

## Statystyka II. Egzamin komputerowy 04.09.2007.

1. Zbuduj model analizy wariancji dla danych `quine` z biblioteki `MASS`. W modelu tym zmienną objaśnianą jest `Days` - liczba opuszczonych dni nauki przez australijskie dzieci. Dobierz transformację zmiennej `Days` korzystając z funkcji `boxcox` i `logtrans` oraz wykonaj automatyczną selekcję zmiennych.
2. Na danych o zawartości pierwiastków w kilku rodzajach szkła: `library(MASS); data(fgl); x=fgl[1:185,]`, porównaj w eksperymencie weryfikacji krzyżowej (5/6 danych do trenowania) metody klasyfikacji zaimplementowane w funkcjach `rpart`, `nnet` oraz regresję logistyczną. Cecha przewidywana przyjmuje dwie wartości: `y = ifelse(X[,10]=="WinF",1,0)`. Policz współczynnik Goodmana-Kruskala między `y` oraz `y_pred` dla tych trzech metod.
3. Napisz funkcję rysującą wykres konturowy 2-wymiarowej gęstości  $f(x,y)$  tak, że warstwy ograniczają obszary o zadanych prawdopodobieństwach  $p=(p_1, \dots, p_k)$ . Na przykład, jeśli  $p=c(1,2,3,4)/5$ , to pierwsza warstwa jest zadana przez zbiór  $\{(x,y): f(x,y)=z, P(f(x,y) \geq z)=1/5\}$  dla pewnego  $z$ . Policz gęstość ( $Mg, Al$ ) za pomocą `kde2d` i narysuj warstwy dla danego wyżej  $p$  na danych o zawartości pierwiastków w kilku rodzajach szkła (`fgl[1:185,]`).
4. Załóżmy, że czas życia żarówki ( $C$ ) ma rozkład gamma z parametrami: `skala=2, kształt=4`. Wylosuj 1000 niezależnych obserwacji z tego rozkładu i oszacuj prawdopodobieństwo, że  $C$  jest nie mniejszy niż  $d=1.5*z_{0.25}$ , gdzie  $z_{0.25}$  jest kwantylem próbkowym rzędu 0.25 tego rozkładu (wskazówka: funkcja `quantile` oblicza kwantyle próbkowe). Policz  $d$  numerycznie (bez losowania).
5. Wczytaj z archiwum `MacierzePot11_3kol.zip` potencjały kontaktowe (PK), czyli 12 symetrycznych macierzy  $20 \times 20$  opisujących oddziaływania 210 par aminokwasów w białkach. PK są zapisane w plikach trójkolumnowych tak, że w pierwszej kolumnie znajduje się numer wiersza, w drugiej – numer kolumny, a w trzeciej - wartość odpowiedniego elementu macierzy PK o tych współrzędnych. Napisz ogólną funkcję `three2Vec`, która przekształca plik trójkolumnowy na wektor złożony z elementów macierzy PK należących do jej górnej części (ewentualnie dolnej) oraz przekątnej. Ze zbioru wektorów  $x_1, \dots, x_{12}$  otrzymanych z 12 PK, zbuduj macierz danych  $X=[x_1, \dots, x_{12}]$ . Sprawdź, czy w zbiorze  $X$  -- 210 obserwacji 12-wymiarowych są obserwacje odstające.