Statystyka II Mat. Egzamin komputerowy 20.06.07

1. Wczytaj z archiwum *MacierzePot4.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?

2. Wykonaj klasteryzację aglomeracyjną obserwacji ze zbioru X metodą McQuitty'ego. 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) ?
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 metody klasyfikacji `nnet` i `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. 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.

5. Zbuduj model regresji liniowej wielu zmiennych dla przefiltrowanych danych Cars93:

```
library(MASS)
X=na.omit(Cars93[,c(2,7,12:15,17,19,21,22,24,25)])
row.names(X)=X[,1]; X=X[,-1]
```