

### 3.1

Zbiór danych *SAheart.data* (South African Heart Disease) zawiera dane dotyczące zapadalności na zawał serca wśród mężczyzn pomiędzy 15 a 64 rokiem życia. Zmienna **chd** oznacza że wystąpił (wartość 1) lub nie wystąpił (wartość 0) zawał serca. Dokładny opis danych znajduje się w pliku *SAheart.info*. Podzielić zbiór obserwacji na 2 podzbiory: uczący (pierwsze 324 obserwacje) i testowy (pozostałe obserwacje).

a) Rozważmy jedynie zmienne **chd** oraz **famhist**. Oblicz iloraz szans (ang. Odds Ratio) aby zbadać zależność między tymi dwoma zmiennymi. Podaj interpretację.

b)

- Dopasować model regresji logistycznej na podstawie zbioru uczącego.
- Które zmienne są istotne statystycznie w modelu pełnym.
- Używając metody eliminacji wstecznej z kryterium AIC oraz BIC dokonać selekcji zmiennych. Które zmienne zostają wybrane.
- Przetestować hipotezę że model pełny może być zastąpiony przez model mniejszy.

c) Oblicz iloraz szans (ang. odds ratio) w modelu logistycznym w przypadku kiedy wartości wszystkich zmiennych są ustalone, natomiast zwiększamy wiek pacjenta o jeden rok.

d) Na podstawie modelu regresji logistycznej (pełnego oraz wybranych na podstawie kryterium AIC i BIC) skonstruować klasyfikatory (na podstawie próby uczącej) i dokonać ich oceny obliczając błąd klasyfikacji na próbie testowej.

e) Na podstawie wszystkich zmiennych skonstruować na podstawie próby uczącej klasyfikatory oparte na metodach LDA i QDA. Porównać z klasyfikatorem opartym na dyskryminacji logistycznej obliczając błąd klasyfikacji na próbie testowej.

### 3.2

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) Dopasować model logistyczny. Oblicz procent poprawnych klasyfikacji dla próby uczącej.
- b) Wykonać  $n$ -krotną krosvalidację dokonując oceny metod opartych na dyskryminacji logistycznej, LDA i QDA.

### 3.3

Dane *earthquake.txt* dotyczą klasyfikacji wstrząsów na podstawie danych sejsmologicznych. Wykonaj wykres rozproszenia dla zmiennych **body** i **surface** z zaznaczeniem przynależności do klas. Dopasuj model regresji logistycznej opisujący zależność zmiennej **popn** od zmiennych **body** i **surface**. Jak wyjaśnić fakt że p-wartości statystyk Walda wskazują na nieistotność zmiennych objaśniających?

### 3.4

Dane Leukemia (w pliku Leukemia.RData) zawierają informację dotyczącą 72 pacjentów chorych na dwa rodzaje białaczki ( $y = 1$  lub  $y = 0$ ). Dla każdego pacjenta mamy 3571 wartości ekspresji genów. Użyj regularyzowanej wersji regresji logistycznej. Skorzystaj z pakietu **glmnet**. Przedstaw na wykresie profile estymowanych współczynników w zależności od wartości parametru kary.

### 3.5 (ZADANIE DOMOWE)

- Zaimplementuj metodę największej wiarygodności estymacji współczynników w modelu logistycznym (metoda nosi nazwę iteracyjnej metody ważonych najmniejszych kwadratów). Dokładny opis można znaleźć w książce 'The Elements of Statistical Learning', rozdział 4.4.1.
- Argumenty wejściowe: wektor obserwacji  $y$ , macierz eksperymentu  $X$ , punkt startowy  $\beta^0$ , liczba iteracji, warunek stopu.
- Można korzystać z gotowych implementacji metody ważonych najmniejszych kwadratów (np. w R funkcja `lm()` z argumentem `weights`).
- Porównać działanie z funkcją `glm()` w R: wartości estymowanych parametrów, czas działania.