Statystyka II Mat. Egzamin komputerowy 20.06.07

1. Wczytaj dane o zawartości pierwiastków w kilku rodzajach szkła: `library(MASS);X=fgl[1:185,].` Policz współczynnik Goodmanna-Kruskala tau(Mg, Al). Porównaj na tych danych klasyfikację za pomocą modelu wielomianowego `multinom` z biblioteki `nnet` oraz drzewa decyzyjne z biblioteki `rpart.` Zakładamy, że cecha przewidywana `y=ifelse(X[,10]=="WinF",1,0)`. Wykonaj weryfikację krzyżową (4/5 danych do trenowania) i policz prawdopodobieństrwo poprawnej predykcji dla obu metod.
2. Napisz odpowiednik, uogólnienie funckji `lm` dla modelu $y_i = \beta^T \mathbf{x}_i + \varepsilon_i$ w którym $\varepsilon \sim N(0, \sigma^2 W)$, gdzie W jest znaną macierzą dodatnio określoną.

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

4. Wczytaj z archiwum *MacierzePot5.zip* 12 potencjałów kontaktowych (PK). PK są to symetryczne macierze 20 x 20 opisujące oddziaływania 210 par aminokwasów w białkach. Napisz ogólną funkcję `upperVec`, która przekształca macierz na wektor złożony z elementów 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 $X=[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{12}]^T$. Czy w zbiorze X -- 210 obserwacji 12-wymiarowych są obserwacje odstające?

5. Wykonaj klasteryzację aglomeracyjną obserwacji ze zbioru X metodą Warda. 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) ?