DATA MINING 2014 **Laboratorium 6 (ESTYMACJA GĘSTOŚCI PRAWDOPODOBIEŃSTWA+METODA K-NN)**

6.1
Wygeneruj 200- elementową mieszankę rozkładów

$$0.9 * N(5,1) + 0.1 * N(10,1).$$

a) Wyznacz estymator jądrowy który dość dobrze przybliżałby gęstość teoretyczną f. Narysuj wykresy gęstości mieszanki i estymatora gęstości oraz histogram.

b) Oblicz wartość empiryczną błędu średniokwadratowego

$$\frac{1}{512}\sum_{i=1}^{512}[f(x_i) - \hat{f}_n(x_i)]^2,$$

gdzie x_i, $i = 1, \ldots, 512$ są równoodległymi punktami podziału odcinka $[2, 12]$.

6.2
Dane `geyser(MASS)` dotyczą wybuchów gejzerów. Narysuj wykres estymatora gęstości dwuwymiarowej pary zmiennych (**duration**, **waiting**). Rozstępy wyznacz metodą Shealtera-Jonesa dla każdej zmiennej oddzielnie. Sporządź wykres dla wyznaczonego estymatora gęstości.

6.3
Dane: w pliku *earthquake.txt.*

a) Wyznacz estymatory gęstości dla zmiennej **body** w obu populacjach. Współczynnik wygładzający ustaw arbitralnie na **bw=0.2**. Sporządź wykresy uzyskanych estymatorów gęstości $f1$ i $f2$.
W podobny sposób wykonaj wykresy $\pi_1 f1$ i $\pi_2 f2$, uwzględniając prawdopodobieństwa apriori przynależności klasowej π_1, π_2. Sformułuj postać reguły Bayesowskiej, podaj przybliżoną wartość progu.

b) Wyznacz estymatory gęstości dwuwymiarowych pary zmiennych (**body**,**density**) w obu populacjach. Sporządź wykresy estymatorów i odpowiadające im wykresy konturowe, zastosuj funkcje **persp** i **contour**.

c) Dokonaj klasyfikacji metodą k-nn z $k = 3$. Jako próby uczącej użyj wszystkich obserwacji, a próbą testową będą te same obserwacje, czyli dokonamy reklasyfikacji. Wykonać wykres rozproszenia dla zmiennych **body** i **surface**. Obiekty z klasy *equake* oznaczyć literą "Q", a obiekty z klasy *explosn* literą "X". Wyrysować krzywą rozdzielającą klasy.

6.4
Dane **geny3PC** zawierają 62 obserwacje zmiennych objaśniających **X1**, **X2**, **X3** oraz zmiennej grupującej **grupa**. Obserwacje zostały uzyskane w wyniku przeprowadzonia redukcji danych Alizaheda zawierających 4026 wartości ekspresji genów dla każdego z 62 pacjentów chorych na jedną z 3 chorób: chłoniaka olbrzymiokomórkowego (DLCL), chłoniaka grudkowego (FL) oraz przewlekłą białaczkę limfatyczną (CLL). Najpierw wybrano 19 najbardziej isotnych genów. Zastosowano w tym celu metodę wielokrotnego podziału próby na uczącą i testową, a jako podzbiór zmiennych wybrano te które występowały najczęściej jako zmienne biorące udział w klasyfikacji za pomocą drzew decyzyjnych. Następnie dla oryginlanych wartości 19 genów wyznaczono 3

pierwsze składowe główne, które stanowią wartości zmiennych **X1**, **X2**, **X3**. Zmienna **grupa** przyjmuje wartości 1 (DLCL), 2 (CLL) lub 3 (FL). Skrypt który służy do wczytywania danych znajduje się w pliku *geny3PC.r*.

a) Wyznaczyć estymatory gęstości dwuwymiarowych dla pary zmiennych (**X1**, **X2**) w poszczególnych klasach DLCL, CLL i FL. Sporządzić wykresy estymatorów i wykresy konturowe.

b) Skonstruować empiryczny klasyfikator Bayesowski oparty o estymatory gęstości z punktu (a). Sporządzić wykresy konturowe oraz wyznaczyć krzywą rozdzielającą obszary klasyfikacji.

6.5

Porównaj działanie klasyfikatorów: Naiwnego Bayesa (funkcja `naiveBayes(e1071)`) oraz k-nn (funkcja `knn(class)`) na wybranym zbiorze danych.

6.6

Program Weka. Stosując metodę kroswalidacji 10-krotnej porównaj działanie omawianych dotychczas klasyfikatorów (model logistyczny, drzewa decyzyjne, Naiwny klasyfikator Bayesa) na wybranym zbiorze danych z repozytorium UCI.