Klasyfikacja bez pomiaru

Musimy dokonać klasyfikacji nie znając żadnych cech obiektu. Jedyną przesłanką, która pozwala podjąć decyzję klasyfikacyjną, jest prawdopodobieństwo pojawiania się obiektów różnych klas.

W przypadku dwóch klas c_1 i c_2 uznajemy, że obiekt należy do klasy:

c_1 jeśli $P(c_1) > P(c_2)$

c_2 wpp.

$P(c_1)$ i $P(c_2)$ są nazywane prawdopodobieństwami *a priori*.

Klasyfikator podejmuje oczywiście **zawsze** taką samą decyzję, zależną jedynie od p. *a priori*.

Błąd klasyfikacji: $P_e = \min[P(c_1), P(c_2)]$

W przypadku klasyfikacji płci bieżącej edycji uczestników wykładu, dostaniemy klasyfikator z błędem 3/32 !

Funkcja gęstości prawdopodobieństwa

W przypadku, gdy do klasyfikacji dysponujemy jedną ciągłą cechą, możemy ją potraktować jako ciągłą zmienną losową X.
Rozkład wartości x jest opisany funkcją gęstości prawdopodobieństwa $p(x)$ (pdf - *probability density function*).

Dla każdego przedziału $\langle x_1, x_2 \rangle$ prawdop. wystąpienia wartości X z tego przedziału liczymy ze wzoru: $P(c \in C : x_1 \le X \le x_2) = \int_{x_1}^{x_2} p(x)dx$

Dystrybuanta ciągłej zm.l.: $F(x) = P(c \in C : X < x) = \int_{-\infty}^{x} p(x)dx$

Oczywiście: $\int_{-\infty}^{+\infty} p(x)dx = 1$

Klasyfikacja z wykorzystaniem pdf

Dla potrzeb klasyfikacji bardziej interesujący od ogólnego rozkładu prawdopodobieństwa jest rozkład X w zależności od klasy, do której należą obiekty. Jest to funkcja warunkowego rozkładu gęstości prawdopodobieństwa $p(x \mid c_i)$.

Poszukujemy pr. $P(c_i \mid x)$ wystąpienia klasy c_i, przy warunku pomiaru cechy obiektu o wartości x: $P(c_i \mid x) \sim p(x \mid c_i) P(c_i)$

Wartości $P(c_i \mid x)$ muszą się sumować do 1 po wszystkich klasach.

Wprowadzamy czynniki normalizujący: $p(x) = \sum_i p(x \mid c_i) P(c_i)$

Ostatecznie, poszukiwane prawdopodobieństwo *a posteriori*:

$$P(c_i \mid x) = \frac{p(x \mid c_i) P(c_i)}{p(x)} = \frac{p(x \mid c_i) P(c_i)}{\sum_i p(x \mid c_i) P(c_i)}$$ (por. Tw. Bayesa)

Reguła Bayesa (1)

Mając prawdop. *a posteriori* uznajemy, że obiekt jest klasy

c_1 jeśli $P(c_1 \mid x) > P(c_2 \mid x)$

c_2 wpp.

$p(x)$ w poprzednim wzorze jest tylko czynnikiem skalującym, możemy zatem powyższą regułę zapisać jako:
obiekt jest klasy

c_1 jeśli $p(x \mid c_1)P(c_1) > p(x \mid c_2)P(c_2)$

c_2 wpp.

Warunek decyzji można dalej przekształcać:

$\frac{p(x \mid c_1)}{p(x \mid c_2)} > \frac{P(c_2)}{P(c_1)}$, albo $\ln p(x \mid c_1) - \ln p(x \mid c_2) > \ln P(c_2) - \ln P(c_1)$

...

Straty i ryzyko

Decyzje klasyfikacyjne mogą wiązać się z różnym kosztem - stratą - w systemie klasyfikacji.

Dla każdej decyzji klasyfikacyjnej $\alpha_i, i = 1 \ldots a$ wprowadzamy funkcję straty $\lambda(\alpha_i \mid c_j)$. Jest to strata związana z podjęciem decyzji α_i, gdy faktycznie obiekt jest klasy c_j.

Całkowita strata dla decyzji α_i: $R(\alpha_i \mid x) = \sum_{j=1}^{a} \lambda(\alpha_i \mid c_j) P(c_j \mid x)$ jest nazywana ryzykiem warunkowym.

Ryzyko całkowite klasyfikatora - dla wszystkich wartości x, z uwzględnieniem rozkładu $p(x)$:

$$R = \int R(\alpha(x) \mid x) p(x) dx$$

Reguła Bayesa (2)

Klasyfikator optymalny minimalizuje ryzyko całkowite. Uzyskamy to, podejmując dla każdej wartości x, decyzję minimalizującą ryzyko warunkowe $R(\alpha(x)|\, x)$.

Reguła decyzyjna Bayesa:
Oblicz $R(\alpha_i \,|\, x)$ dla $i = 1 \ldots a$ i wybierz akcję α_i, dla której $R(\alpha_i \,|\, x)$ jest minimalne.

Dla zadanych funkcji strat, reguła decyzyjna Bayesa daje **najlepsze** wyniki klasyfikacji (stąd klasyfikator optymalny Bayesa).

Dlaczego nie klasyfikować wszystkiego optymalnie?!

Klasyfikator minimalizujący wsp. błędów

Minimum-error-rate classifier

Często spotyka się funkcję strat, która ma tylko dwie wartości: dla decyzji poprawnej i decyzji błędnej.

Formalnie: $\lambda(\alpha_i \mid c_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$

Ryzyko warunkowe w tym przypadku:

$$R(\alpha_i \mid x) = \sum_{j=1}^{a} \lambda(\alpha_i \mid c_j) P(c_j \mid x) = \sum_{i \neq j} P(c_j \mid x) = 1 - P(c_i \mid x)$$

Zwróćmy uwagę, że warunek decyzyjny w tym przypadku sprowadza się ponownie do prawdopodobieństwo *a posteriori*, tzn. wybieramy c_i gdy $P(c_i \mid x) > P(c_j \mid x), i \neq j$

Funkcje decyzyjne

F. rozróżniające, discriminant functions

Określmy dla każdej klasy funkcję decyzyjną $g_i(x), i = 1 \ldots a$.

Uznajemy, że obiekt jest klasy c_i jeśli $g_i(x) > g_j(x), i \neq j$

W tej reprezentacji klasyfikator składa się z a modułów obliczających funkcje decyzyjne dla wszystkich klas oraz selektora maksimum wybierającego klasę.

F. decyzyjne dzielą przestrzeń cech na *obszary decyzyjne* $\Re_1 \ldots \Re_c$ poszczególnych klas, rozdzielone *powierzchniami decyzyjnymi*.

Obszary $\Re_i$ i $\Re_j$ (odpowiednio klas c_i i c_j) są rozdzielone powierzchnią o równaniu: $g_i(x) - g_j(x) = 0$.

Funkcje decyzyjne

Klasyfikator Bayesa: $g_i(x) = -R(\alpha_i \mid x)$

Klasyfikator minimalizujący wsp. błędu: $g_i(x) = P(c_i \mid x)$

Zwróćmy uwagę, że powyższe przypisanie funkcji decyzyjnych nie jest jedyne. Każda nowa funkcja $\dot{g}_i(x) = s g_i(x) + t, s > 0$ da nam takie same wyniki klasyfikacji.

Ogólnie, jeżeli mamy funkcję f monotonicznie rosnącą, to funkcja decyzyjna $\dot{g}_i(x) = f(g_i(x))$ nie zmienia klasyfikacji.

Np.:

$$g_i(x) = P(c_i \mid x) = p(x \mid c_i) P(c_i) = \ln p(x \mid c_i) + \ln P(c_i)$$

Parametry rozkładu prawdopodobieństwa

Wartość oczekiwana:

Dla ciągłej zm.l. X o rozkładzie $p(x)$: $$EX = \int_{-\infty}^{\infty} x\, p(x)\, dx$$

Dla dyskretnej zm.l. X o zbiorze skoków W

i skokach $p_i = P(X = x_i)$: $$EX = \sum_{x_i \in W} x_i\, p_i$$

Wariancja: $\sigma^2 = E(X - EX)^2$

Momentem rzędu *r* względem stałej *s* jest liczba $E(X - s)^r$
względem $s = 0$ momenty normalne,
względem $s = EX$ momenty centralne

Jednowymiarowy rozkład normalny

$$p(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

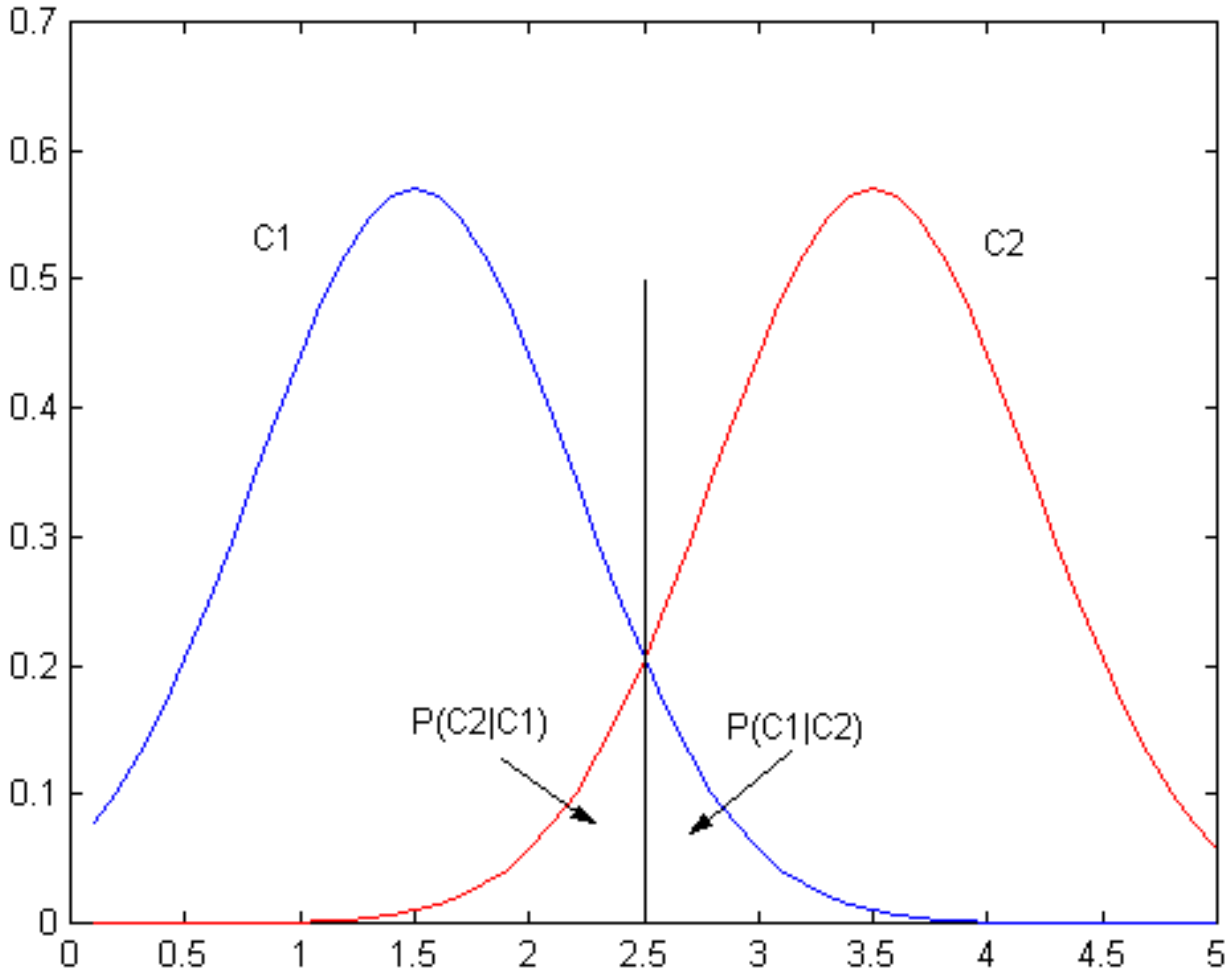
Przykład:

μ_1= 1.5 σ_1= 0.7

μ_2= 3.5 σ_2= 0.7

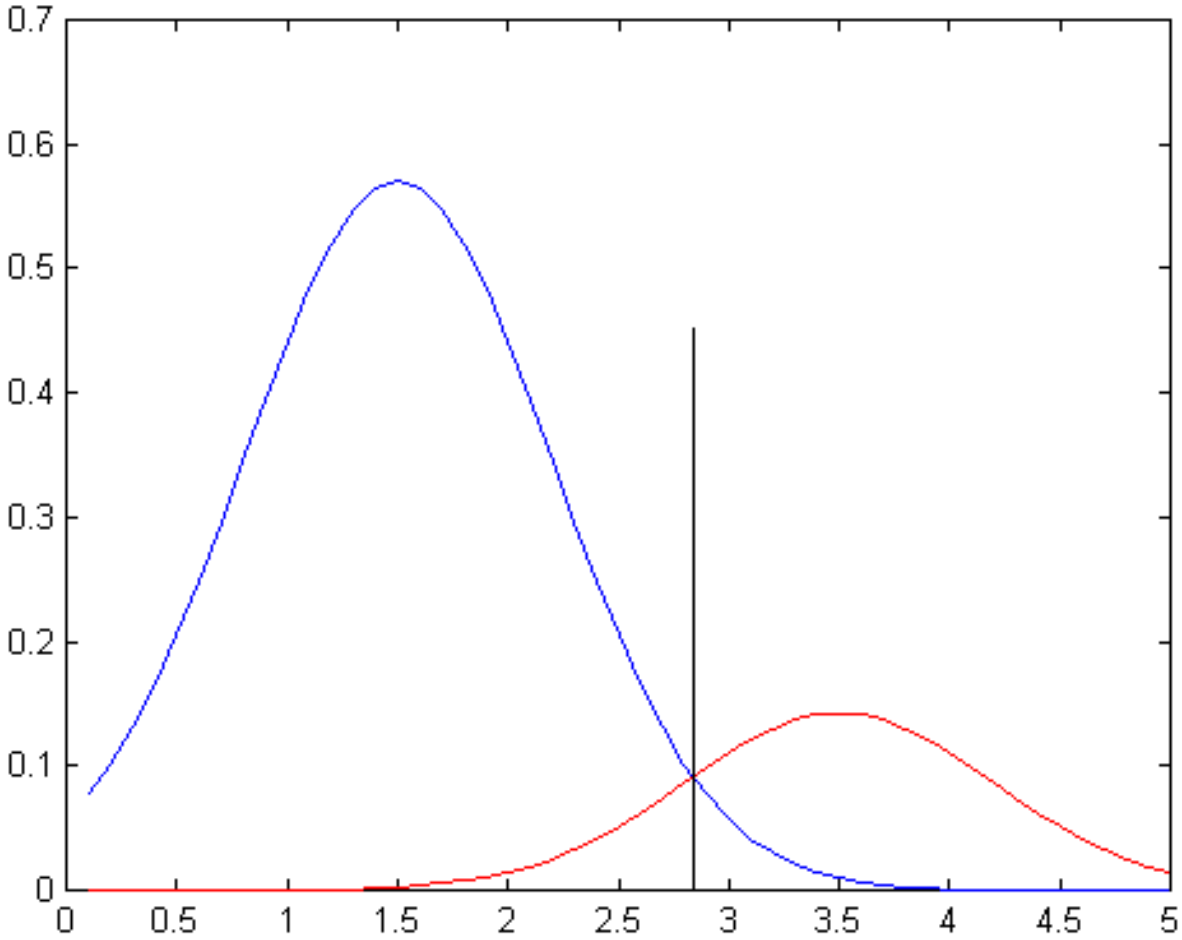
dla $P(c_1)=P(c_2)=0.5$ $\quad x_{opt}=\frac{\mu_2+\mu_1}{2}=2.5$

dla $P(c_1)=0.8\ P(c_2)=0.2$ $\quad x_{opt}=\frac{\mu_2+\mu_1}{2}-\frac{\sigma^2\ln(P_2/P_1)}{\mu_2-\mu_1}=2.5+0.34=2.84$



$P(C_2|C_1)=F_{C2}(x_{opt})=0,0766$ $P(C_1|C_2)= 1-F_{C1}(x_{opt})=0,0766$

Prawd. błędu klasyfikatora: $0.5 * P(C_2|C_1) + 0.5*P(C_1|C_2) = 0.0766$



$P(C_2|C_1)=F_{C2}(x_{opt})=0.1729$ $P(C_1|C_2)= 1-F_{C1}(x_{opt})=0.0278$

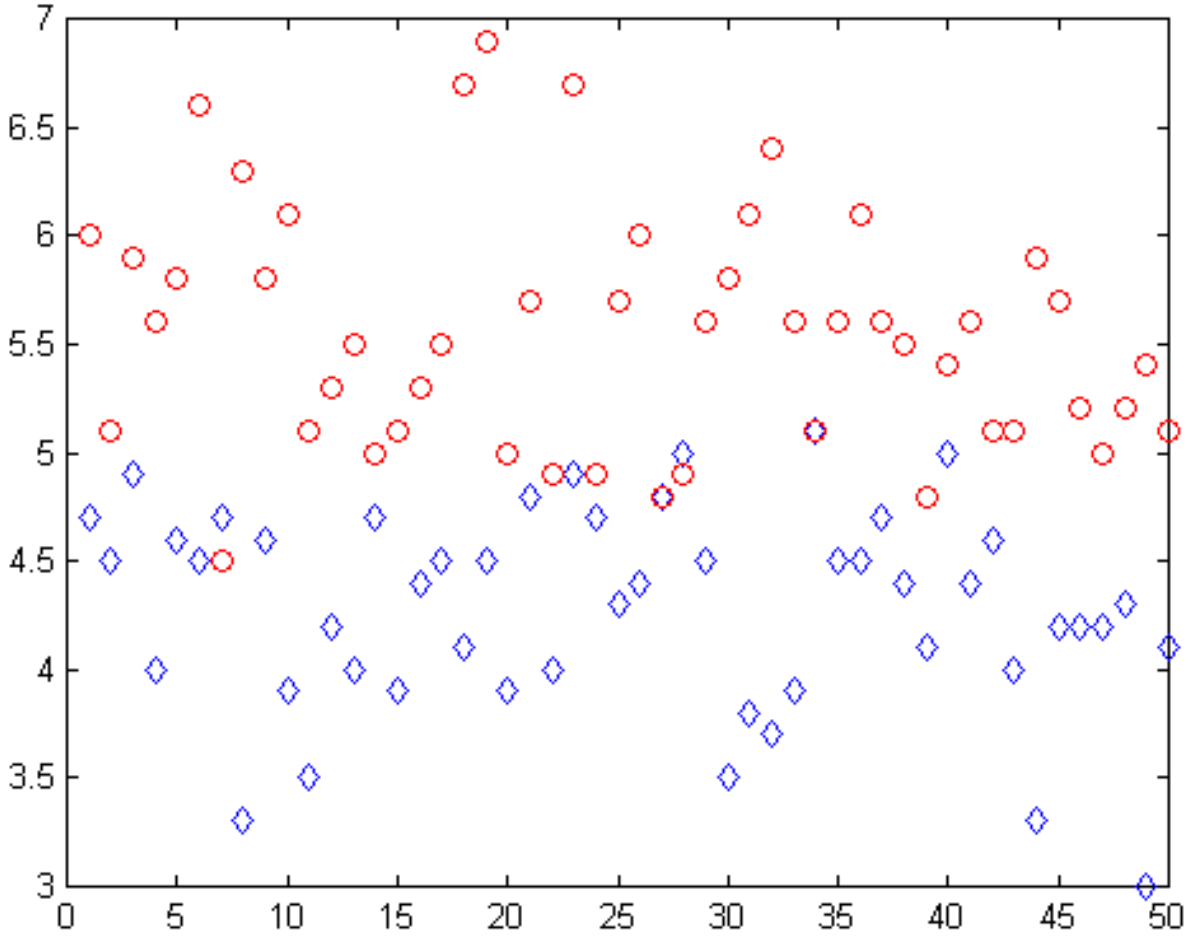
Prawd. błędu klasyfikatora: $0.2 * P(C_2|C_1) + 0.8*P(C_1|C_2) = 0.0568$

Klasyfikacja kwiatów

Iris virginica

Iris versicolor
(kosaciec różnobarwny)

Cecha: długość płatka



Wartości:

$\mu_1 = 5.55$

$\sigma_1 = 0.55$

$\mu_2 = 4.29$

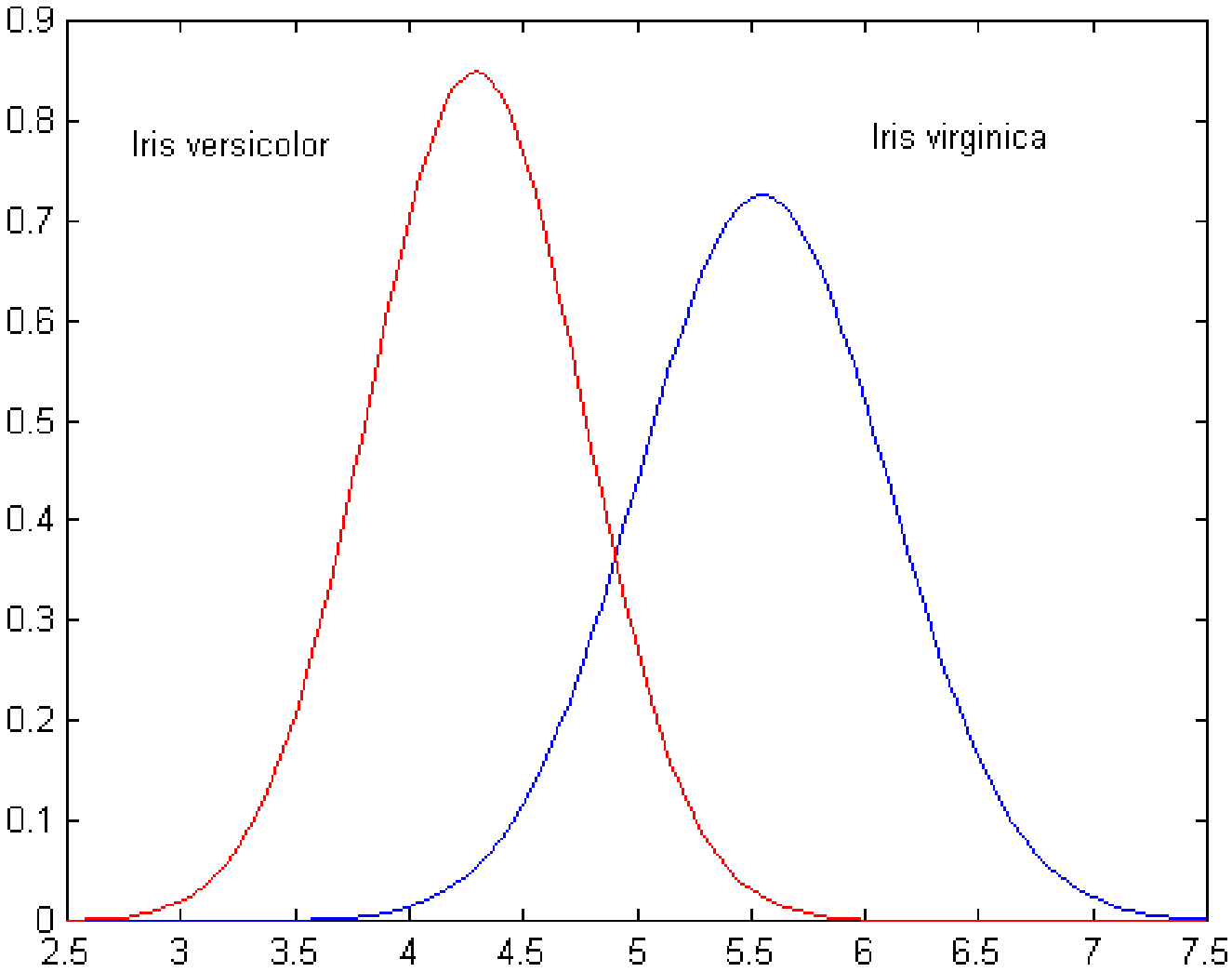
$\sigma_2 = 0.47$

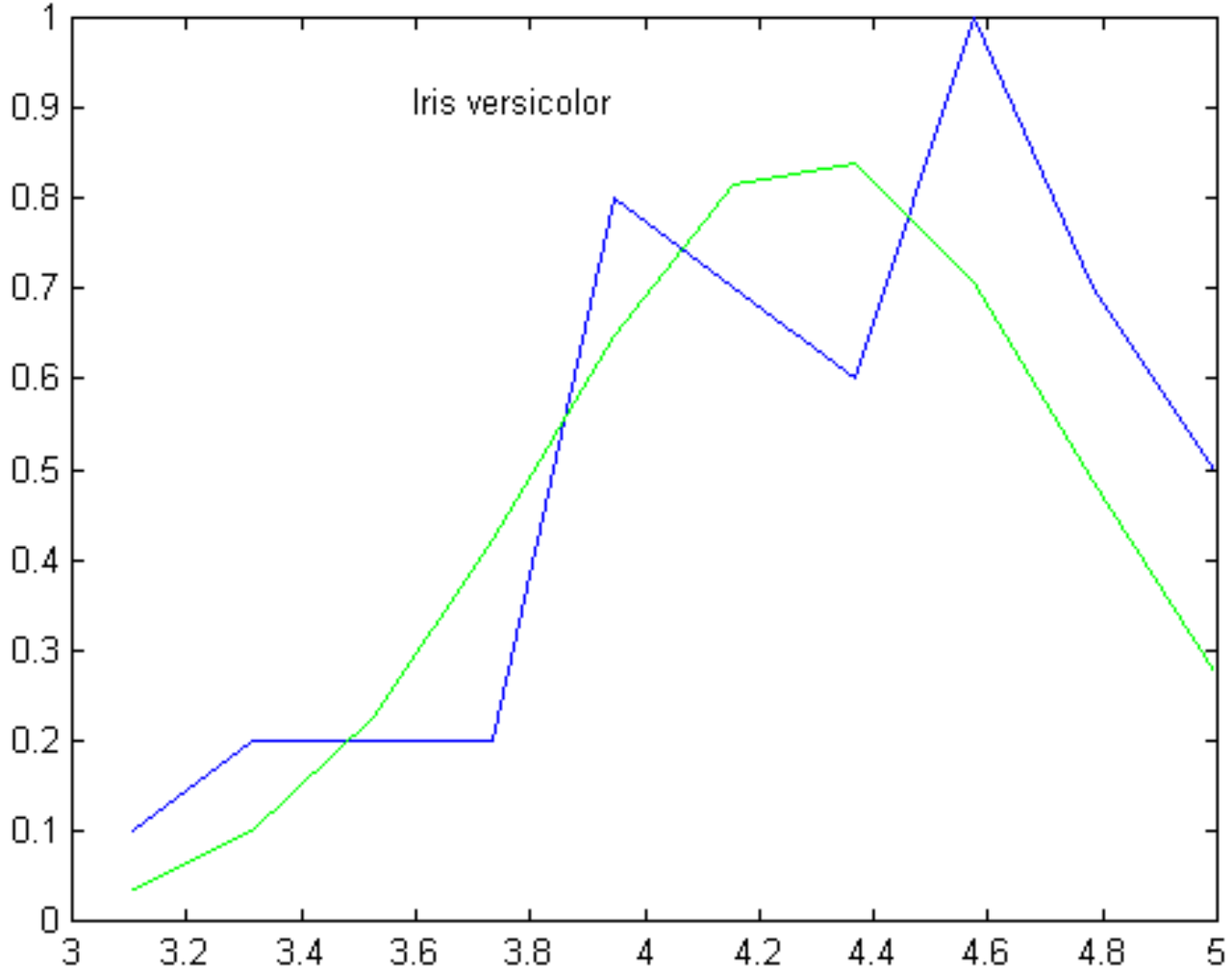
$x_{opt} \approx 4.91$

$P_{err} \approx 0.11$

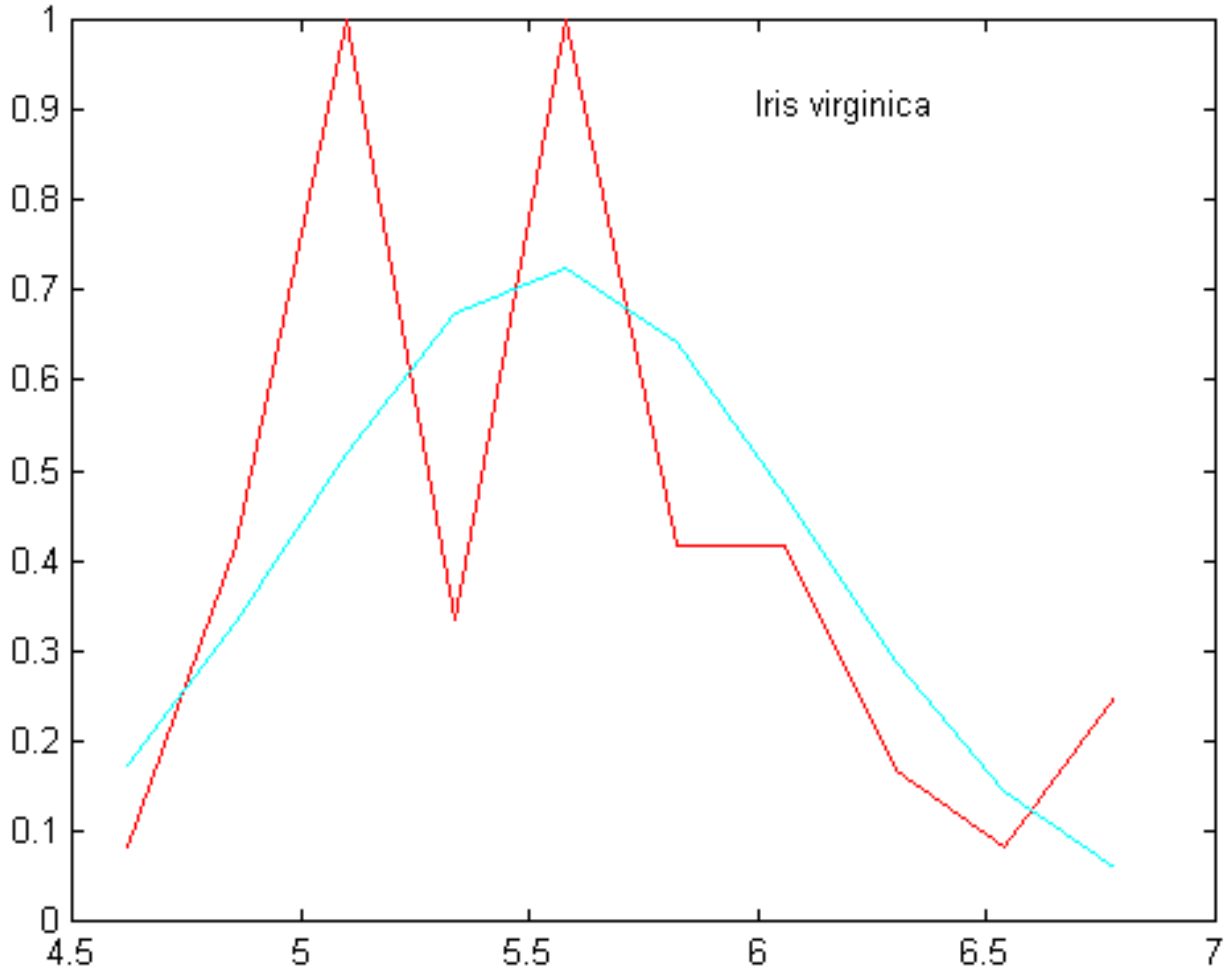
Zmierzony błąd:

$P_{err} \approx 0.09$





P_{err}=0.5*0.0961=0.048 Błąd zmierzony = 0.03



P_{err}=0.5*0.1197=0.0599 Błąd zmierzony = 0.06

Wielowymiarowy rozkład normalny

$$p(x)=\frac{1}{\sqrt{(2\pi)^d\,|\Sigma|}}\exp\left[-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right]$$

Wektor cech: $\mathbf{x}^T=(x_1,x_2,\ldots,x_d)$

Wartość oczekiwana: $E\mathbf{X}=\boldsymbol{\mu}=\frac{1}{N}\sum_{i=1}^{N}x_i$

Wariancja $\rightarrow$ macierz wariancji-kowariancji

$$\boldsymbol{\Sigma}=E\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right]=\{\sigma_{ij}\}$$

$$\sigma_{ij}=E\left[(x_i-\mu_i)(x_j-\mu_j)^T\right]$$

$\sigma_{ii}=\sigma_i^2$ wariancja cechy i

σ_{ij} kowariancja między cechami i oraz j ($\sigma_{ij}=\sigma_{ji}$)

Właściwości kowariancji

Cechy są nieskorelowane, gdy

$$E\left[x_i\, x_j\right] = E\left[x_i\right] E\left[x_j\right]$$

Cechy są ortogonalne, gdy

$$E\left[x_i\, x_j\right] = 0$$

Cechy są niezależne, gdy

$$p(x_i, x_j) = p(x_i)\, p(x_j)$$

Współczynnik korelacji: $\rho_{ij} = \dfrac{\operatorname{cov}(x_i, x_j)}{\sqrt{\operatorname{var}(x_i)\operatorname{var}(x_j)}} = \dfrac{\sigma_{ij}}{\sqrt{\sigma_{ii}\,\sigma_{jj}}}$

Populacja - próba

Estymator wartości oczekiwanej: $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Wariancja próby: $\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$

Macierz wariancji-kowariancji próby:

$$\sigma_{ij} = \frac{1}{N-1}\sum_{k=1}^{N}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)$$

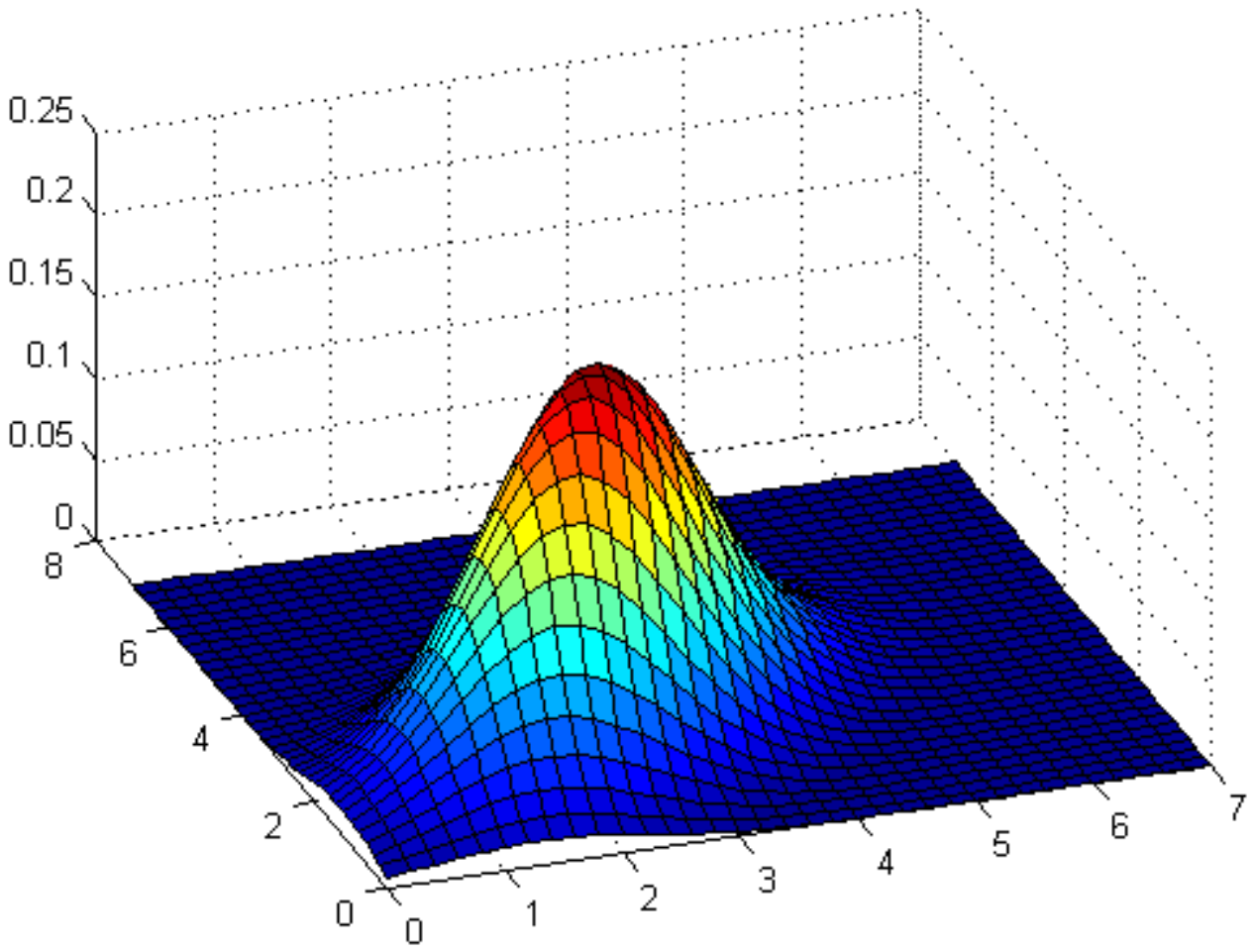
$$\sigma_{ij} = \frac{1}{N-1}\sum_{k=1}^{N} d_{ik} d_{jk}, \quad d_{ik} = x_{ik} - \bar{x}_i$$

Rozkłady

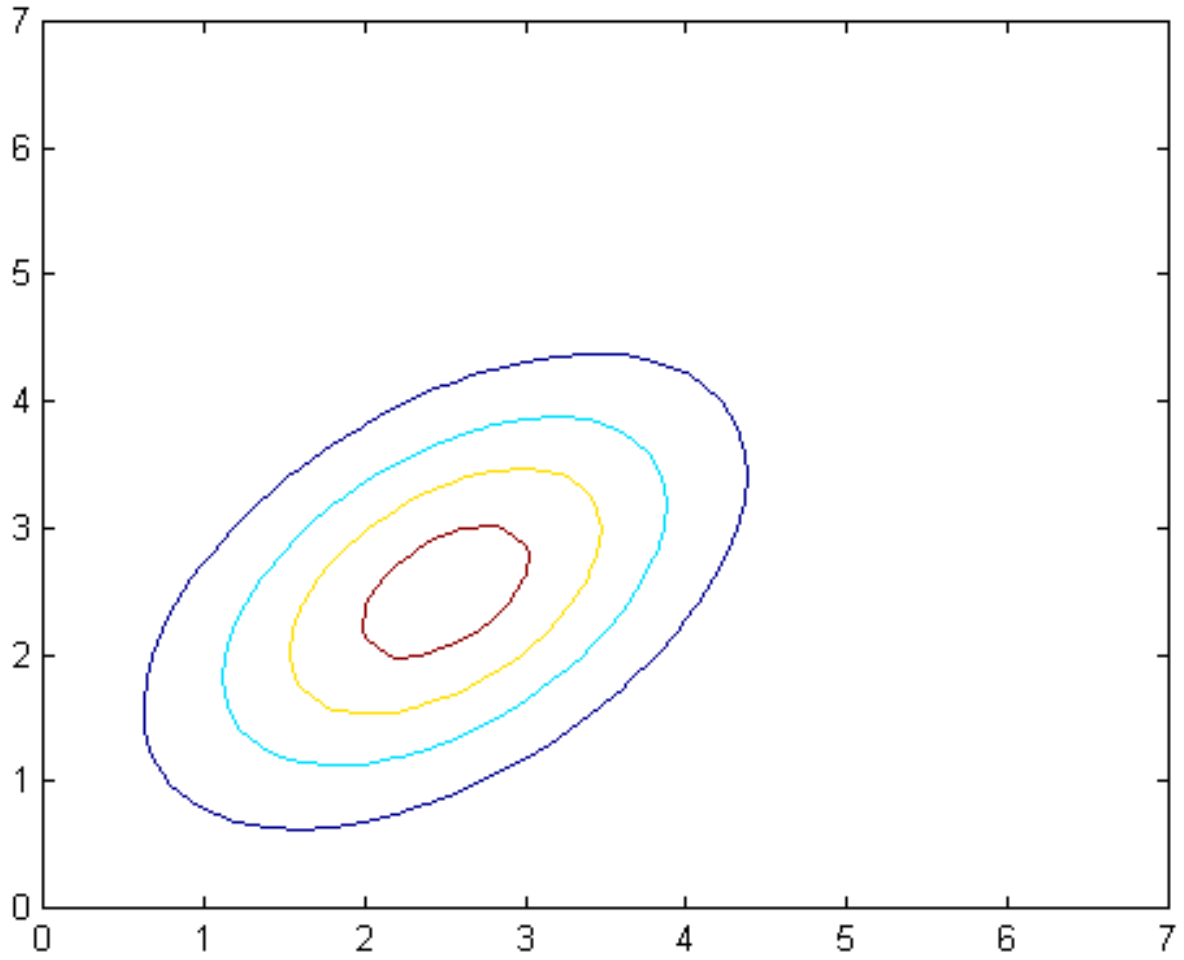
Rozkład brzegowy:
Jeśli X jest wielowymiarowym rozkładem normalnym, to każdy rozkład brzegowy też jest normalny.

Rozkład warunkowy:
Jeśli X jest wielowymiarowym rozkładem normalnym, to każdy rozkład warunkowy też jest normalny.

Rozkład kombinacji:
Jeśli X jest wielowymiarowym rozkładem normalnym, to każdy rozkład liniowej kombinacji cech też jest normalny.

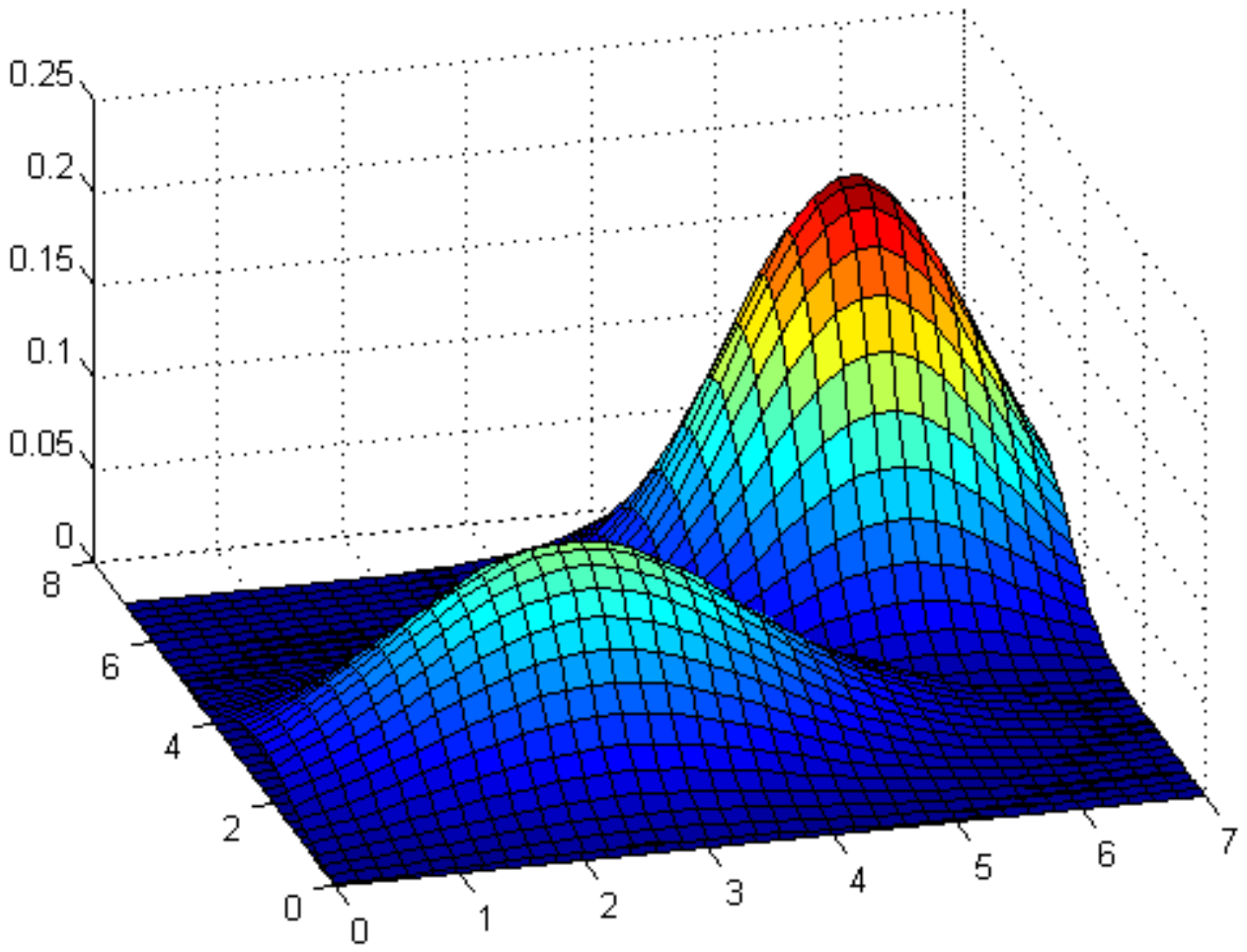


Dwuwymiarowy rozkład normalny

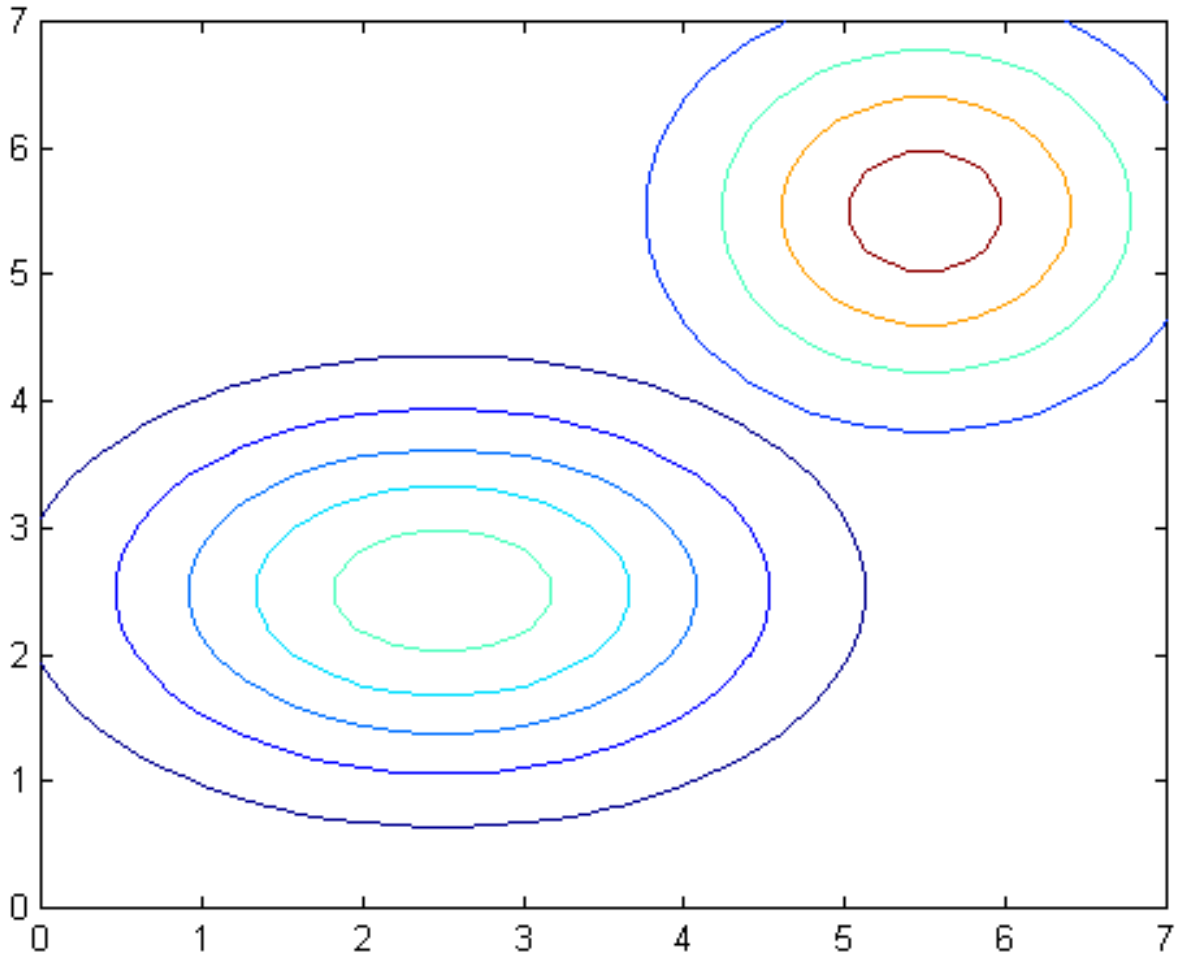


... i jego wykres konturowy

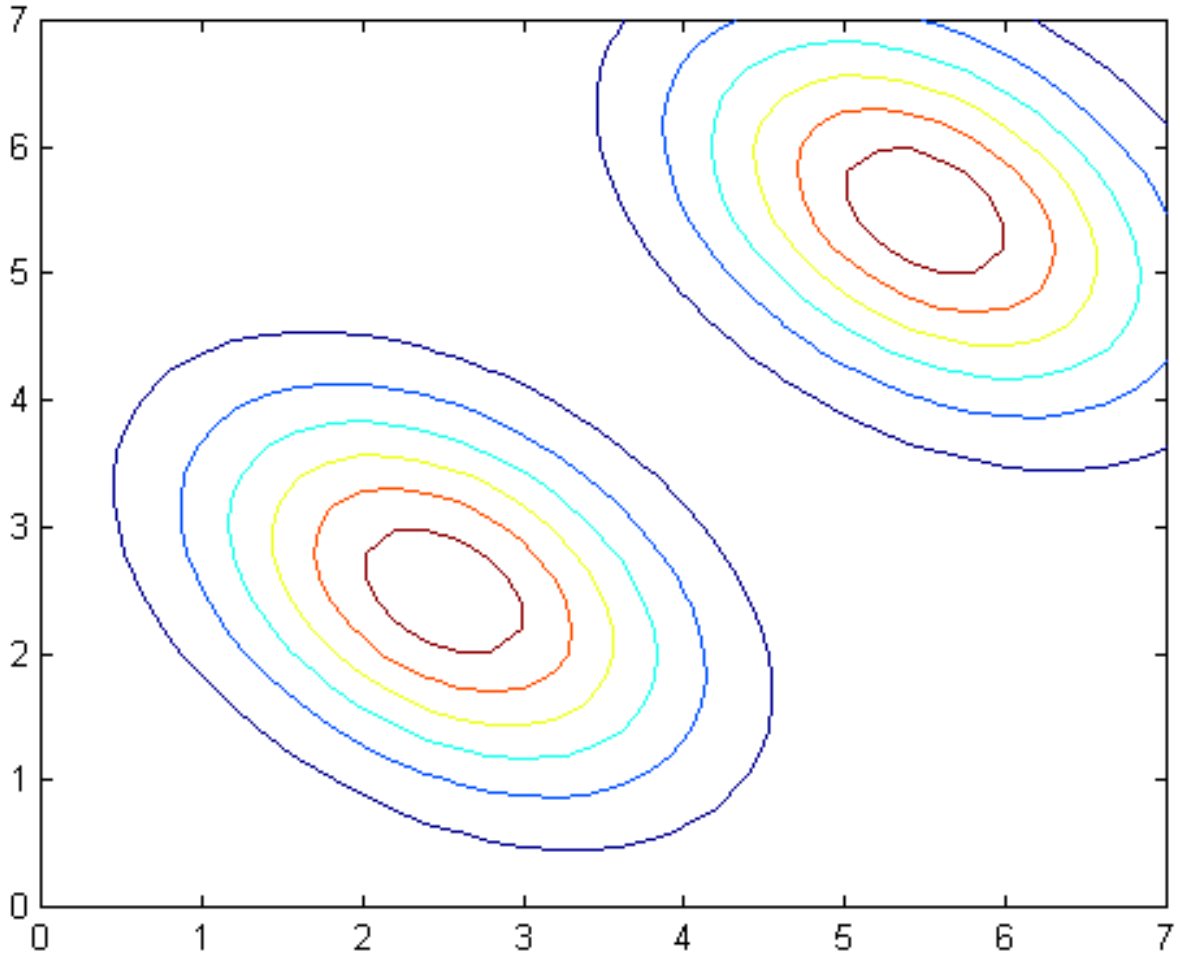
Odległość Mahalanobisa r: $r^2 = (\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})$



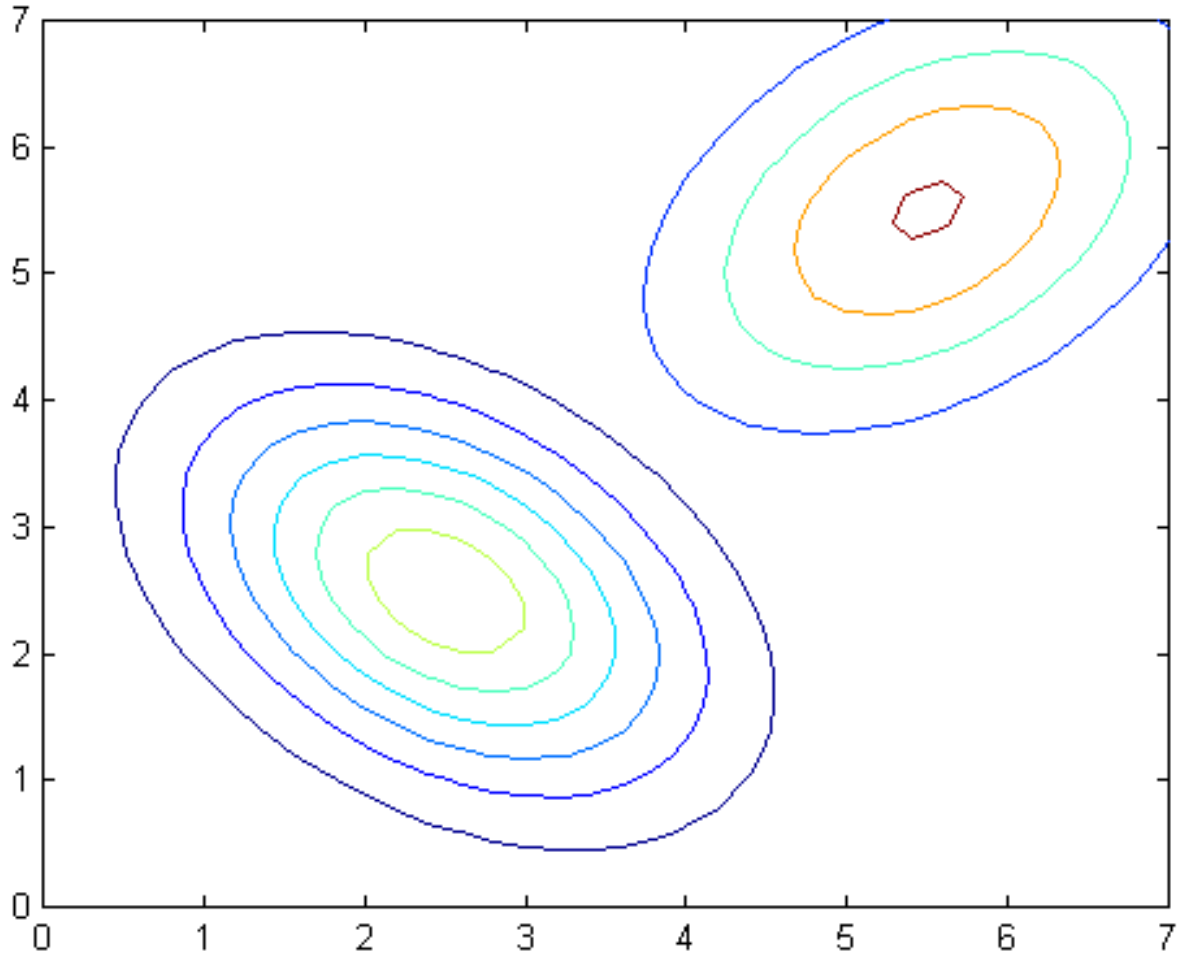
Dwie klasy o rozkładach normalnych



... i ten sam wykres konturowo



Równe macierze wariancji-kowariancji



Nierówne macierze wariancji-kowariancji

Funkcje decyzyjne dla rozkładu normalnego

Weźmiemy pod uwagę klasyfikator minimalizujący wsp. błędu (tzn. przyjmiemy zerojedynkową funkcję strat).

Funkcje decyzyjne: $g_i(\mathbf{x}) = \ln p(\mathbf{x} \mid c_i) + \ln P(c_i)$

Podstawiamy: $p(\mathbf{x} \mid c_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \ln P(c_i)$$

$\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}$ liniowa granica między klasami

$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ liniowa granica między klasami

$\boldsymbol{\Sigma}_i$ granica jest powierzchnią drugiego stopnia!

Test zgodności χ^2

Załóżmy, że chcemy sprawdzić, czy rozkład danych, którymi dysponujemy jest zgodny z rozkładem normalnym o zadanych parametrach. Oczywiście parametry tego rozkładu: średnią i odchylenie standardowe wyznaczamy z naszych danych. Podstawowa procedura przeprowadzenia testu polega na podziale danych na *k* koszyków. Dla każdego koszyka będziemy mieć liczbę N_i oznaczającą liczbę egzemplarzy danych w koszyku. Znając rozkład teoretyczny (w przykładzie normalny) i jego parametry, możemy dla stosownych zakresów koszyków wyznaczyć liczby n_i, które reprezentują liczbę egzemplarzy danych z rozkładu teoretycznego należących do poszczególnych koszyków. Możemy teraz wyznaczyć statystykę χ^2:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(N_i - n_i\right)^2}{n_i}$$

Test zgodności χ^2

Wartość statystyki służy do **odrzucenia** hipotezy zerowej, że badane dane są zgodne z pewnym rozkładem teoretycznym. Generalnie im większa wartość statystyki χ^2, tym bardziej prawdopodobne, że hipotezę zerową należy odrzucić. Wartości graniczne odczytuje się z tablic (policzenie rozkładu χ^2 jest dość skomplikowane), przy czym trzeba uwzględnić liczbę **stopni swobody** w teście. Wyjściowo, liczba stopni swobody jest taka jak liczba koszyków (minus 1), ale musimy odjąć od niej liczbę parametrów rozkładu teoretycznego obliczoną na podstawie danych. W przykładzie są dwa parametry: średnia i odchylenie standardowe. Dodatkowo, liczba elementów w zbiorze danych jest ustalona, co oznacza, że liczbę elementów w jednym koszyku możemy wyznaczyć na podstawie liczby elementów w innych koszykach – zmniejsza to liczbę stopni swobody o 1.

Test χ^2 dla kosaćców

Do testu użyję 4 koszyków, położonych wokół wartości średniej.

1: $x<\mu-\sigma/2$ 2: $\mu-\sigma/2 \geq x < \mu$ **3:** $\mu \geq x < \mu+\sigma/2$ **4:** $x \geq \mu+\sigma/2$

Wybór przedziałów nie jest bez znaczenia. Na ogół zaleca się by dla każdego koszyka był spełniony warunek $n_i \geq 5$ (w wielu źródłach pojawia się $n_i \geq 10$).

Daje mi to następujące liczności w koszykach:

Koszyk	N_i	p_i	n_i	χ^2_i
1	15	0.3048	15.2	0.0026
2	7	0.1952	9.8	0.8000
3	13	0.1952	9.8	1.0449
4	15	0.3048	15.2	0.0026

W sumie χ^2=1.8502

Liczba stopni swobody: 4 (koszyki) – 2 (parametry) – 1 = 1

Z tablic rozkładu χ^2 czytamy wartość dla wybranego poziomu istotności (powiedzmy α=0.05): 3.841 (>1.85).

Nie można wykluczyć (z istotnością 0.05), że rozkład jest normalny!

Problemy reguły Bayesa

- Kryterium może być skomplikowane nawet dla rozkładu normalnego.
- Jeszcze gorzej, gdy rozkład nie jest normalny.
- Może być trudno ustalić rozkład.
- Klasy o małym p. a priori słabo wpływają na kryterium decyzyjne.
- Skąd wziąć p. a priori?
- Kompletność ?! - decyzja wymijająca.

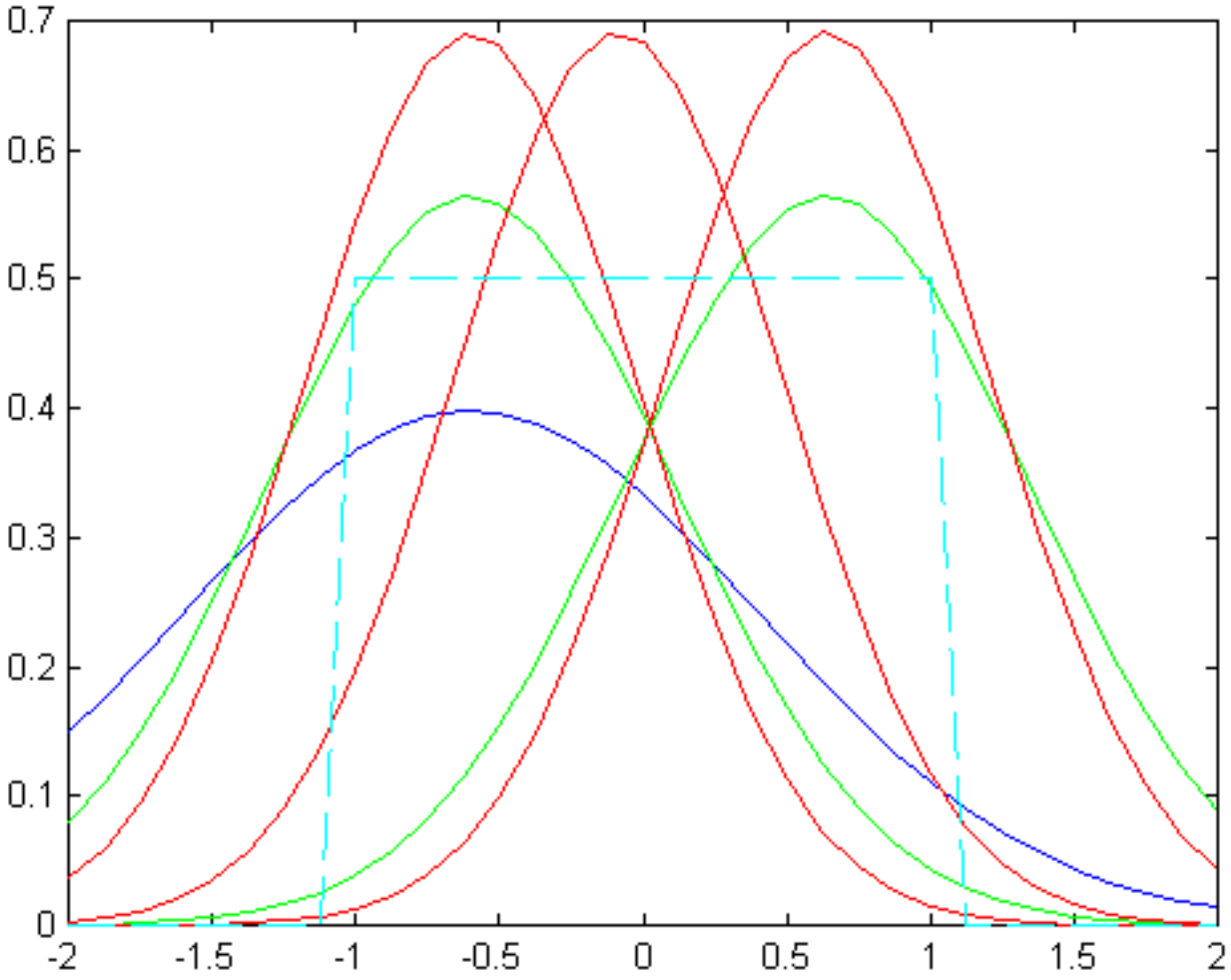
Estymacja pdf z oknem Parzena

Możemy wyznaczyć przybliżony rozkład gęstości prawdopodobieństwa przy użyciu tzw. Okna Parzena. Zasada polega na "budowaniu" nieznanego rozkładu ze składowych wprowadzanych przez punkty ze zbioru uczącego. "Częściówki" daje nam funkcja okna $\varphi(u)$. Nie ma specjalnych ograniczeń na postać tej funkcji, ale musi być typu *pdf* (nieujemna; całka po całej dziedzinie = 1).

Można użyć gaussowskiej funkcji okna: $\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$

Dodatkowo, wprowadzimy parametr h_1, nazywany szerokością okna, który będzie skalowany liczbą punktów w zbiorze uczącym: $h_n = \frac{h_1}{\sqrt{n}}$

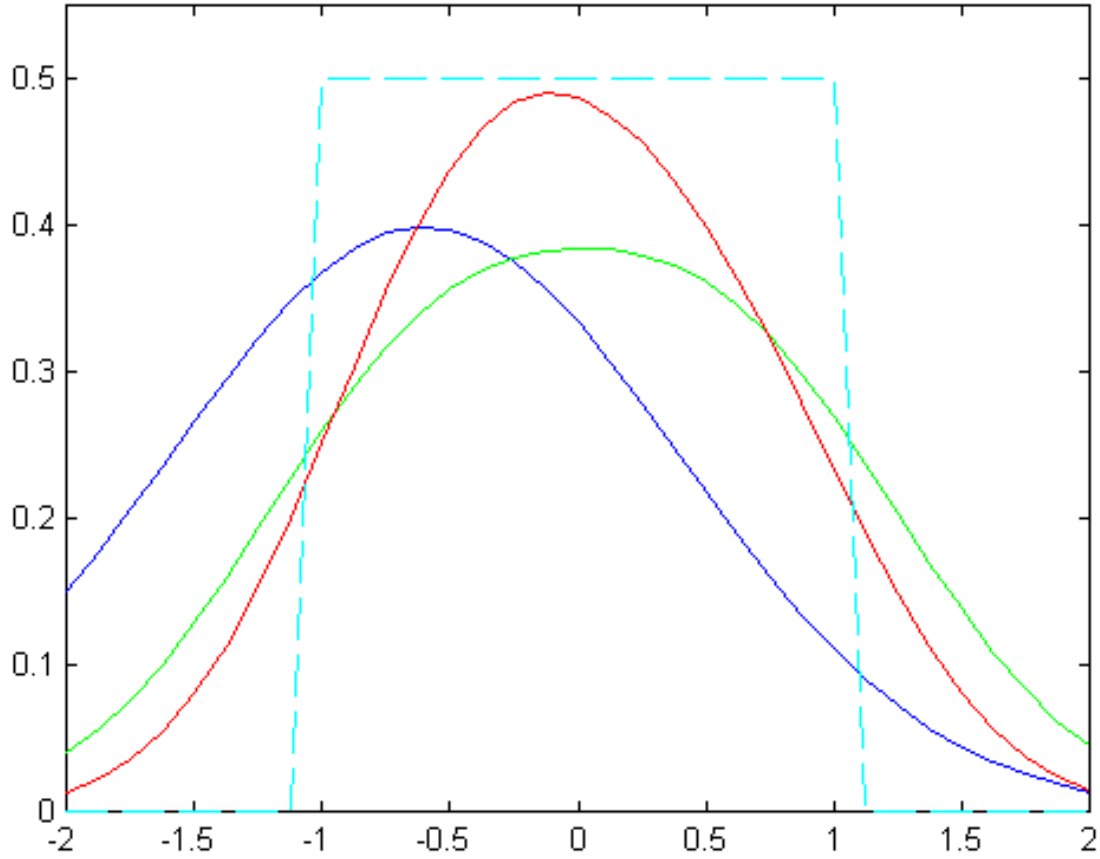
Estymacja pdf z oknem Parzena



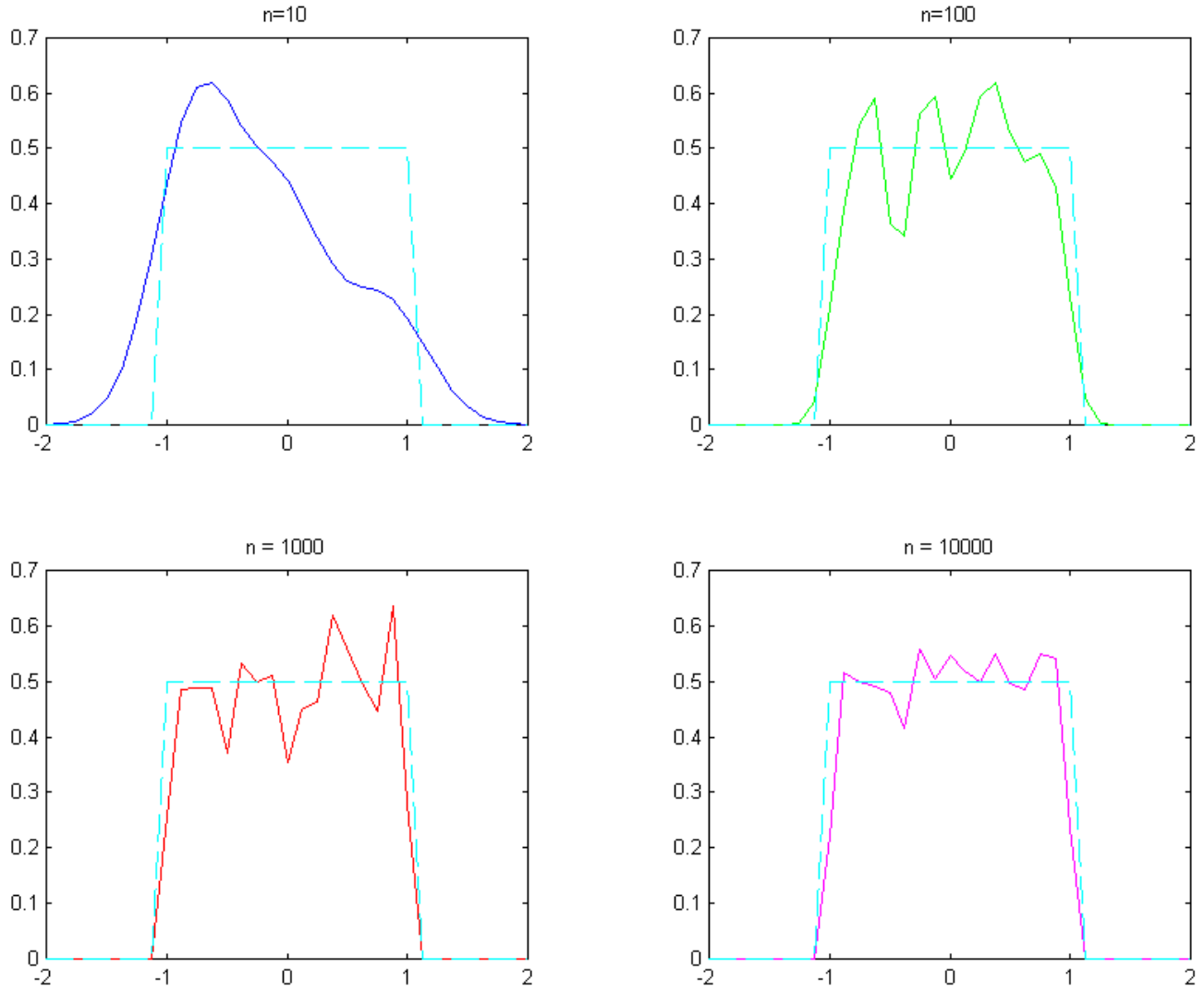
Funkcja okna dla 1, 2 i 3 próbek zbioru uczącego.

Estymacja pdf z oknem Parzena

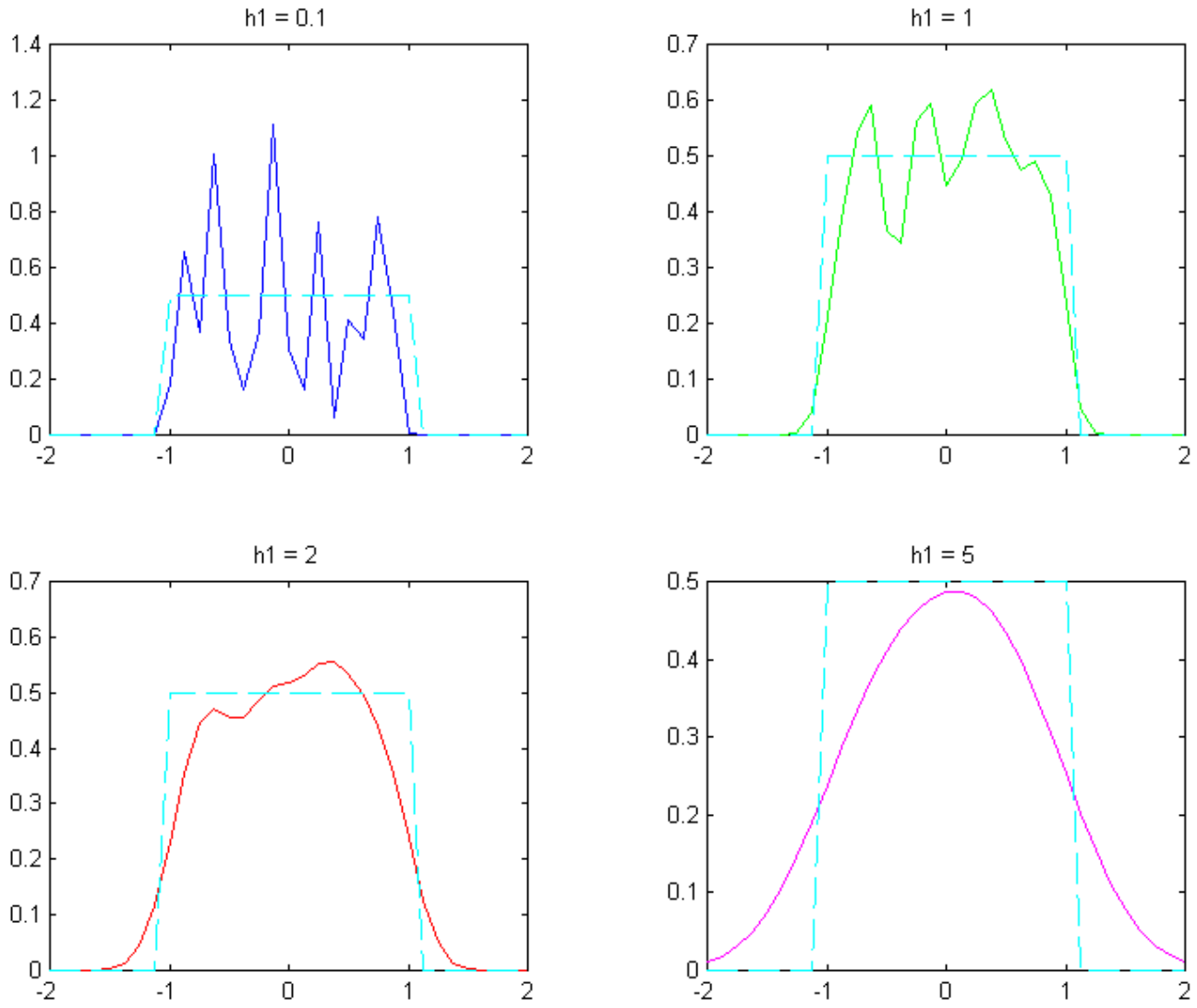
Ostatecznie, estymata *pdf* ma postać: $p_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\varphi\left(\frac{x - x_i}{h_n}\right)$



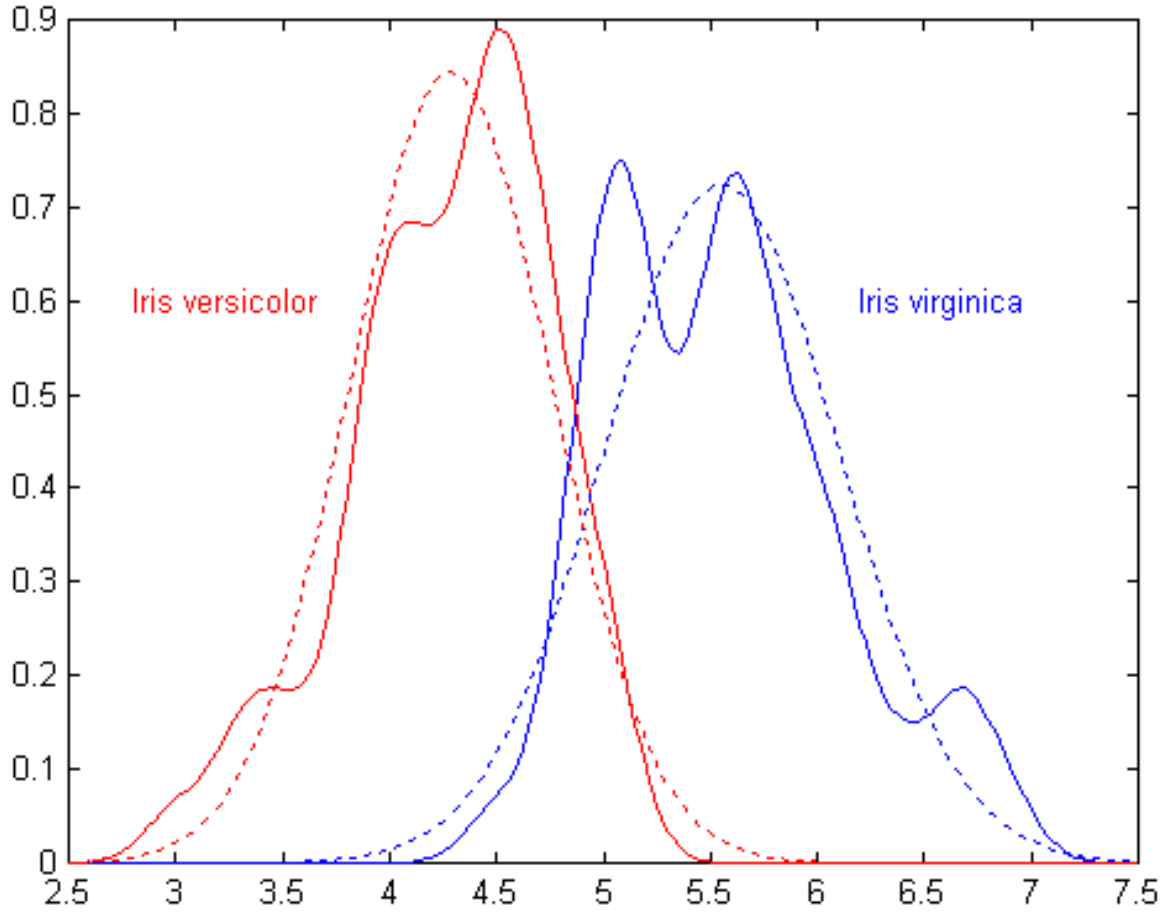
Okno Parzena dla różnych liczb próbek



Różne szerokości okna Parzena



Okno Parzena i dane kosaćców



Proszę zwrócić uwagę na bardzo małą różnicę granicy decyzyjnej w obu przypadkach.