Statystyka II Mat. Egzamin komputerowy 13.06.08

1. Porównaj metody klasyfikacji `lda`, `qda`, `glm` (regresja logistyczna) w eksperymencie kroswalidacji pięciokrotnej 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`.

2. 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]. Cechą przewidywaną jest `MPG.city`.

3. Dla modelu regresji z poprzedniego zadania:
   - 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`.

4. Napisz funkcję rysującą wykres konturowy 2-wymiarowej gęstości $f(x,y)$ tak, że warstwice ograniczają obszary o zadanych prawdopodobieństwach $p=(p_1,...p_k)$. Na przykład, jeśli $p = c(1,2,3,4)/5$, to pierwsza warstwica jest zadana przez zbiór $\{(x,y): f(x,y) = z,\ P(f(x,y) \geq z) = 1/5\}$ dla pewnego z. Dla danych z zadania 2. policz gęstość `(MPG.city,EngineSize)` za pomocą `kde2d` i narysuj warstwice dla danego wyżej p.

5. Wczytaj z archiwum `MacierzePot16_3kol.zip` potencjały kontaktowe (PK), czyli 12 symetrycznych macierzy 20 x 20 opisujących oddziaływania 210 par aminokwasów w białkach. PK są zapisane w plikach trójkolumnowych tak, że w pierwszej kolumnie znajduje się numer wiersza, w drugiej – numer kolumny, a w trzeciej – wartość elementu macierzy PK o tych współrzędnych. Napisz ogólną funkcję `three2Vec`, która przekształca plik trójkolumnowy na wektor złożony z elementów macierzy PK należących do jej dolnej części bez przekątnej. Ze zbioru wektorów $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{12}$ otrzymanych z 12tu 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?