Statystyka II Mat. Egzamin komputerowy 12.06.10.

1. Porównaj metody klasyfikacji *lda*, *qda*, *glm* (regresja logistyczna) w eksperymencie kroswalidacji pięciokrotnej na danych `biopsy` z biblioteki MASS:
   `Data=na.omit(biopsy[,-1]);`
   `y=ifelse(Data[,10]=="malignant",1,0);  Data=Data[,-10]`
   Jako miarę dokładności klasyfikacji przyjmij wspólną informację zawartą w `y` oraz `y_pred`.

2. Niech dany będzie model regresji logistycznej dla danych z poprzedniego zadania. Policz (asymptotyczną) p-wartość testu ilorazu wiarygodności hipotezy: *współczynniki przy V1, V2 i V3 są sobie równe.*

3. Wczytaj z katalogu
   http://www.mimuw.edu.pl/~pokar/StatystykaII/DANE/DaneEgzaminacyjne/
   archiwum *MacierzePot3.zip* 12 potencjałów kontaktowych (PK). PK są to symetryczne macierze 20 x 20 opisujące oddziaływania 210 par aminokwasów w białkach. Napisz ogólną funkcję lower`Vec`, która przekształca macierz na wektor złożony z elementów należących do jej dolnej części bez przekątnej. Ze zbioru wektorów $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{12}$ otrzymanych z 12 PK, zbuduj macierz danych X=[$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{12}$] (190 obserwacji 12-wymiarowych). Za pomocą funkcji `boxplot.stat` sprawdź, czy w zbiorze X są obserwacje odstające?

4. Wykonaj klasteryzację metodą k-medoidów danych z poprzedniego zadania. Na podstawie wykresu separowalności oszacuj liczbę klastrów. Wskazówka: http://www.mimuw.edu.pl/~pokar/StatystykaII/EKSPLORACJA/sep_syl.r
   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`=85 ?

5. Niech dany będzie model liniowy $\mathbf{y} = \boldsymbol{\beta}^T\mathbf{X} + \boldsymbol{\varepsilon}$. Wykorzystując rozkład QR macierzy **X** bez użycia pętli, bez mnożenia i odwracania macierzy, napisz własną funkcję `hatvalues`. Argumentem funkcji ma być tylko macierz cech obserwowalnych (predyktorów) X.

6. Zbuduj model analizy wariancji dla danych `quine` z biblioteki MASS. Cechą objaśnianą jest `Days`. Sprawdź, czy transformacja logarytmiczna `Days` poprawia dopasowanie modelu. Policz p-wartość testu F-Snedecora hipotezy: *czynnik Age można pominąć.*