

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 krosvalidację 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 krosvalidację.
- b) Oszacuj prawdopodobieństwo poprawnej klasyfikacji stosując metodę wielokrotnego podziału próby na część uczącą i testową.
- c) Oblicz prawdopodobieństwo poprawnej rekasyfikacji, 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:

- Czulość (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

$$\hat{p}(1|x) > t\hat{p}(0|x),$$

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 specyficzność (ang. specificity) 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 LIFT (skumulowana liczba obserwacji z klasy wyróżnionej posortowana względem prawdopodobieństw aposteriori vs. liczba obserwacji) dla obu powyższych modeli.

5.4

Wczytaj dane **diabetes.arff** (znajdują się w katalogu UCIarff; użyj funkcji **read.arff(foreign)**). Podziel dane losowo na część treningową i testową, na części treningowej dopasuj model logistyczny. Wyznacz krzywą ROC, parameter AUC korzystając z biblioteki **ROCR**.

5.5

Wczytaj dane **iris.data**. Dokładny opis danych znajduje się na stronie <http://archive.ics.uci.edu/ml/datasets/iris>.

Rozpatrzmy klasyfikator oparty na metodzie najbliższego sąsiada KNN. Dokonaj estymacji prawdopodobieństwa błędnej klasyfikacji używając estymatora **bootstrap 0.632**.