Statystyka II Mat. Egzamin komputerowy 16.06.15.

Wszystkie potrzebne dane są w katalogu
http://www.mimuw.edu.pl/~pokar/StatystykaII/DANE/DaneEgzaminacyjne/

1. Dla danych *diabetes* zbuduj predykcyjny model liniowy dla kolumny `y`, to znaczy policz estymator $\boldsymbol{\beta}$ za pomocą regresji grzbietowej. Karę wybierz w eksperymencie pięciokrotnej kroswalidacji.

2. Zbuduj pełny model logistyczny dla danych *Prostate Cancer*, w którym zmienną wyjaśnianą jest *Tumor Penetration*. Predyktory *Rase* oraz *Tumor Volume* przekształć za pomocą `as.numeric`. (i) Za pomocą ilorazu wiarygodności przetestuj na poziomie ufności 0.95 hipotezę, że w modelu pełnym zbiór predyktorów: *Age*, *Rase*, *Tumor Volume* oraz *Capsular Involvement* jest niepotrzebny. (ii) Wybierz (pod)model za pomocą kryterium Akaike'go.

3. Porównaj metody klasyfikacji *`lda`, `qda`, `glm`* w eksperymencie kroswalidacji dziesięciokrotnej na danych *Prostate Cancer*. Cecha przewidywaną `y` jest *Tumor Penetration*. Jako miarę dokładności klasyfikacji przyjmij wspólną informację zawartą w `y` oraz `y_pred`.

4. Wczytaj archiwum *MacierzePot12_3kol.zip* 12 potencjałów kontaktowych (PK) w formacie 3-kolumnowym: numer_wiersza numer_kolumny element_macierzy. 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). Wykonaj klasteryzację metodą k-means danych z poprzedniego zadania. Na podstawie wykresu separowalności oszacuj liczbę klastów.

5. 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=90` ?