Statystyka II Mat. Egzamin komputerowy 20.06.07

1. Napisz (bez wykorzystania `eigen` oraz `svd`) funkcję `rnormMV(n,m,S)` generującą próbę prostą o liczności `n` z wielowymiarowego rozkładu normalnego o wektorze wartości oczekiwanej `m` oraz nieosobliwej macierzy kowariancji `S`.

2. Wczytaj z pliku *MacierzePot7.zip* 12 potencjałów kontaktowych (PK). PK są to symetryczne macierze 20 x 20 opisujące oddziaływania 210 par aminokwasów w białkach. Napisz funkcję `upperVec`, która przekształca macierz na wektor złożony z elementów należących do jej górnej części oraz przekątnej. Ze zbioru wektorów $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{12}$ otrzymanych z 12 PK, zbuduj macierz danych $X=[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{12}]^T$. Czy w zbiorze X -- 210 obserwacji 12-wymiarowych są obserwacje odstające?

3. Wykonaj klasteryzację aglomeracyjną obserwacji ze zbioru X metodą complete linkage. Dendrogram zapisz do pliku .e*ps*. Narysuj oddziaływania między aminokwasami na płaszczyźnie dwóch pierwszych składowych głównych i policz udział wariancji tych składowych w całkowitej wariancji danych. Ile jest klastrów (odpowiedź uzasadnij wykresem separowalności) ?

4. Zbuduj model regresji liniowej wielu zmiennych dla danych `X=data.frame(state.x77,row.names=state.abb,check.names=T)`. Cechą objaśnianą ma być `Life.Exp`.

5. Wczytaj dane demograficzne o szkołach holenderskich: `library(nlme); data(MathAchSchool);X=MathAchSchool[,-c(1,6)].` Policz współczynnik Goodmanna-Kruskala tau(PRACAD, DISCLIM). Porównaj na tych danych model logistyczny i drzewa decyzyjne z `rpart.` Zakładamy, że cecha przewidywana `y=ifelse(X[,2]=="Catholic",1,0)`. Wykonaj weryfikację krzyżową (4/5 danych do trenowania) i policz prawdopodobieństrwo poprawnej predykcji dla obu metod.