Statystyka II Mat. Egzamin komputerowy 20.06.07

1. Zbuduj model regresji liniowej wielu zmiennych dla danych bodyfatJohnson.txt bez 1szej kolumny. Cechą przewidywaną ma byc "Bodyfat".

2. 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`.

3. Wczytaj dane demograficzne o szkołach holenderskich: `library(nlme); data(MathAchSchool);X=MathAchSchool[,-c(1,6)].` Policz wspólną informację M(PRACAD, DISCLIM). Porównaj na tych danych model wielomianowy multinom z biblioteki `nnet` 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.

4. Wczytaj z pliku *MacierzePot6.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 $\boldsymbol{x}_1$, … , $\boldsymbol{x}_{12}$ otrzymanych z 12 PK, zbuduj macierz danych X=[$\boldsymbol{x}_1$, …, $\boldsymbol{x}_{12}$]^T. Czy w zbiorze X -- 210 obserwacji 12-wymiarowych są obserwacje odstające?

5. Wykonaj klasteryzację aglomeracyjną obserwacji ze zbioru X metodą complete linkage. Dendrogram zapisz do pliku *.eps*. 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) ?