DATA MINING 2017/2018

Laboratorium 5 (METODY OCENY KLASYFIKATORÓW)

5.1

Dane `BreastCancer(mlbench)` zawiera informację o 699 pacjentkach z podejrzeniem nowotworu piersi. Celem analizy jest stwierdzenie czy dany guz jest złośliwy (zmienna **Class="malignant"**) czy łagodny (zmienna **Class="benign"**). Każda z pacjentek jest opisywana 10 zmiennymi. W analizie należy pominąć pierwszą zmienną (**Id**).

a) Napisz skrypt który wykonuje kroswalidację K-krotną (K niech będzie parametrem). Wykorzystaj pakiet `cvTools`. Przetestuj wybrane metody klasyfikacji (drzewa, LDA, QDA, model logistyczny) szacując prawdopodobieństwo poprawnej klasyfikacji w oparciu o kroswalidację.

b) Oblicz prawdopodobieństwo poprawnej reklasyfikacji, porównaj z wynikami z punktów (a) i (b). W którym przypadku różnice są największe?

5.2

Dokonaj losowego podziału danych `BreastCancer(mlbench)` na część uczącą i testową. Na części treningowej dopasuj wybrany model klasyfikacyjny. Oszacuj:

- Czułość (sensitivity, recall): $P(\hat{Y} = +|Y = +)$,
- Specyficzność (specificity): $P(\hat{Y} = -|Y = -)$.
- Precyzje (precission): $P(Y = +|\hat{Y} = +)$.

5.3

Dane *urine.txt* dotyczą własności fizykochemicznych moczu. W zbiorze znajdują się następujące zmienne:

- **presence**- obecność kryształów (no, yes)
- **sg**- ciężar właściwy
- **ph**- wartość pH
- **mosm**- (ang. osmolarity)
- **mmho**- przewodnictwo
- **urea**- stężenie mocznika
- **calcium**- stężenie wapnia

Celem analizy jest stwierdzenie obecności kryształów (które mogą świadczyć o rozwoju kamieni nerkowych) na podstawie danych fizykochemicznych.

a) Rozpatrzmy regułę klasyfikacyjną $d_t(x)$ następującej postaci: Klasyfikuj element x do pierwszej populacji jeżeli

$$\hat{p}(1|x) > t,$$

gdzie prawdopodobieństwa aposteriori estymujemy używając modelu logistycznego uwzględniając wszystkie zmienne, natomiast t jest pewną nieujemną wartością progową. Dla wszystkich wartości t progów będących wartościami estymowanych prawdopodobieństw aposteriori dla elementów próby (uzupełnionych o wartości 0 i 1) wyznacz czułość (ang. sensitivity) oraz specy-

ficzność (and. specifity) a następnie narysuj krzywą ROC.

b) Powtórz polecenie z punktu (a) dla modelu mniejszego uwzględniającego jedynie zmienne **sg**, **mmho** oraz **urea**. Oba wykresy nanieść na jeden rysunek i porównać wyniki.

c) Obliczyć wskaźnik AUC (Area Under Curve) dla obu modeli.
d) Wyznacz krzywą ROC, parameter AUC korzystając z biblioteki `ROCR`