Statystyka II Mat. Egzamin komputerowy 12.06.09.

1. Wczytaj z pliku
   http://www.mimuw.edu.pl/~pokar/StatystykaII/DANE/AAindex
tabele własności (wiersze tabeli) fizykochemicznych aminokwasów. Za pomocą funkcji `boxplot.stat` sprawdź, czy w zbiorze X są obserwacje odstające?. Wykonaj klasteryzację hierarchiczną `hclust` metodą `"centroid"` tych danych. Dendrogram zapisz do pliku .*eps*. Na podstawie wykresu separowalności oszacuj liczbę klastów. Napisz ogólną funkcję `reduceDim(X,proc)`, która przekształca macierz danych X na macierz złożoną z jej k-pierwszych składowych głównych, gdzie k jest minimalną liczbą składowych, których łączna wariancja stanowi przynajmniej `proc` procent całkowitej wariancji X. Jakie jest k dla `proc`=80 ?

2. Porównaj metody klasyfikacji *`lda`*, *`qda`*, *`glm`* (regresja logistyczna) w eksperymencie kroswalidacji sześciokrotnej na danych `X=na.omit(biopsy[,-1])` z biblioteki `MASS`. Cecha przewidywana `y` przyjmuje 2 wartości: „benign" i „malignant. Jako miarę efektywności predykcji przyjmij wspólną informację zawartą w `y` oraz `y_pred`.

3. Niech dany będzie model liniowy $\mathbf{y} = \boldsymbol{\beta}^T\mathbf{X} + \boldsymbol{\varepsilon}$. Wykorzystując rozkład SVD macierzy **X** (funkcja `svd`), bez użycia pętli, bez mnożenia i odwracania macierzy, napisz własną wersję obliczającą wpływy obserwacji h[i]=H[i,i]. Argumentem funkcji ma być tylko macierz cech obserwowalnych (predyktorów) X.

4. Napisz odpowiednik, uogólnienie funckji `lm` dla modelu $\mathbf{y} = \boldsymbol{\beta}^T\mathbf{X} + \boldsymbol{\varepsilon}$. w którym $\boldsymbol{\varepsilon} \sim N(0,\sigma^2 W)$, gdzie W jest znaną macierzą dodatnio określoną.

5. Dla modelu regresji $\mathbf{y} = \boldsymbol{\beta}^T\mathbf{X} + \boldsymbol{\varepsilon}$:
   - policz (dowolną metodą) p-wartość testu F-Snedecora hipotezy: *współczynniki przy trzech ostatnich cechach są równe zero*;
   - policz p-wartość tej samej statystyki wykorzystując tylko `y=X[,1]`, rozkład QR macierzy X=`cbind(1,X[,-1])` oraz funkcję `pf`.

6. Zbuduj model regresji liniowej wielu zmiennych dla przefiltrowanych danych Cars93: library(MASS) ; X=na.omit(Cars93[,c(2,7,12:15,17,19,21,22,24,25,26)]); row.names(X)=X[,1]; X=X[,-1]. Cechą przewidywaną jest `MPG.city`. Proszę nie NIE przeprowadzać transformacji cech `boxcox` i `logtrans`.