Zadanie uczenia.

Podział metod uczenia

- Uczenie pod nadzorem – w przypadku gdy zbiór uczący składa się z wektorów cech (atrybutów) oraz wektorów odpowiedzi.
    - wektor wejściowy – wektor cech (zmiennych, atrybutów) opisujących
    - wektor wyjściowy – wektor zmiennych opisywanych
    - <u>cel uczenia</u> – nauczenie systemy odpowiedzi na wektory wejściowe
    - <u>Przykład</u> – zadanie klasyfikacji. Wyjściem etykieta klasy

Zadanie uczenia.

Podział metod uczenia

- Uczenie nienadzorowane – zbiór uczący składający się tylko z wejściowych wektorów cech
    - <u>cel uczenia</u> – opisanie, objaśnienie zbioru danych wejściowych tylko na podstawie ich samych. Wykrycie wewnętrznej struktury danych, wykrycie współzależności
    - <u>Przykład</u> – zadanie grupowania danych w rozłączne klasy (klasyfikacja nienadzorowana)

Zadanie uczenia klasyfikacji…

<u>Założenia:</u>

- Dysponujemy N niezależnymi obserwacjami (próbkami) pochodzącymi z k populacji (klas, grup)
- Wszystkie obserwacje są wektorami losowymi (wektory cech) o tym samym skończonym wymiarze p

<u>Zadanie klasyfikacji polega na przypisaniu obiektowi numeru klasy na podstawie jego wektora cech.</u>

Zadanie uczenia klasyfikacji…

<u>Definicje</u>

- Zbiór obserwacji (próbek) nazywamy *zbiorem wektorów uczących* - oznaczamy symbolem V
- Każdy element zbioru V opisywany jest przez zestaw atrybutów (cech) $a_1, \dots, a_p$ ze zbioru A oraz przez klasę, którą reprezentuje
- Atrybuty tworzą tak zwaną *przestrzeń cech*

Zadanie uczenia klasyfikacji…

- Zakładamy, że przestrzeń cech jest podzbiorem R^p, a każdy obiekt reprezentowany jest p wymiarowym wektorem cech
- Zakładamy, że klasę identyfikujemy poprzez etykietę oznaczającą jej numer:
- Przy powyższych założeniach

  $V = \{(x_1, y_1), \dots, (x_N, y_N)\}$, gdzie $x_i = [x_{i1}, \dots, x_{im}]^T$,

  $y_i \in \{1, \dots, k\}$ dla $i = 1, \dots, N$,

Zadanie uczenia klasyfikacji…

<u>Definicja:</u>

Algorytmem klasyfikacji (regułą decyzyjną) nazywamy funkcję odwzorowującą przestrzeń cech w zbiór numerów klas, to znaczy (przy przyjętych założeniach) funkcję:

$$\psi : R^p \to \{1, \dots, K\}$$

taką, że $\psi(x) = i$ gdy wektor x jest elementem klasy i.

- Każda próbka może należeć tylko i wyłącznie do jednej klasy

Zadanie uczenia klasyfikacji…

- Reguła decyzyjna generuje rozkład przestrzeni cech na tak zwane obszary decyzyjne $D_x^i = \{ x \in R^p : \psi(x) = i \}$

  o następujących własnościach:

  1. $D_x^i \cap D_x^j = \Phi$ dla $i, j \in \{1, \ldots, k\}$, $i \neq j$, co znaczy, że żaden obiekt nie może należeć do dwóch klas jednocześnie
  2. $\bigcup_{i \in \{1,\ldots,k\}} D_x^i = R^p$, co znaczy, że każdy obiekt musi należeć do jednej z klas.

Zadanie uczenia klasyfikacji…

Definicja:

Funkcja klasyfikującą (dyskryminującą) *i-tej klasy* nazywamy funkcję odwzorowujących przestrzeń cech w zbiór liczb rzeczywistych $g_i : R^p \to R$, taką, że:

$$\underset{i \in \{1,\ldots,k\}}{\forall} \; \underset{x \in D_x^i}{\forall} \; g_i(x) = \max_{j \in \{1,\ldots,k\}} g_j(x)$$

- Definicja oznacza, że, na *i*-tym obszarze decyzyjnym wartość *i*-tej funkcji klasyfikującej jest największa.

Zadanie uczenia klasyfikacji…

- Zadanie klasyfikacji sprowadza się w ten sposób do przydzielenia wektora cech *x* do klasy, dla której odpowiednia funkcja klasyfikująca na wektorze *x* osiąga największą wartość:

$$\psi(x) = i \text{, gdy } g_i(x) = \max_{j \in \{1,\ldots,k\}} g_j(x)$$

- wystarczy więc wyliczyć wartości wszystkich funkcji klasyfikujących, znaleźć wartość największą, a jako wynik działania algorytmu zwrócić indeks funkcji dla której ta wartość została osiągnięta.
- **Uczenie klasyfikacji** – polega na znalezieniu postaci funkcji klasyfikujących

Przykłady zadania klasyfikacji

- Wspomaganie diagnostyki medycznej
  - klasyfikacja pacjenta ze względu na jednostkę chorobową
  - określenie przynależności pacjenta do grupy o podwyższonym ryzyku
- Rozpoznawanie cyfr / pisma
- Określenie zdolności kredytowej klienta banku
- Rozpoznawanie niechcianych wiadomości elektronicznych (spamu)

Optymalny klasyfikator statystyczny

- W praktyce zadanie klasyfikacji na ogół nie jest jednoznaczne, ponieważ zdarza się, że obiekty należące do dwóch różnych klas reprezentowane są przez ten sam wektor cech.
- Klasyfikowany obiekt przydziela się do klasy w której prawdopodobieństwo jego wystąpienia jest największe **-optymalny klasyfikator statystyczny**

Optymalny klasyfikator statystyczny

Model matematyczny zadania klasyfikacji:

Zakładamy, że wektor $x = (x_1, \ldots, x_p)^T \in R^p$ oraz numer klasy *j* do której należy jest realizacją pary zmiennych losowych (X, J),

$X : \Omega \to R^p$, gdzie Ω przestrzeń klasyfikowanych obserwacji,

J - dyskretna zmienna losową o wartościach w zbiorze $\{1, \ldots, K\}$

Przykład:

$$x = \begin{bmatrix} 0,7875 \\ 1,561 \\ -2,432 \end{bmatrix} \qquad P(J = j \mid X = x) = ??$$

klasyfikowany obiekt **wektor cech** **prawdopodobieństwa przynależności do klas**

Prawdopodobieństwo całkowite, wzór Bayesa.

Twierdzenie (o prawdopodobieństwie całkowitym)

Niech $A, B_1, ..., B_n \in \Im \subset \Omega$, $P(B_i) > 0$ dla $i = 1, ..., n$,

$B_1 \cup ... \cup B_n = \Omega$ oraz $B_i \cap B_j = \phi$ dla $i \neq j$. Wtedy:

$$P(A) = P(A \mid B_1)P(B_1) + ... + P(A \mid B_n)P(B_n).$$

Prawdopodobieństwo $P(A)$ nazywamy całkowitym (zupełnym) prawdopodobieństwem zajścia zdarzenia *A*.

Prawdopodobieństwo całkowite, wzór Bayesa.

Twierdzenie (Bayesa)

Niech $A, B_1, ..., B_n \in \Im \subset \Omega$, $P(B_i) > 0$ dla $i = 1, ..., n$,

$B_1 \cup ... \cup B_n = \Omega$ oraz $B_i \cap B_j = \phi$ dla $i \neq j$. Wtedy:

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A)}.$$

Uwaga: Na mocy twierdzenia o prawdopodobieństwie całkowitym:

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A \mid B_1)P(B_1) + ... + P(A \mid B_n)P(B_n)}$$

Prawdopodobieństwo całkowite, wzór Bayesa.

Przykład:

Jeden z testów na obecność w organizmie wirusa HCV daje pozytywny rezultat w 97% przypadków osób zarażonych tą choroba i mylnie wskazuje na jego obecność u osób zdrowych w 3% przypadków. Ile wynosi prawdopodobieństwo, że osoba, u której wykryto tym testem obecność wirusa HCV, jest faktycznie chora, jeśli wiadomo, ze u 2% populacji stwierdza się tego wirusa.

Zmienna losowa

Definicja

Niech $(\Omega, \Im, P)$ będzie dowolną przestrzenią probabilistyczną.

Zmienną losową nazywamy każdą funkcję *X* określoną na przestrzeni zdarzeń elementarnych Ω o wartościach rzeczywistych, taką że każdemu przedziałowi wartości zmiennej *X* postaci $(-\infty, x)$ odpowiada zdarzenie losowe, to znaczy $\{\omega : X(\omega) < x\} \in \Im$ dla każdego $x \in R$

Gdy przestrzeń zdarzeń elementarnych jest skończona, a każdy jej podzbiór jest zdarzeniem elementarnym to każda funkcja $X : \Omega \to R$ jest zmienną losową.

Zmienna losowa

Oznaczenia

- $P(X = x)$ - prawdopodobieństwo, że zmienna losowa *X* przyjmuje wartość *x*, czyli $P(X = x) = P\{\omega \in \Omega : X(\omega) = x\}$
- $P(X < x)$ - prawdopodobieństwo, że zmienna losowa *X* przyjmuje wartości mniejsze od *x*.
- $P(X \geq x)$ - prawdopodobieństwo, że zmienna losowa *X* przyjmuje wartości większe lub równe *x*.
- $P(a \leq X \leq b)$ - prawdopodobieństwo, że zmienna losowa *X* przyjmuje wartości z przedziału $[a, b]$

Zmienna losowa

Uwaga

- Zmienną losową, która przyjmuje skończoną bądź przeliczalną liczbę wartości nazywamy zmienną losową dyskretną (typu dyskretnego)
- Zmienną losową, która przyjmuje wartości z pewnego przedziału nazywamy zmienną losową ciągłą (typu ciągłego)

Zmienna losowa

Definicja

Niech $X:\Omega \to R$ będzie zmienną losową. Funkcję $F:R\to[0,1]$ określoną wzorem $F(x)=P(X<x)$ nazywamy dystrybuantą zmiennej losowej *X*.

Własności:

1. $0\le F(x)\le 1$ dla każdego $x\in R$
2. $\lim\limits_{x\to-\infty} F(x)=0$ oraz $\lim\limits_{x\to\infty} F(x)=1$
3. $F(x)$ jest funkcją niemalejącą
4. $F(x)$ jest funkcją lewostronnie ciągłą, to znaczy $\lim\limits_{x\to x_0^-} F(x)=F(x_0)$
5. $P(a\le X<b)=P(X\in[a,b))=F(b)-F(a)$
6. $P(X=x_0)=\lim\limits_{x\to x_0^+} F(x)-F(x_0)$
7. Każda funkcja spełniająca własności 2-4 jest dystrybuantą pewnej zmiennej losowej.

Zmienna losowa

Zmienna losowa dyskretna

- Zmienna losowa, która przyjmuje skończoną bądź przeliczalną liczbę wartości $x_1,\dots,x_n,\dots$
- $P(X=x_i)=p_i>0$ - to znaczy p_i oznacza prawdopodobieństwo przyjęcia przez zmienną losową *X* wartości x_i
- $\sum\limits_i p_i=1$

Zmienna losowa

Definicja

Niech *X* będzie zmienną losową dyskretną. Funkcję *p* określoną na zbiorze wartości zmiennej losowej określoną równością $p(x_i)=P(X=x_i)=p_i$ nazywamy ***funkcją rozkładu prawdopodobieństwa*** zmiennej losowej *X* lub krócej ***rozkładem*** *zmiennej losowej X*.

Równoważnie rozkład prawdopodobieństwa dyskretnej zmiennej losowej podaje się najczęściej w postaci tablicy

x_i	x_1	x_2	…	x_n	…
p_i	p_1	p_2	…	p_n	…

Oczywiście dla dowolnego rozkładu $\sum\limits_i x_i$

Zmienna losowa

Definicja

Zmienną losową *X* przyjmującą wszystkie wartości z pewnego przedziału liczbowego, bądź przedziałów, dla której istnieje nieujemna funkcja $f:R\to R^+\cup\{0\}$ taka, że dystrybuantę zmiennej losowej *X* przedstawić można w postaci

$F(x)=\int\limits_{-\infty}^{x} f(t)\,dt$ nazywamy ***zmienną losową typu ciągłego***,

funkcję *f* natomiast nazywamy ***gęstością rozkładu*** zmiennej losowej *X*.

Zmienna losowa

Zmienna losowa typu ciągłego. Definicja.

Mówimy że dany jest rozkład zmiennej losowej *X* typu ciągłego gdy dana jest jej funkcja gęstości bądź dystrybuanta.

Własności

1. $F'(x)=f(x)$
2. $\int\limits_{-\infty}^{\infty} f(x)=1$
3. $P(a\le X\le b)=F(b)-F(a)=\int\limits_a^b f(x)\,dx$
4. zarówno funkcja gęstości oraz dystrybuanta są funkcjami ciągłymi

Zmienna losowa

Charakterystyki liczbowe zmiennej losowej

- Opisują takie własności zmiennej losowej jak wartość najbardziej prawdopodobna, rozrzut wartości, kształt histogramu czy krzywej gęstości
- Podstawowe charakterystyki liczbowe zmiennej losowej
  - Wartość oczekiwana (przeciętna)
  - Wariancja
  - Odchylenie standardowe

Zmienna losowa

Charakterystyki liczbowe zmiennej losowej

1. Wartość oczekiwana zmiennej losowej *X*, oznaczana $EX = m$, określa wartość wokół której skupiają się realizacje zmiennej losowej uzyskiwane w wyniku wielokrotnego powtarzania tego samego eksperymentu.
2. Niech *X* będzie zmienną losową dyskretną. Wówczas jej wartość oczekiwana jest równa $EX = \sum_i x_i p_i$.
3. Niech *X* będzie zmienną losową ciągłą. Wówczas jej wartość oczekiwana wyraża się wzorem $EX = \int_R x f(x)\,dx$, gdzie $f(x)$ jest gęstością.

Zmienna losowa

Własności wartości oczekiwanej

1. $E(C) = C$, gdzie *C* jest stałą
2. $E(CX) = CE(X)$
3. $E(X \pm Y) = E(X) \pm E(Y)$
4. $E(XY) = E(X) \cdot E(Y)$ gdy *X*, *Y* są zmiennymi niezależnymi.

Definicja.

Zmienne losowe *X*, *Y* określone na tej samej przestrzeni zdarzeń elementarnych nazywamy niezależnymi gdy dla dowolnych x, y niezależne są zdarzenia $\{X < x\}$ oraz $\{Y < y\}$ tzn. gdy:

$$P(X < x \wedge Y < y) = P(X < x) \cdot P(Y < y)$$

Zmienna losowa

Charakterystyki liczbowe zmiennej losowej

1. Wariancja zmiennej losowej oznaczana D^2X jest miarą rozproszenia wokół wartości średniej, wzrasta wraz ze wzrostem rozproszenia kolejnych realizacji zmiennej losowej.
2. Wariancję zmiennej losowej określamy za pomocą następujących wzorów:

   $D^2X = E(X - EX)^2$

   $D^2X = E(X^2) - (EX)^2$ (Dowód. ćwiczenie)

- W przypadku dyskretnym: $D^2X = \sum_i x_i^2 p_i - (EX)^2$
- W przypadku ciągłym: $D^2X = \int_R x_i^2 f(x)\,dx - (EX)^2$

Zmienna losowa

Własności wariancji

1. $D^2(C) = 0$, gdzie *C* jest stałą
2. $D^2(CX) = C^2 D^2(X)$
3. $D^2(X \pm C) = D^2(X)$
4. $D^2(X \pm Y) = D^2(X) + D^2(Y)$ gdy *X*, *Y* są zmiennymi niezależnymi.

Zmienna losowa

Wybrane rozkłady ciągłe

1. Rozkład normalny:
   Mówimy, że zmienna losowa *X* ma rozkład normalny w wartością oczekiwaną μ i odchyleniem standardowym σ, co oznaczamy $X \sim N(\mu, \sigma)$ jeśli jej funkcja gęstości określona jest wzorem: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

   Uwaga: Dystrybuanta rozkładu normalnego nie wyraża się za pomocą funkcji elementarnych – jest stablicowana

Zmienna losowa

Własności rozkładu normalnego

- Symetryczny względem prostej $x=\mu$ co oznacza, że $P(X<\mu)=P(X>\mu)=0,5$
- Funkcja gęstości osiąga maksimum w punkcie $x=\mu$ wynoszące $f(\mu)=\frac{1}{\sqrt{2\pi}\sigma}$
- Prawdopodobieństwo, że zmienna losowa *X* przyjmuje wartości z przedziału $[m-3\sigma, m+3\sigma]$ jest w przybliżeniu równe 1
- Rozkład normalny $N(0,1)$ nazywamy rozkładem normalnym standardowym

Optymalny klasyfikator statystyczny…

Rozkład zmiennej losowej *I* scharakteryzowany jest prawdopodobieństwami a priori klas:

$$P(J=j)=p_j \text{ dla } j=1,...,k.$$

Wektor losowy X dla każdego $j\in\{1,...,k\}$ ma natomiast rozkład prawdopodobieństwa wyrażony gęstością, zwaną gęstością warunkową cech w klasie.

$$f(x\mid j)=f_j(x), \text{ dla } x\in R^p.$$

Optymalny klasyfikator statystyczny…

Gęstość bezwarunkowa zmiennej *X* wyraża się wzorem:

$$f(x)=\sum_{j\in\Theta} p_j f_j(x)=\sum_{j=1}^{M} p_j f_j(x).$$

Przyjmujemy, że $f(x)>0$ dla każdego $x\in R^p$.

Obliczmy prawdopodobieństwo warunkowe, że obiekt, któremu odpowiada wektor cech *x* należy do klasy *j*, to znaczy

$$P(J=j\mid X=x) \text{ dla } j\in\{1,...,k\},\ x\in R^p.$$

Optymalny klasyfikator statystyczny…

Stosując wzór Bayesa, mamy:

$$P(J=j\mid X=x)=\frac{P(X=x\mid J=j)\cdot P(J=j)}{\sum_{j\in\{1,...,k\}} P(X=x\mid J=j)\cdot P(J=j)},$$

Oznaczamy:

$$p_j=P(J=j)$$
$$p_j(x)=P(J=j\mid X=x),$$
$$f(x)=\sum_{j\in\{1,...,k\}} P(X=x\mid J=j)\cdot P(J=j)$$

Optymalny klasyfikator statystyczny…

Skąd:

$$p_j(x)=\frac{p_j f_j(x)}{f(x)}.$$

$p_j(x)$ - nazywamy prawdopodobieństwem a posteriori klasy *j*.

- Optymalny klasyfikator statystyczny (klasyfikator Bayesa) obiekt reprezentowany przez wektor cech *x* przyporządkowuje do klasy, dla której wartość prawdopodobieństwa a posteriori jest największa, czyli do klasy w której prawdopodobieństwo jego wystąpienia jest największe

Optymalny klasyfikator statystyczny…

Ponieważ dla dowolnego *x* wartość $f(x)$ jest stała, reguła decyzyjna optymalnego klasyfikatora Bayesa jest postaci:

$$\psi^*(x)=i, \text{ gdy } p_i(x)=\max_{j\in\{1,...,k\}} p_j(x),$$

czyli:

$$\psi^*(x)=i, \text{ gdy } p_i f_i(x)=\max_{1\le j\le k} p_j f_j(x).$$

Ponieważ logarytm jest funkcją rosnącą regułę powyższą można zapisać w równoważnej postaci:

$$\psi^*(x)=i, \text{ gdy } \ln\left(p_i f_i(x)\right)=\max_{1\le j\le k}\ln\left(p_j f_j(x)\right).$$

Optymalny klasyfikator statystyczny…

Przypadek normalności rozkładu cech w klasach:

Często spotykamy się z sytuacją gdy rozkłady wektorów cech obiektów w klasach pochodzą z wielowymiarowego rozkładu normalnego. Wtedy funkcje gęstości rozkładów cech w klasach wyrażają się wzorami:

$$f_j(x) = (2\pi)^{-\frac{p}{2}} \left|\Sigma_j\right|^{1/2} \exp\left[-\frac{1}{2}(x-m_j)^T \Sigma_j^{-1}(x-m_j)\right],\ \text{dla } j=1,2,\ldots M,$$

gdzie: m_j p-wymiarowy wektor wartości oczekiwanych,

Σ_j macierz kowariancji rozkładu w klasie j.

Optymalny klasyfikator statystyczny…

Podstawiając (patrz: 195) otrzymujemy:

$$\ln\left(g_j(x)\right) = \ln\left(p_j (2\pi)^{-\frac{p}{2}} \left|\Sigma_j\right|^{1/2} \exp\left[-\frac{1}{2}(x-m_j)^T \Sigma_j^{-1}(x-m_j)\right]\right) =$$

$$= \ln(p_j) + \ln\left((2\pi)^{-\frac{p}{2}}\right) + \ln\left(\left|\Sigma_j\right|^{1/2}\right) - \frac{1}{2}(x-m_j)^T \Sigma_j^{-1}(x-m_j) =$$

$$= \ln(p_j) - \frac{p}{2}\ln(2\pi) + \frac{1}{2}\ln\left|\Sigma_j\right| - \frac{1}{2}(x-m_j)^T \Sigma_j^{-1}(x-m_j)$$

pomijając stałą $-\frac{p}{2}\ln(2\pi)$ otrzymujemy funkcje klasyfikujące:

$$g_j(x) = \ln(p_j) + \frac{1}{2}\ln\left|\Sigma_j\right| - \frac{1}{2}(x-m_j)^T \Sigma_j^{-1}(x-m_j),\ j=1,2,\ldots M,$$

Optymalny klasyfikator statystyczny…

- W szczególnych przypadkach funkcje klasyfikujące mogą dalej się redukować:
  - gdy prawdopodobieństwa a priori wystąpienia obiektu są równe:

    $$g_j(x) = \frac{1}{2}\ln\left|\Sigma_j\right| - \frac{1}{2}(x-m_j)^T \Sigma_j^{-1}(x-m_j)$$

  - w przypadku, gdy macierze kowariancji rozkładów w każdej z klas są identyczne:

    $$g_j(x) = \ln(p_j) - \frac{1}{2}(x-m_j)^T \Sigma^{-1}(x-m_j)$$

Optymalny klasyfikator statystyczny…

Przykład

Rozważmy problem klasyfikacji do dwóch klas, wiadomo, że rozkłady cech w klasach są dwuwymiarowymi rozkładami normalnymi z wektorami wartości oczekiwanych dla klas 1, 2 odpowiednio:

$$m_1 = \begin{bmatrix}-3\\-3\end{bmatrix},\ m_2 = \begin{bmatrix}3\\3\end{bmatrix}$$

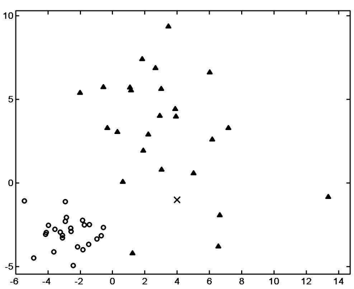
i macierzami kowariancji

$$\Sigma_1 = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix},\ \Sigma_2 = \begin{bmatrix}9 & 0\\0 & 9\end{bmatrix}$$

Prawdopodobieństwo wystąpienia każdej z klas wynosi 0,5.

Przykład…

Po 25 próbek z wymienionych rozkładów ilustruje rysunek. Krzyżykiem zaznaczono przykładową próbkę podlegającą klasyfikacji opisaną wektorem cech równym $\begin{bmatrix}4 & -1\end{bmatrix}^T$.



Przykład…

Rozkłady cech w klasach są dwuwymiarowymi rozkładami normalnymi. Uczenie klasyfikatora polega na wyznaczeniu funkcji klasyfikujących, które dla rozkładu normalnego wyrażają się wzorem: $g_j(x) = \ln(p_j) - \frac{1}{2}\ln\left|\Sigma_j\right| - \frac{1}{2}(x-m_j)^T \Sigma_j^{-1}(x-m_j)^T$.

Wstawiając odpowiednie wartości za m oraz Σ otrzymujemy:

$$g_1(x) = \ln(0,5) - \frac{1}{2}\ln|1| - \frac{1}{2}\left(\begin{bmatrix}x_1+3 & x_2+3\end{bmatrix}\cdot\begin{bmatrix}x_1+3\\x_2+3\end{bmatrix}\right) =$$

$$= \ln(0,5) - \frac{1}{2}\left((x_1+3)^2 + (x_2+3)^2\right) =$$

$$= \ln(0,5) - \frac{1}{2}\left(x_1^2 + x_2^2 + 6x_1 + 6x_2 + 18\right)$$

Przykład…

$$g_2(x) = \ln(0,5) - \frac{1}{2}\ln|81| - \frac{1}{2}\left(\begin{bmatrix} x_1 - 3 & x_2 - 3 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{9} & 0 \\ 0 & \frac{1}{9} \end{bmatrix} \cdot \begin{bmatrix} x_1 - 3 \\ x_2 - 3 \end{bmatrix} \right) =$$

$$= \ln(0,5) - \frac{1}{2}\ln|81| - \frac{1}{2}\left(\begin{bmatrix} \frac{x_1 - 3}{9} & \frac{x_2 - 3}{9} \end{bmatrix} \cdot \begin{bmatrix} x_1 - 3 \\ x_2 - 3 \end{bmatrix} \right) =$$

$$= \ln(0,5) - \frac{1}{2}\ln|81| - \frac{1}{2}\left(\frac{(x_1 - 3)^2}{9} + \frac{(x_2 - 3)^2}{9} \right) =$$

$$= \ln(0,5) - \frac{1}{2}\ln|81| - \frac{1}{2}\left(\frac{x_1^2}{9} + \frac{x_2^2}{9} - \frac{6}{9}x_1 - \frac{6}{9}x_2 + 2 \right)$$

Przykład…

Uwzględniając fakt, że prawdopodobieństwa a priori klas są równe otrzymujemy następujące funkcje klasyfikujące:

$$g_1(x) = -\frac{1}{2}\left(x_1^2 + x_2^2 + 6x_1 + 6x_2 + 18\right),$$

$$g_2(x) = -2\ln|3| - \frac{1}{2}\left(\frac{x_1^2}{9} + \frac{x_2^2}{9} - \frac{6}{9}x_1 - \frac{6}{9}x_2 + 2 \right).$$

Powierzchnia decyzyjna jest postaci $g_1(x) = g_2(x)$, a zatem wstawiając wyliczone funkcje klasyfikujące otrzymujemy:

$$\left(x_1^2 + x_2^2 + 6x_1 + 6x_2 + 18\right) = \frac{1}{2}\ln|81| + \left(\frac{x_1^2}{9} + \frac{x_2^2}{9} - \frac{6}{9}x_1 - \frac{6}{9}x_2 + 2 \right)$$

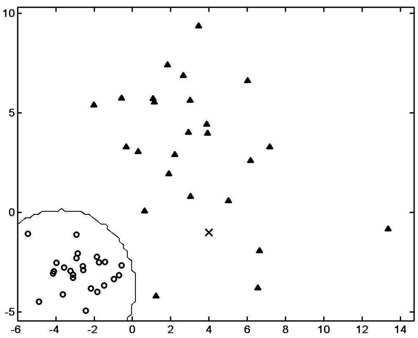
$$8x_1^2 + 8x_2^2 + 60x_1 + 60x_2 + 144 - 9\ln|81| = 0\,.$$

$$2x_1^2 + 2x_2^2 + 15x_1 + 15x_2 + 36 - 9\ln|3| = 0$$

Przykład…

Powierzchnię decyzyjną $2x_1^2 + 2x_2^2 + 15x_1 + 15x_2 + 36 - 9\ln|3| = 0$

ilustruje rysunek:



Przykład…

Zaklasyfikujmy punkt $\begin{bmatrix} 4 & -1 \end{bmatrix}^T$. Mamy

$$g_1(x) = -\frac{1}{2}\left(x_1^2 + x_2^2 + 6x_1 + 6x_2 + 18\right),$$

$$g_1\left(\begin{bmatrix} 4 \\ -1 \end{bmatrix}\right) = \frac{16 + 1 + 24 - 6 + 18}{-2} = -\frac{53}{2}$$

oraz $$g_2(x) = -2\ln|3| - \frac{1}{2}\left(\frac{x_1^2}{9} + \frac{x_2^2}{9} - \frac{6}{9}x_1 - \frac{6}{9}x_2 + 2 \right).$$

$$g_2\left(\begin{bmatrix} 4 \\ -1 \end{bmatrix}\right) = -2\ln|3| - \frac{1}{2}\left(\frac{-1}{9} + 2 \right) \approx -3{,}142$$

Klasyfikowany obiekt należy więc do klasy 2.

Parametryczny klasyfikator Bayesa

- Optymalna klasyfikacja bayesowska wymaga znajomości, rozkładu zmiennej losowej *J* i warunkowej gęstości rozkładu prawdopodobieństwa cech w klasach.
- W praktycznych zadaniach klasyfikacji tej wiedzy na ogół nie mamy
  - o Znana jest postać rozkładu lecz nieznane jego parametry
  - o Nie znana jest ani postać rozkładu
- W przypadku, gdy znamy postać warunkowego rozkładu cech w klasach, natomiast nie znamy jego parametrów, brakującą wiedzę rekompensujemy informacjami zawartymi w zbiorze uczącym.

Parametryczny klasyfikator Bayesa…

- Zbiór uczący $V = \{(x_1, y_1), ..., (x_N, y_N)\}$ - zbiór *N* niezależnych realizacji pary zmiennych losowych (X, J)
- Niech $V_i = \{x_j : y_j = i\}$ oznacza zbiór tych wektorów cech, które są elementami klasy *i*, to znaczy pochodzą z populacji o warunkowej gęstości $f_j(x)$. Niech $|V_j| = N_j$.
- Zadanie polega na estymacji brakujących parametrów i wstawieniu uzyskanych wartości do funkcji klasyfikujących optymalnego algorytmu Bayesa.
- Algorytm uzyskany w ten sposób nazywa się parametrycznym klasyfikatorem Bayesa.

Parametryczny klasyfikator Bayesa…

- Prawdopodobieństwa a priori klas przybliżamy częstościami występowania poszczególnych klas w zbiorze uczącym:

$$p_j = \frac{N_j}{N}$$

- Estymacja parametrów gęstości warunkowych cech w klasach zależy od samej postaci tych gęstości i może odbywać się na przykład metodą największej wiarogodności.

Przykład:

Załóżmy, że rozkład cech w klasach jest dwuwymiarowym rozkładem normalnym o nieznanych parametrach oraz dany jest zbiór uczący:

x_1	x_2	kl.	x_1	x_2	kl.	x_1	x_2	kl.	x_1	x_2	kl.
-4,887	-4,489	1	-1,755	-2,504	1	1,241	-4,208	2	-0,56	5,748	2
-2,928	-1,104	1	-2,586	-2,703	1	-0,318	3,301	2	6,636	-1,92	2
-0,707	-3,145	1	-4,112	-2,949	1	6,026	6,620	2	3,015	5,635	2
-2,566	-2,883	1	-2,852	-2,046	1	3,895	4,423	2	1,826	7,418	2
-2,442	-4,940	1	-1,481	-3,658	1	2,210	2,911	2	6,180	2,612	2
-1,861	-2,228	1	-3,633	-4,108	1	2,652	6,866	2	3,049	0,808	2
-3,582	-2,765	1	-1,434	-2,475	1	13,373	-0,834	2	3,464	9,363	2
-3,976	-2,535	1	-5,462	-1,062	1	1,895	1,942	2	1,152	5,554	2
-3,247	-2,940	1	-4,157	-3,055	1	0,636	0,080	2	0,294	3,051	2
-1,844	-3,978	1	-3,100	-3,281	1	1,072	5,711	2	7,186	3,293	2
-2,176	-3,810	1	-0,565	-2,653	1	6,560	-3,787	2	5,028	0,581	2
-2,928	-2,286	1	-0,974	-3,332	1	-2,024	5,388	2	3,926	3,987	2
-3,099	-3,135	1				2,936	4,015	2			

Przykład:

Funkcje klasyfikujące są postaci:

$$g_j(x) = \ln(p_j) - \frac{1}{2}(x - m_j)^T \Sigma^{-1} (x - m_j)$$

Estymatory wartości średniej i macierzy kowariancji wyliczamy na podstawie ze zbioru uczącego. Estymator największej wiarogodności dla wektora wartości średnich dany jest wzorem:

$$\hat{m}_j = \frac{1}{|V_j|} \sum_{i \in V_j} x_i$$

natomiast dla macierzy kowariancji:

$$\hat{\Sigma}_j = \frac{1}{|V_j|} \sum_{i \in V_j} (x_i - m_j)(x_i - m_j)^T .$$

Przykład:

Otrzymujemy odpowiednio:

$$\hat{m}_1 = \begin{bmatrix} -2,734 \\ -2,963 \end{bmatrix}, \ \hat{m}_2 = \begin{bmatrix} 3,254 \\ 3,1420 \end{bmatrix}$$

$$\hat{\Sigma}_1 = \begin{bmatrix} 1,575 & -0,134 \\ -0,134 & 0,827 \end{bmatrix}, \ \hat{\Sigma}_2 = \begin{bmatrix} 10,421 & -3,763 \\ -3,762 & 11,804 \end{bmatrix}$$

Wyznaczniki i macierze odwrotne do macierzy kowariancji:

$$|\hat{\Sigma}_1| = 1,284 , \ |\hat{\Sigma}_2| = 108,86 ,$$

$$\hat{\Sigma}_1^{-1} = \begin{bmatrix} 0,644 & 0,107 \\ 0,107 & 1,227 \end{bmatrix}, \ \hat{\Sigma}_2^{-1} = \begin{bmatrix} 0,108 & 0,035 \\ 0,035 & 0,096 \end{bmatrix}$$

Przykład:

Gęstości rozkładu cech w klasach:

$$f_1\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \frac{1}{2\pi\sqrt{1,284}} \exp\left[-\frac{1}{2}\begin{bmatrix} x_1 + 2,734 \\ x_2 + 2,963 \end{bmatrix}^T \cdot \begin{bmatrix} 0,644 & 0,107 \\ 0,107 & 1,227 \end{bmatrix} \cdot \begin{bmatrix} x_1 + 2,734 \\ x_2 + 2,963 \end{bmatrix}\right] =$$

$$= \frac{1}{2\pi\sqrt{1,284}} \exp\left[-0,322x_1^2 - 0,614x_2^2 - 0,107x_1x_2 - 2,078x - 3,928x_2 - 8,66\right]$$

oraz

$$f_2\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \frac{1}{2\pi\sqrt{108,86}} \exp\left[-\frac{1}{2}\begin{bmatrix} x_1 - 3,254 \\ x_2 - 3,142 \end{bmatrix}^T \cdot \begin{bmatrix} 0,108 & 0,035 \\ 0,035 & 0,096 \end{bmatrix} \cdot \begin{bmatrix} x_1 - 3,254 \\ x_2 - 3,142 \end{bmatrix}\right] =$$

$$= \frac{1}{2\pi\sqrt{108,86}} \exp\left[-0,054x_1^2 - 0,048x_2^2 - 0,035x_1x_2 + 0,461x_1 + 0,415x_2 - 1,403\right]$$

Przykład:

Uwzględniając fakt, że prawdopodobieństwa a priori klas są równe oraz opuszczając stałą $1/(2\pi)$ otrzymujemy funkcje klasyfikujące

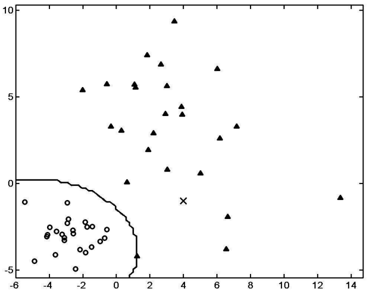
$$g_1\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \ln\frac{1}{\sqrt{1,284}} - 0,322x_1^2 - 0,614x_2^2 - 0,107x_1x_2 - 2,078x_1 - 3,928x_2 - 8,66$$

$$g_2\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \ln\frac{1}{\sqrt{108,86}} - 0,054x_1^2 - 0,048x_2^2 - 0,035x_1x_2 + 0,461x_1 + 0,415x_2 - 1,4$$

Przykład:

Powierzchnia decyzyjna w przypadku dwóch klas spełnia zależność $g_1(x) = g_2(x)$, a zatem także $\ln g_1(x) = \ln g_2(x)$, skąd wstawiając wyliczone funkcje klasyfikujące otrzymujemy:

$$\ln\sqrt{84,782} - 0,268x_1^2 - 0,566x_2^2 - 0,072x_1x_2 - 2,539x_1 - 4,343x_2 - 7,257 = 0$$



Przykład:

- Zaklasyfikujmy wektor: $x = \begin{bmatrix} 4 & -1 \end{bmatrix}^T$. Mamy

$$g_1\left(\begin{bmatrix} 4 \\ -1 \end{bmatrix}\right) = \ln\frac{1}{\sqrt{1,284}} - 0,322 \cdot 16 - 0,614 \cdot 1 - 0,107 \cdot 4 \cdot (-1) - 2,078 \cdot 4 +$$

$$-3,928 \cdot (-1) - 8,66 = \ln\frac{1}{\sqrt{1,284}} - 5,152 - 0,614 - 0,428 - 8,312 + 3,928 - 8,66 =$$

$$= -19,363$$

$$g_2\left(\begin{bmatrix} 4 \\ -1 \end{bmatrix}\right) = \ln\frac{1}{\sqrt{108,86}} - 0,054 \cdot 16 - 0,048 \cdot 1 - 0,035 \cdot 4 \cdot (-1) + 0,461 \cdot 4 +$$

$$+0,415 \cdot (-1) - 1,403 = \ln\frac{1}{\sqrt{108,86}} - 0,864 - 0,048 + 0,14 + 1,844 - 0,415 - 1,403 =$$

$$= -7,059$$

Zgodnie z regułą klasyfikatora klasa 2