Elementy sztucznej inteligencji

Sprawozdanie nr 1

Klasyfikator Bayesa 1 i 2 wymiarowy

Maciej Ciupa
maciek.ciupa@gmail.com

1 Treść zadania

Należy napisać program, który dla wybranej ilości (w zadaniu domyślnie dwóch) próbek zaklasyfikuje dane do odpowiednich klas, wcześniej estymując te dane i wyświetlając je na wykresach. Każdą z danych potrzebnych do klasyfikacji wg metody Bayesa można zmienić w programie.

2 Selekcja informacji

„Obserwacja przedmiotów, stanów różnego rodzaju obiektów i zjawisk wiąże się często z przeprowadzeniem ich klasyfikacji. Jest to dzielenie na grupy, do których zalicza się przedmioty, stany obiektów lub zjawiska różniące się wprawdzie, ale mające cechy wspólne.” ([1] str. 7)

Zbiór obserwowanych przedmiotów, stanów, zaliczanych do jednej grupy (ze względu na rozpatrywaną cechę) nazywamy ***klasą***. Każdy element klasy – ***obiektem***.

„Abstrahując od konkretnej treści rozpoznawanego obiektu będziemy przyjmować, że można go zawsze opisać za pomocą pewnego ***zbioru cech***. Przez pojęcie cech należy rozumieć nie same obiekty, lecz istotne i reprezentatywne charakterystyki lub ich odwzorowanie w postaci liczb, formuł, symboli itd.

Przypuśćmy, że w wyniku badania obiektu można z każdą z analizowanych jego ***N*** cech związać wartość liczbową. Wobec tego, formalnym opisem obiektu może być zbiór ***N*** liczb (…)” ([1] str. 8) tzw. wektor cech

$$\overline{\mathrm{x}} = \left(x_1, x_2, \ldots, x_\mathrm{N}\right)$$

Systemy, które dzielą obiekty (dane im odpowiadające) na klasy nazywane są ***klasyfikatorami***. Jednym z nich jest klasyfikator Bayesa wykorzystujący twierdzenie o prawdopodobieństwie zupełnym.

3 Teoretyczne podstawy klasyfikatora Bayesa

Klasyfikator ten wykorzystując

- macierz strat $\Lambda \equiv \{\lambda_{ij}\}$, gdzie λ_{ij} oznacza stratę wynikłą z przypisania obiektu klasy *i*-tej do klasy *j*-tej
- prawdopodobieństwa *a priori* $P(A_i)$ wystąpienia *i*-tej klasy
- rozkłady prawdopodobieństwa $P(\overline{\mathrm{x}} \mid A_i)$ wystąpienia obiektu $\overline{\mathrm{x}}$ w *i*-tej klasie.

przyporządkowuje obiekt $\mathrm{\bar{x}}$ do pewnej klasy $A_k, k \in [\mathrm{N}]$, dla której ryzyko jest najmniejsze tj.:

$$\mathrm{\bar{x}} \in A_k \Leftrightarrow k = \min\{i : R_i(\mathrm{\bar{x}}) \wedge i = 1,2,\ldots,\mathrm{N}\}$$

gdzie:

$$R_i(\mathrm{\bar{x}}) = \sum_{k=1}^{\mathrm{N}} \lambda_{ki} \mathrm{P}(\mathrm{A}_k \mid \mathrm{\bar{x}}), \qquad \mathrm{P}(A_k \mid \mathrm{\bar{x}}) = \frac{\mathrm{P}(A_k)\mathrm{P}(\mathrm{\bar{x}} \mid A_k)}{\sum_{i=1}^{\mathrm{N}} \mathrm{P}(A_i)\mathrm{P}(\mathrm{\bar{x}} \mid A_i)}$$

Błąd klasyfikacji

$$r = \frac{ilosc\_blednie\_zaklasyfikowanych}{ilosc\_wszystkich\_obiektów}$$

Funkcja graniczna dla dwóch klas

$$R_1 = R_2$$

$$(\lambda_{11} - \lambda_{12}) P(A_1 \mid x) = (\lambda_{22} - \lambda_{21}) P(A_2 \mid x)$$

$$\frac{P(A_1) \cdot P(x \mid A_1)}{P(A_2) \cdot P(x \mid A_2)} = \frac{(\lambda_{22} - \lambda_{21})}{(\lambda_{11} - \lambda_{12})}$$

a) przypadek jedno-wymiarowy $\bar{x} \equiv x$

gdzie $X \sim N(\mu_k, \sigma_k)$, $P(x \mid A_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x - \mu_k)^2}{2 \cdot \sigma_k^2} \right)$, $P(A_k) = p_k$

$$\boxed{Ax^2 + Bx + C = 0}$$

$$A = \frac{(\sigma_1^2 - \sigma_2^2)}{\sigma_1^2 \cdot \sigma_2^2}, \; B = -2\frac{(\sigma_1^2 \mu_2 - \sigma_2^2 \mu_1)}{\sigma_1^2 \cdot \sigma_2^2}, \; C = \frac{2}{\sigma_1^2 \cdot \sigma_2^2}\left(\ln\left(\frac{-p_1(\lambda_{1,1} - \lambda_{1,2})\sigma_2}{p_2(\lambda_{2,1} - \lambda_{2,2})\sigma_1} \right)\sigma_1^2\sigma_2^2 + \sigma_1^2\mu_2^2 - \sigma_2^2\mu_1^2 \right)$$

czyli:

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

b) przypadek dwu-wymiarowy

gdzie $(x_1, x_2) = X \sim N(\bar{\mu}_k, \Sigma_k)$, $P(\bar{x} \mid A_k) = \frac{1}{2\pi\sqrt{\det(\Sigma_k)}} \exp\left(-\frac{1}{2}(\bar{x} - \bar{\mu}_k)\Sigma_k^{-1}(\bar{x} - \bar{\mu}_k)^T \right)$, $P(A_k) = p_k$

$$\boxed{Ax_1^2 + Bx_2^2 + Cx_1 + Dx_2 + Ex_1x_2 + F - G = 0}$$

$\Sigma_k \equiv \{\sigma_{k,i,j}\}$, $\bar{\mu}_k \equiv (\mu_{k,1}, \mu_{k,2})$

$$A = \sigma_{1,1,1} - \sigma_{2,1,1}$$
$$B = \sigma_{1,2,2} - \sigma_{2,2,2}$$
$$C = -2\sigma_{1,1,1}\mu_{1,1} - \mu_{1,2}(\sigma_{1,1,2} + \sigma_{1,2,1}) + \sigma_{2,1,1}\mu_{2,1} + \mu_{2,2}(\sigma_{2,1,2} + \sigma_{2,2,1})$$
$$D = -\mu_{1,1}(\sigma_{1,1,2} + \sigma_{1,2,1}) - 2\sigma_{1,2,2}\mu_{1,2} + \mu_{2,1}(\sigma_{2,1,2} + \sigma_{2,2,1}) + 2\sigma_{2,2,2}\mu_{2,2}$$
$$E = \sigma_{1,1,2} + \sigma_{1,2,1} - \sigma_{2,1,2} - \sigma_{2,2,1}$$
$$F = \sigma_{1,1,1}\mu_{1,1}^2 + \mu_{1,1}\mu_{1,2}(\sigma_{1,1,2} + \sigma_{1,2,1}) + \sigma_{1,2,2}\mu_{1,2}^2 - \sigma_{2,1,1}\mu_{2,1}^2 - \mu_{2,2}(\mu_{2,1}(\sigma_{2,1,2} + \sigma_{2,2,1}) + \sigma_{2,2,2}\mu_{2,2})$$
$$G = 2\ln\left(\frac{p_1\sqrt{\det(\Sigma_2)}(\lambda_{1,1} - \lambda_{1,2})}{p_2\sqrt{\det(\Sigma_1)}(\lambda_{2,2} - \lambda_{2,1})}\right)$$

4 Klasyfikator 1-wymiarowy

Do sprawozdania załączony jest program ***Bayes1W*** napisany w środowisku Matlab, który dla

- zewnętrznego zbioru klas (menu *Plik* ►*Import danych z pliku …*), liczba klas dowolna (teoretycznie, powyżej 8 mała czytelność)
- prawdopodobieństwa a priori wystąpienia każdej z klas (pole *Spis prawdopodobieństw*)
- parametrów rozkładów prawdopodobieństwa wystąpienia obiektów w danych klasach (pole *Spis parametrów rozkładów*)
- macierzy strat (pole *Macierz strat*)

generuje funkcje ryzyka dla każdego z obiektów i klasyfikuje je do tej klasy, dla którego ryzyko jest najmniejsze. Oblicza przedziały wystąpienia dla każdej z klas, w przypadku dwóch klas generuje funkcje graniczną i po logarytmowaniu kwadratową. Wyznacza również ilość poprawnie/błędnie zaklasyfikowanych obiektów tj. błąd klasyfikacji.

Program uruchamiamy z linii poleceń, przy ustawionym aktualnym katalogu roboczym na ten, w którym znajdują się pliki programu:

- ***Bayes1W.m*** – główny program
- ***Bayes1W.fig*** – graficzne okno

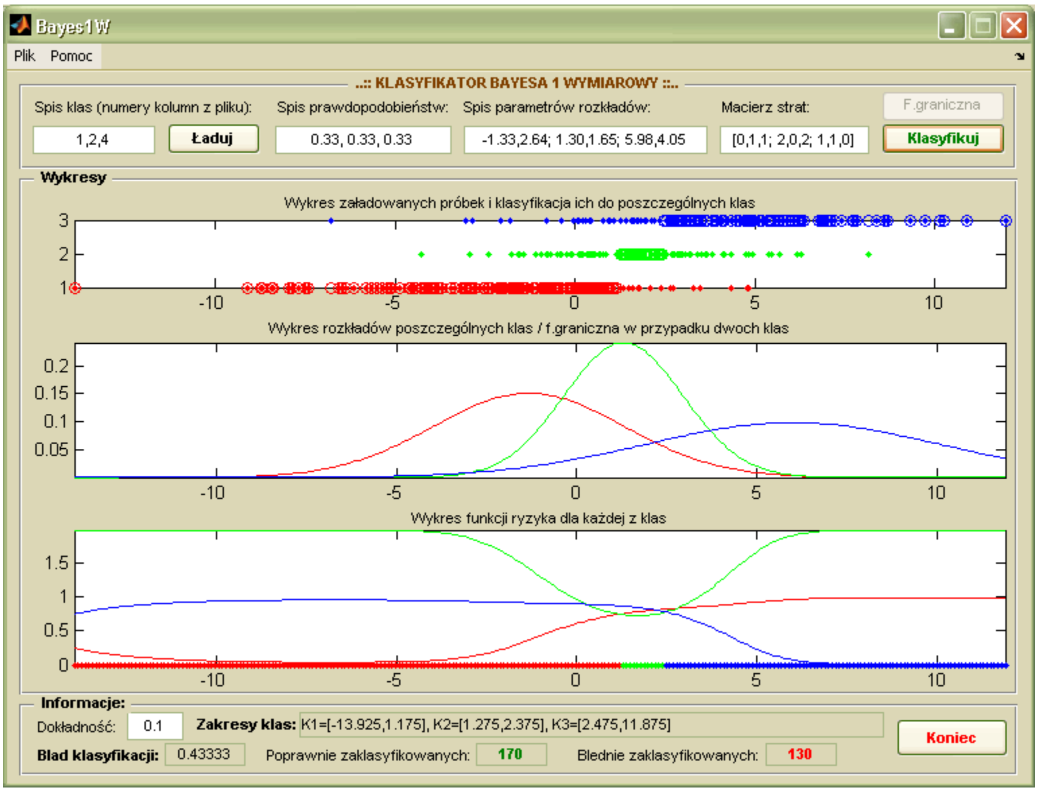
- ***getParams.m*** – funkcja zwracająca wektor średniej *mi* i odchylenie standardowe (macierz kowariancji) *sigma* 10% danych wybieranych losowo z zadanej próbki *Data*
  ```
  function [mi, sigma] = getParams(DATA)
  ```

- ***fnorm1W.m*** – funkcja zwracająca wartość gęstości rozkładu normalnego jednowymiarowego przy zadanych parametrach $N(\mu, \sigma)$
  ```
  function Y = fnorm(X, MI, SIGMA)
  ```

Algorytm klasyfikacji

1. Program pobiera dane z pliku i za pomocą funkcji ***getParams()*** określa parametry potrzebne do wyznaczenia rozkładu prawdopodobieństwa wystąpienia obiektu danej klasy.
2. Następnie korzystając z macierzy strat i prawdopodobieństwa całkowitego obliczane są funkcje ryzyka przypisania każdego z obiektów do każdej z klas.

3. Poszukiwana jest funkcja, dla której dla danego elementu wartość jest najmniejsza – do tej klasy obiekt zostaje zaklasyfikowany.



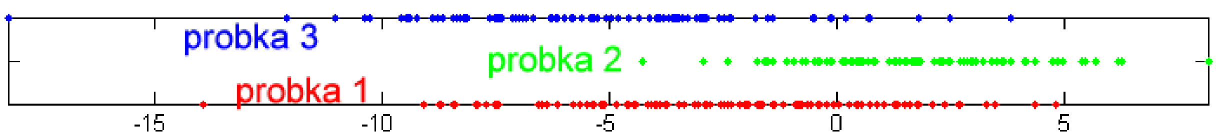
(rys. 1)

5 Analiza danych 1-wymiarowych

5.1 Domyślna klasyfikacja

Dane wejściowe:

Plik *dane.txt*, kolumny *1, 2, 3* (rys.2)



(rys. 2)

Estymacja danych (z 10% przypadkowo wybranych danych):

Próbka A_1: $\mu = -1.45$, $\sigma = 2.18$

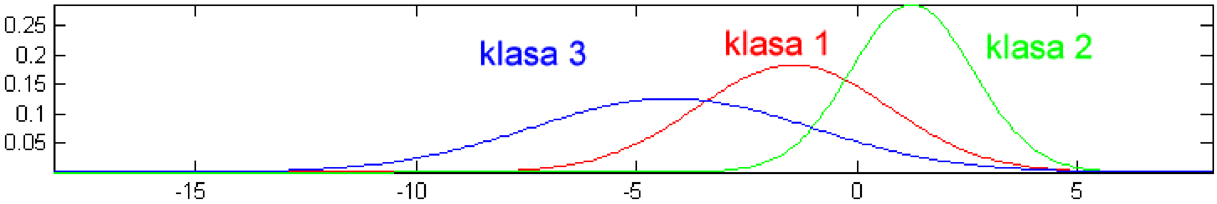
Próbka A_2: $\mu = 1.25$, $\sigma = 1.39$

Próbka A_3: $\mu = -4.21$, $\sigma = 3.17$

Macierz strat $$\Lambda = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

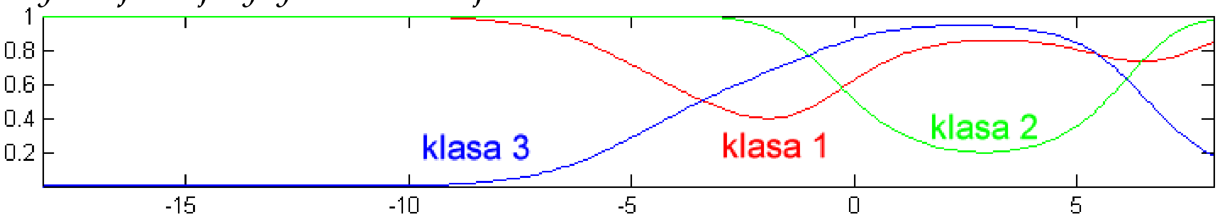
Prawdopodobieństwa a priori: $P(A_i) = 1/3$ dla $i = 1,2,3$

Wykres rozkładów $X_{A_i} \sim N(\mu, \sigma)$ dla $i = 1,2,3$



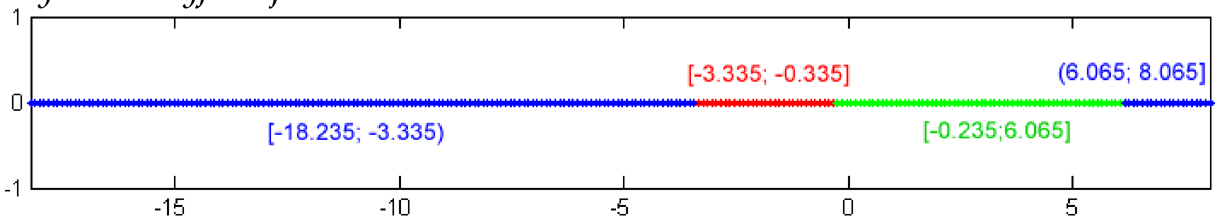
(rys. 3)

Wykres funkcji ryzyka dla każdej z klas:



Czyli wybieramy te klasy, dla których wartość funkcji jest najmniejsza.

Wykres klasyfikacji obiektów:



Program obliczył przedziały z dokładnością $\varepsilon = 0.1$

Czyli wynik klasyfikacji przedstawia się następująco:

Błąd klasyfikacji: $r = 0.38667$
Ilość poprawnie zaklasyfikowanych: 184
Ilość błędnie zaklasyfikowanych: 116

Błąd jest „*spory*", w następnym punkcie dokonamy zmian różnych parametrów i przyjrzymy się zmianie błędu klasyfikacji w szczególności.

5.2 Wpływ zmian parametrów na klasyfikację

5.2.1 Zmiana prawdopodobieństwa a priori

Dane początkowe: $P(A_1) = 0.33$, $P(A_2) = 0.33$, $P(A_3) = 0.33$
$r = 0.38667$

Dane końcowe: $P(A_1) = 0.24$, $P(A_2) = 0.38$, $P(A_3) = 0.38$
$r = 0.37667$

Choć generalnie zmiany tych parametrów powiększają błąd klasyfikacji.

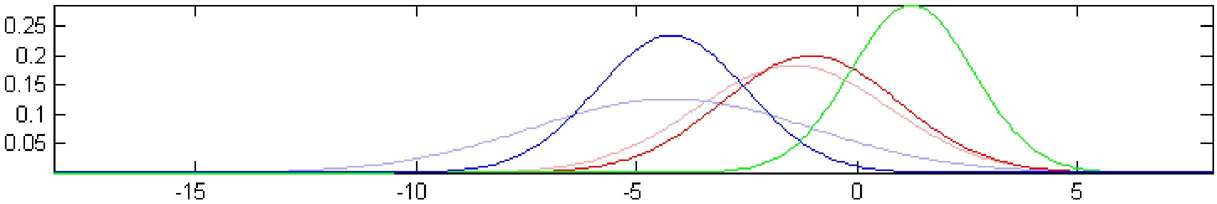
5.2.2 Zmiana parametrów rozkładów

Dane początkowe: $A_1: \mu = -1.45, \sigma = 2.18$, $A_2: \mu = 1.25, \sigma = 1.39$, $A_3: \mu = -4.21, \sigma = 3.17$

$r = 0.38667$

Dane końcowe: $A_1: \mu = -1.05, \sigma = 2.0$, $A_2: \mu = 1.25, \sigma = 1.39$, $A_3: \mu = -4.21, \sigma = 1.7$

$r = 0.36667$



(rys. 7)

Bledszym kolorem zostały oznaczone poprzednie rozkłady.
I tu również, zmiana na inne parametry powiększa błąd klasyfikacji.

5.2.3 Zmiana macierzy strat

Macierz zerojedynkowa (domyślna) dla tych danych znów okazuję się praktycznie najbardziej optymalna. Dla innych ustawień powiększamy tylko błąd klasyfikacji.

6 Klasyfikator 2-wymiarowy

Do sprawozdania załączony jest również klasyfikator 2-wymiarowy ***Bayes2W*** napisany w środowisku Matlab 7.2.0, który umożliwia

- Załadowanie dwóch klas z plików tekstowych odpowiedniego formatu (załączone do treści zadania - pliki *dom*.txt*), oznaczone odpowiednio Próbka 1 i Próbka 2 wraz z parametrami rozkładu normalnego opisującego owe dane.
- Ręczną zmianę parametrów rozkładów lub estymację 10% losowo wybranych danych
- Ręczną zmianę innych parametrów potrzebnych do klasyfikacji Bayesa, tj. macierzy strat i prawdopodobieństwa a priori
- Zaklasyfikowanie każdego z obiektów do jednej z dwóch klas
- Wygenerowanie funkcji granicznej dzielącej przestań $\boldsymbol{R}^2$ na dwa (w szczególności brak podziału, gdy minimalne ryzyko dla całej dziedziny jest przypisane jednej klasie) podzbiory klasyfikujące jej elementy do odpowiedniej klasy
- Obliczenie błędu klasyfikacji

Program uruchamiamy z linii poleceń, przy ustawionym aktualnym katalogu roboczym na ten, w którym znajdują się pliki programu:

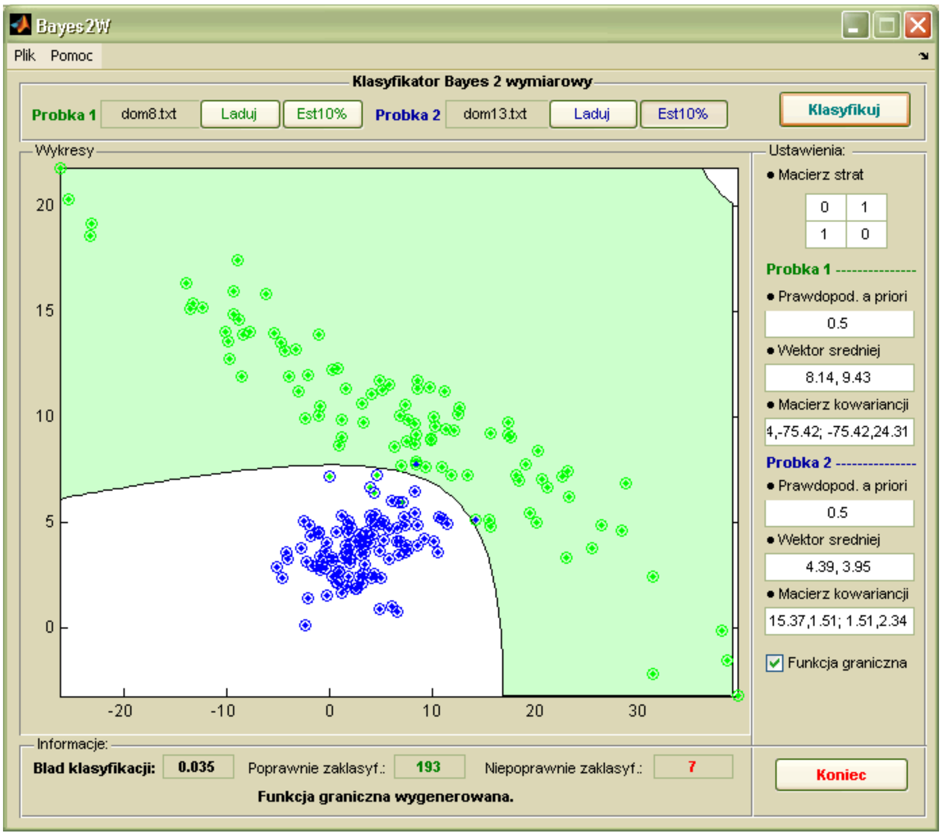
- ***Bayes2W.m*** – główny program
- ***Bayes2W.fig*** – graficzne okno

- ***getParams.m*** – funkcja zwracająca wektor średniej *mi* i odchylenie standardowe (macierz kowariancji) *sigma* 10% danych wybieranych losowo z zadanej próbki *Data*

```
function [mi, sigma] = getParams(DATA)
```

- *fnorm2W.m* – funkcja zwracająca wartość gęstości rozkładu normalnego dwuwymiarowego przy zadanych parametrach $N(\mu, \Sigma)$
  ```
  function Y = function Y = fnorm2W(X, Mi, invCov, detCov)
  ```
  gdzie invCov to odwrócona macierz kowariancji a detCov to wyznacznik macierzy kowariancji (aby przy każdym wywoływaniu nie obliczał jej ponownie)

Algorytm klasyfikacji jest analogiczny do tego w jednowymiarowym przypadku, z tym że funkcja graniczna jest funkcją dwóch zmiennych.



7 Analiza danych dwu-wymiarowych

Dla przykładu weźmy próbki:
Próbka 1 - dom1.txt
Probka 2 – dom13.txt

Domyślne dane rozkładów (pobrane z pliku) wynoszą odpowiednio:

Próbka 1: $\bar{\mu} = (5.3, 5.12)$, $\Sigma = \begin{bmatrix} 0.8 & 0.4 \\ 0.4 & 1.29 \end{bmatrix}$

Probka 2: $\bar{\mu} = (5.19, 5.07)$, $\Sigma = \begin{bmatrix} 0.24 & 0.36 \\ 0.36 & 1.18 \end{bmatrix}$

Błąd klasyfikacji: $r = 0.22$

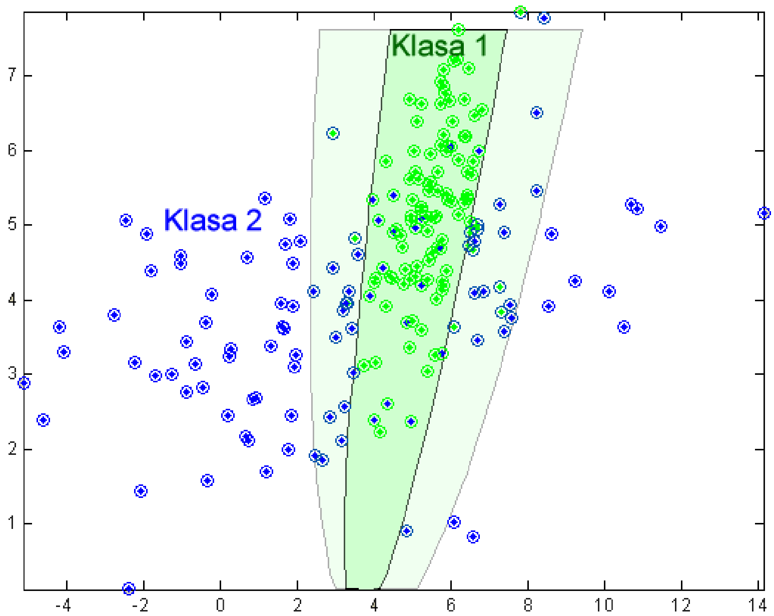
Dane uzyskane po estymacji 10% danych (wybranych przypadkowo) dla klasy 1:

Próbka 1: $\bar{\mu} = (5.19, 5.07)$, $\Sigma = \begin{bmatrix} 0.24 & 0.36 \\ 0.36 & 1.18 \end{bmatrix}$

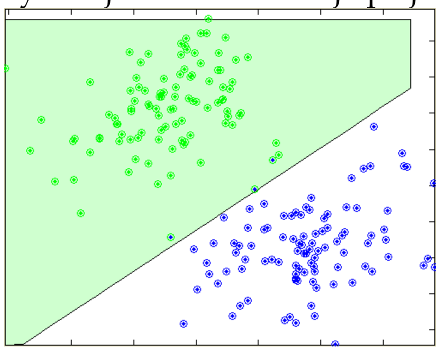
Probka 2: $\mu = (2.62, 3.61)$, $\Sigma = \begin{bmatrix} 14.2 & 1.84 \\ 1.84 & 1.58 \end{bmatrix}$

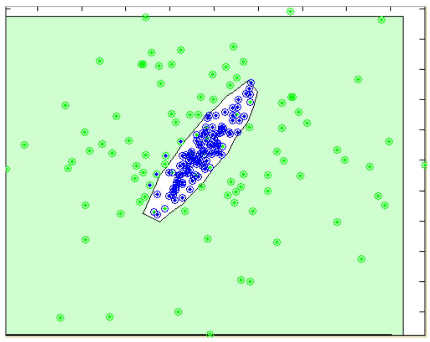
Błąd klasyfikacji: r = 0.145

Czyli modyfikacja parametrów może polepszyć wynik klasyfikacji. Wszystko zależy od rozmieszczenia klas względem siebie.

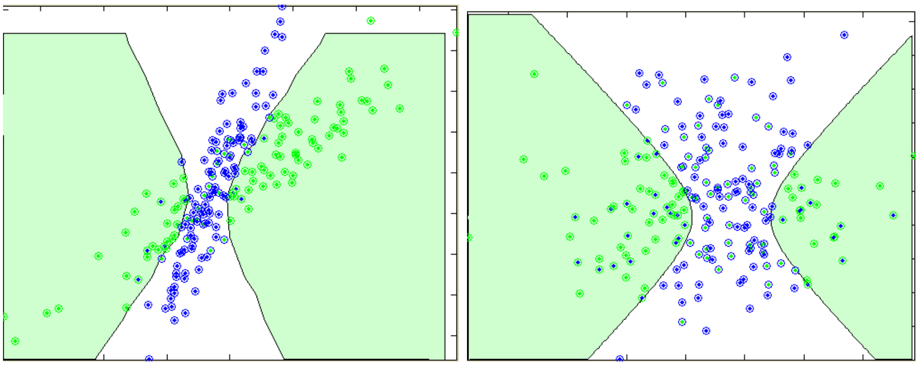


Dla klas zlokalizowanych w różnych obszarach lub dla danych o różnym współczynniku skupienia, klasyfikacja radzi sobie najlepiej.





Natomiast dla danych przenikających się o podobnym współczynniku skupienia błąd klasyfikacji rośnie (w drugim przypadku dochodzi nawet do 0.5)



8 Literatura

[1] Metody Selekcji informacji, Wojciech Sobczak, Witold Malina, WN-T W-wa 1978
[2] Klasyfikator Bayesa, materiały do laboratoriów ESI, mgr J. Rafałko, http://jrafalko.webpark.pl/kbi.doc