\, dx && (15) \\ &= 1 - P_c[\Psi] && (16) \\ &= P_e[\Psi]. && (17) \end{aligned}$$

5. Algorytm bayesowski

Algorytm bayesowski Ψ^* (optymalny) minimalizuje ryzyko średnie

$$\mathscr{R}[\Psi^*] = \min_{\Psi} \mathscr{R}[\Psi]. \tag{18}$$

Dla dwóch klas oraz dla 0-1 funkcji strat minimalizacja ryzyka algorytmu Ψ^* oznacza minimalizację średniego prawdopodobieństwa błędnej klasyfikacji P_e, ponieważ

$$\mathscr{R}[\Psi^*] = P_e[\Psi^*] = p_1 \int_{\mathscr{D}_{\mathscr{X}}^{*(2)}} f_1(x)\,dx + p_2 \int_{\mathscr{D}_{\mathscr{X}}^{*(1)}} f_2(x)\,dx. \tag{19}$$

Jednocześnie warunek ten oznacza maksymalizację średniego prawdopodobieństwa poprawnej klasyfikacji P_c, ponieważ

$$\mathscr{R}[\Psi^*] = 1 - P_c[\Psi^*] = 1 - \left(p_1 \int_{\mathscr{D}_{\mathscr{X}}^{*(1)}} f_1(x)\,dx + p_2 \int_{\mathscr{D}_{\mathscr{X}}^{*(2)}} f_2(x)\,dx \right). \tag{20}$$

Z twierdzenia Bayesa prawdopodobieństwo *a posteriori* klasy $k \in \{1, 2\}$, gdy zaobserwowano x wynosi

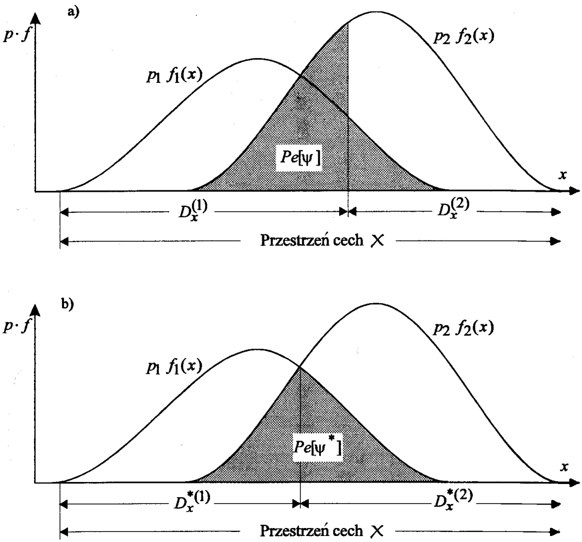
$$P(k\,|\,x) = \frac{p_k f_k(x)}{\sum_{i \in \mathscr{M}} p_i f_i(x)}, \tag{21}$$

gdzie

$$f(x) \quad = \quad p_1 f_1(x) + p_2 f_2(x). \tag{22}$$

Algorytm bayesowski sprowadza się do reguły

$$\Psi^*(x) = \begin{cases} 1, & \text{gdy } p_1 f_1(x) > p_2 f_2(x), \\ 2, & \text{gdy } p_1 f_1(x) < p_2 f_2(x). \end{cases} \tag{23}$$



Rysunek 1. Prawdopodobieństwo błędnej klasyfikacji: a) dowolnego algorytmu, b) algorytmu optymalnego (bayesowskiego).

Źródło: M. Kurzyński, "Rozpoznawanie obiektów. Metody statystyczne", (1997).

7. Oszacowanie ryzyka dla M klas

Dla algorytmu bayesowskiego w przypadku rozłączności nośników funkcji gęstości

$$P_e[\Psi^*] = 0\,. \tag{24}$$

W przypadku $p_1 f_1(x) = p_2 f_2(x)$ dla $M = 2$ klas

$$P_e[\Psi*] = 0.5\,. \tag{25}$$

W problemie M-klasowym zachodzi

$$0 \leq P_e[\Psi*] \leq \frac{M-1}{M}\,. \tag{26}$$

Zadanie

Zad. Wyznacz punkt graniczny oraz ryzyko dla bayesowskiego algorytmu rozpoznawania obrazów, jeżeli funkcja gęstości prawdopodobieństwa cech w klasie 1 wynosi f_1, w klasie 2 - f_2, a p_1 i p_2 to prawdopodobieństwa *a priori* wystąpienia klas.

Dane:

$f_1(x) = (-4x+3)\,1\left[0,\,\frac{3}{4}\right]$

$f_2(x) = (2x)\,1\,[0,\,1]$

a) $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{2}$

b) $p_1 = \frac{2}{3}$, $p_2 = \frac{1}{3}$

Naiwny klasyfikator Bayesa

1. Wektory cech

Rozpoznawanie D-wymiarowych wektorów cech

$$\mathbf{x} = (x_1, x_2, \ldots, x_D)\,.$$

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
		⋮		
5.9	3.0	5.1	1.8	*I. virginica*

Rysunek 1. Zbiór danych *iris*: 4 cechy (długość i szerokość działki kielicha, długość i szerokość płatka), 3 klasy (gatunki irysa).


Iris setosa


Iris versicolor


Iris virginica

Rysunek 2. Klasyfikacja do 3 klas określonych przez gatunek irysa.

2. Klasyfikator Bayesa - przypadek wielowymiarowy

W przypadku obrazów opisanych przez D-wymiarowe wektory cech

$$\mathbf{x} = (x_1, x_2, \ldots, x_D)\,, \tag{1}$$

klasyfikator Bayesa wskazuje na klasę $i \in \mathscr{M}$

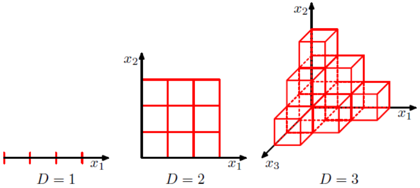
$$\Psi^*(\mathbf{x}) = i, \;\; \text{jeżeli } p_i f_i(\mathbf{x}) = \max_{k \in \mathscr{M}} p_k f_k(\mathbf{x}). \tag{2}$$

($\mathscr{M}$ - zbiór klas)

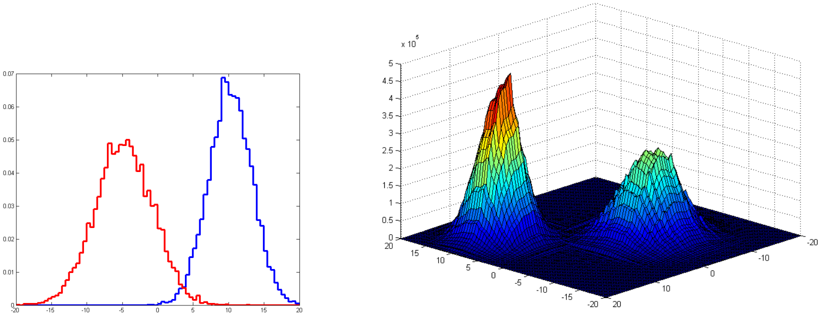
3. Estymacja funkcji gęstości a przekleństwo wymiarowości

Przekleństwo wymiarowości (inaczej zwane *zjawiskiem pustej przestrzeni*)

- związane jest z wykładniczym wzrostem liczby D-wymiarowych kostek, stanowiących podział przestrzeni cech podczas nieparametrycznej estymacji funkcji gęstości, przy zwiększaniu rozmiaru D wektora cech.



Rysunek 3. Ilustracja przekleństwa wymiarowości.

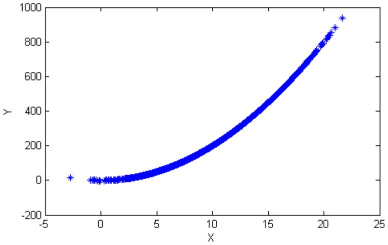


Rysunek 4. Nieparametryczna estymacja funkcji gęstości dla liczby cech $D = 1$ oraz $D = 2$.

4. Naiwny klasyfikator Bayesa

Naiwny klasyfikator Bayesa Ψ_{NB} to klasyfikator Bayesa Ψ^*,

dla którego zakłada się, że cechy $X_1, X_2, \ldots, X_D$ są wzajemnie niezależne!



Rysunek 5. Przykład zmiennych losowych zależnych X i $Y = 2X^2 - 1$.

Definicja 1. Zmienne losowe $X_1, X_2, \ldots, X_D$ są *niezależne* wtedy i tylko wtedy, gdy

$$P\{X_1 < x_1, X_2 < x_2, \ldots, X_D < x_D\} = P\{X_1 < x_1\} P\{X_2 < x_2\} \ldots P\{X_D < x_D\}, \qquad (3)$$

czyli

$$F_{X_1, X_2, \ldots, X_D}(\mathbf{x}) = \prod_{d=1}^{D} F_{X_d}(x_d). \qquad (4)$$

Definicja 2. Zmienne losowe $X_1, X_2, \ldots, X_D$ są *niezależne* wtedy i tylko wtedy, gdy

$$f_{X_1, X_2, \ldots, X_D}(\mathbf{x}) = \prod_{d=1}^{D} f_{X_d}(x_d). \qquad (5)$$

Naiwny klasyfikator Bayesa wskazuje na klasę $i \in \mathscr{M}$ na podstawie zaobserwowanego wektora cech $\mathbf{x} = (x_1, x_2, \ldots, x_D)$

$$\Psi_{NB}(\mathbf{x}) = i, \quad \text{jeżeli } p_i \prod_{d=1}^{D} f_i^{(d)}(x_d) = \max_{k \in \mathscr{M}} p_k \prod_{d=1}^{D} f_k^{(d)}(x_d). \tag{6}$$

Zasada działania pozostaje identyczna jak dla klasyfikatora Bayesa, tzn. maksymalizowane jest prawdopodobieństwo *a posteriori* - patrz wzór (21) z wykładu nr 1. Zakładając niezależność cech otrzymujemy, że funkcja gęstości f_k łącznego rozkładu w klasie $k \in \mathscr{M}$ to iloczyn gęstości brzegowych $f_k^{(d)}, d = 1, 2, \ldots, D,$

$$f_k(\mathbf{x}) = \prod_{d=1}^{D} f_k^{(d)}(x_d). \tag{7}$$

5. Naiwny klasyfikator Bayesa - przypadek dwóch klas

Naiwny klasyfikator Bayesa na podstawie zaobserwowanego wektora cech

$\mathbf{x} = (x_1, x_2, \ldots, x_D)$ wskazuje na klasę

$$\Psi_{NB}(\mathbf{x}) = \begin{cases} 1, & \text{gdy } p_1 \prod_{d=1}^{D} f_1^{(d)}(x_d) > p_2 \prod_{d=1}^{D} f_2^{(d)}(x_d), \\ 2, & \text{w przeciwnym wypadku.} \end{cases} \tag{8}$$

Warunek

$$p_1 \prod_{d=1}^{D} f_1^{(d)}(x_d) > p_2 \prod_{d=1}^{D} f_2^{(d)}(x_d) \tag{9}$$

można przekształcić na warunek równoważny:

$$\frac{p_1 \prod_{d=1}^{D} f_1^{(d)}(x_d)}{p_2 \prod_{d=1}^{D} f_2^{(d)}(x_d)} > 1 \tag{10}$$

$$\ln \frac{p_1 \prod_{d=1}^{D} f_1^{(d)}(x_d)}{p_2 \prod_{d=1}^{D} f_2^{(d)}(x_d)} > \ln 1 \tag{11}$$

$$\ln \frac{p_1}{p_2} + \ln \prod_{d=1}^{D} \frac{f_1^{(d)}(x_d)}{f_2^{(d)}(x_d)} > 0 \tag{12}$$

$$\ln \frac{p_1}{p_2} + \sum_{d=1}^{D} \ln \frac{f_1^{(d)}(x_d)}{f_2^{(d)}(x_d)} > 0 \tag{13}$$

Wyrażenie

$$\delta(\mathbf{x}) = \ln \frac{p_1}{p_2} + \sum_{d=1}^{D} \ln \frac{f_1^{(d)}(x_d)}{f_2^{(d)}(x_d)} \tag{14}$$

będziemy nazywać funkcją dyskryminacyjną między klasami 1 i 2.

Wtedy naiwny klasyfikator Bayesa można zapisać

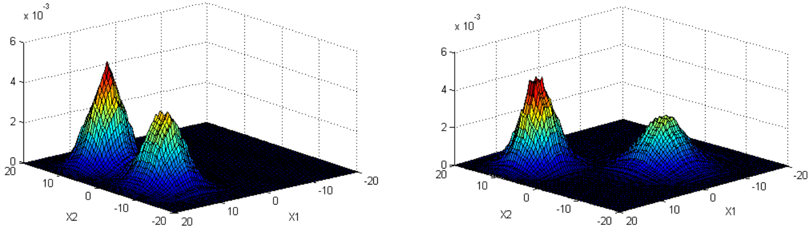
$$\Psi_{NB}(\mathbf{x}) = \begin{cases} 1, & \text{gdy } \delta(\mathbf{x}) > 0, \\ 2, & \text{w przeciwnym wypadku.} \end{cases} \tag{15}$$

Przykład: 2 klasy $\{1,2\}$ i 2 cechy $\mathbf{x} = (x_1, x_2)$,

prawdopodobieństwa a priori równe $p_1 = p_2 = 0.5$

(a) $f_1^{(1)} \sim \mathcal{N}(10,3)$, $f_1^{(2)} \sim \mathcal{N}(10,3)$, $f_2^{(1)} \sim \mathcal{N}(10,3)$, $f_2^{(2)} \sim \mathcal{N}(-5,4)$

(b) $f_1^{(1)} \sim \mathcal{N}(10,3)$, $f_1^{(2)} \sim \mathcal{N}(10,3)$, $f_2^{(1)} \sim \mathcal{N}(-5,4)$, $f_2^{(2)} \sim \mathcal{N}(-5,4)$



a) b)

Rysunek 6. Przykład dyskryminacji między klasami.

W przypadku (a) $p_1 = p_2 = 0.5$ oraz $f_1^{(1)}(x_1) = f_2^{(1)}(x_1)$ dla każdego x_1. Wtedy funkcja dyskryminacyjna otrzymuje postać

$$\begin{aligned}
\delta(\mathbf{x}) &= \ln\frac{p_1}{p_2} + \sum_{d=1}^{D} \ln\frac{f_1^{(d)}(x_d)}{f_2^{(d)}(x_d)} && (16)\\
&= \ln\frac{f_1^{(2)}(x_2)}{f_2^{(2)}(x_2)}, && (17)
\end{aligned}$$

ponieważ $\ln\frac{p_1}{p_2} = 0$ oraz $\ln\frac{f_1^{(1)}(x_1)}{f_2^{(1)}(x_1)} = 0$. Dyskryminacja między klasami odbywa się jedynie na podstawie wartości funkcji gęstości dla drugiej cechy x_2.

W przypadku (b) $p_1 = p_2 = 0.5$, więc $\ln \frac{p_1}{p_2} = 0$. Wtedy funkcja dyskryminacyjna ma postać

$$\delta(\mathbf{x}) \quad = \quad \ln \frac{f_1^{(1)}(x_1)}{f_2^{(1)}(x_1)} + \ln \frac{f_1^{(2)}(x_2)}{f_2^{(2)}(x_2)}. \tag{18}$$

Dyskryminacja między klasami odbywa się na podstawie wartości funkcji gęstości dla obu cech x_1 i x_2.