ROB
Laboratorium 2
Sprawozdanie
Maciej Ruszczyk

1. Wymagane funkcje umieszczone są w odpowiednich plikach i dołączone do sprawozdania.

2. W badanych danych zostały znalezione 2 odstające próbki o numerach kolejno 642 oraz 186. Próbkę 642 charakteryzują najmniejsze wartości, zaś próbkę 186 największe.

3. Klasyfikator Bayesa został stworzony przy wykorzystaniu cech numer 2 i 4. Plik ‘bayes.png’ przedstawia wykres obrazujący rozkład próbek przy wykorzystaniu tych 2 cech. Za szerokość okna Parzena została przyjęta wartość 0.001.
Klasyfikator dał następujące wartości błędów:

	pdfindep	pdfmulti	pdfparzen
ercf	2,63%	0,49%	2,4%

4. Wyniki uzyskane przy badaniu wpływu ilości próbek wykorzystanych do budowy klasyfikatora do jakości (wartości w tabeli oznaczają błędy klasyfikatora):

Procent próbek \ f.gęst.	Pdfindep	Pdfmulti	pdfparzen
10	3.13%	1.3%	9.23%
25	2.92%	0.68%	5.68%
50	2.71%	0.64%	3.73%
100	2.63%	0.49%	2.41%

Eksperyment przeprowadzony był 5 razy, a wyniki uśrednione. Próbki przydzielone są do 8 klas.
Z tabeli można zauważyć, że czym większa liczba próbek została użyta do stworzenia klasyfikatora tym lepsze wyniki daje przy klasyfikacji. Największy wpływ na ilość próbek miała metoda parzena, ponieważ błąd zmalał ok. 4 razy przy 10-krotnym zwiększeniu ilości próbek.
Poniższa tabela przedstawia wyniki tego samego eksperymentu z przydzielaniem próbek do 4 klas.

Procent próbek \ f.gęst.	Pdfindep	Pdfmulti	pdfparzen
10	1.05%	0.92%	6.04%
25	0.89%	0.37%	4.13%
50	0.86%	0.44%	3.07%
100	0.77%	0.33%	2.14%

Jakość klasyfikacji poprawiła się ze względu na fakt, że klasyfikator w tym przypadku przypisywał kartę do jednej talii, a nie do jednej z dwóch jak to było w poprzednim przypadku.

5. Na poniższej tabeli zostały ukazane wyniki uzyskane podczas testowania wpływu parametru h1 na jakość klasyfikacji. Tabela zawiera informacje o błędach podczas klasyfikacji.

Il. klas \ h1	0.0001	0.0005	0.001	0.005	0.01
8	2.85%	1.70%	2.41%	7.95%	13.93%
4	2.30%	1.48%	2.14%	4.33%	7.51%

Błąd przy zwiększaniu okna zmniejsza się, ale tylko do momentu przekroczenia wartości parametru h1. Potem zaczyna się powiększać.
Najmniejsza wartość błędu została uzyskana dla okna Parzena o szerokości 0.0005.
Warto zaznaczyć, że klasyfikacja dla 4 klas daje lepsze wyniki niż dla 8.
Pliki ‘parzen8.png’ oraz ‘parzen4.png’ to wykresy przedstawiające zależność błędu od wielkości h1.

6. W poniższej tabeli ukazano wartości błedu uzyskane dla różnych wartości funkcji prawdopodobieństwa przy badaniu wpływu zmian prawdopodobieństwa a priori. Uzyskane wyniki są średnią z 5 iteracji.

il. klas \ f. gęst	Pdfindep	Pdfmulti	Pdfparzen
8	3.20%	0.66%	1.97%
4	0.72%	0.44%	1.61%

Porównując uzyskane wyniki z wynikami z zadania 4, błąd klasyfikacji dla 8 klas zwiększył się dla metody pdfindep. Błąd przy zastosowaniu okna Parzena zmalał. W przypadku uwzględnienia 4 klas wszystkie błędy są porównywalne.

7. Normalizacja próbek zostaje wykonana zgodnie z poniższym wzorem:
Xnorm(i) = (x(i) – min(x)) / (max(x)-min(x)), gdzie x to wartość cechy, zaś i to indeks próbki.
Normalizacja jest potrzebna, ponieważ różnice między wielkościami cech są znaczne i może to mieć zły wpływ na klasyfikacje.
W poniższej tabeli porównano wielkości błędów uzyskanych przez klasyfikatory:

Ercf \ f.gęst	Pdfindep	Pdfmulti	Pdfparzen	cls1nn
błąd	2.63%	0.49%	28.78%	0.49%

Najbardziej zbliżoną metodą, która zawiera podobne wyniki jak klasyfikator 1-NN jest metoda pdfmulti. Pozostałe metody radzą sobie gorzej.