Statystyka II. Egzamin komputerowy 04.09.2007.

1. Wczytaj z archiwum *MacierzePot14_3kol.zip* potencjały kontaktowe (PK), czyli 12 symetrycznych macierzy 20 x 20 opisujących oddziaływania 210 par aminokwasów w białkach. PK są zapisane w plikach trójkolumnowych tak, że w pierwszej kolumnie znajduje sie numer wiersza, w drugiej – numer kolumny, a w trzeciej - wartośc 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 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{12}$ otrzymanych z 12 PK, zbuduj macierz danych X=[$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{12}$]. Sprawdz, czy w zbiorze X -- 210 obserwacji 12-wymiarowych są obserwacje odstające.

2. Zbuduj model analizy wariancji dla danych `immer` z biblioteki `MASS`. W modelu tym zmienną objaśnianą jest `Y = (Y1+Y2)/2` - średni plon jęczmienia z lat 1931-32. Dobierz transformację zmiennej `Y` korzystając z funkcji `boxcox` i `logtrans` oraz wykonaj automatyczną selekcję zmiennych.

3. Na danych `crabs` z biblioteki `MASS,` 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[,1]=="B",1,0)`. Policz współczynnik Goodmana-Kruskala między `y` oraz `y_pred` dla tych trzech metod.

4. Napisz funkcję rysującą wykres konturowy 2-wymiarowej gęstości $f(x,y)$ tak, że warstwice ograniczają obszary o zadanych prawdopodobieństwach $p=(p_1,\ldots p_k)$. Na przykład, jeśli $p = c(1,2,3,4)/5$, to pierwsza warstwica jest zadana przez zbiór $\{(x,y) : f(x,y) = z,\ P(f(x,y) \geq z) = 1/5\}$ dla pewnego z. Policz gęstosc (*FL*, *RW*) za pomocą `kde2d` i narysuj warstwice dla danego wyżej p na danych `crabs`.

5. Załóżmy, że czas życia żarówki (C) ma rozkład gamma z parametrami: skala=2, kształt=4. Wylosuj1000 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).