Statystyka II Mat. Egzamin komputerowy 16.06.12.

1. Niech dany będzie model liniowy $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$. Wykorzystując rozkład SVD macierzy **X** (funkcja `svd`), bez użycia mnożenia macierzowego i odwracania macierzy, napisz ogólną funkcję obliczającą nieobciążone estymatory wariancji estymatorów elementów wektora **β** otrzymanych metodą najmniejszych kwadratów. Argumentami funkcji mają być **y** oraz **X**.

2. Policz streszczenie modelu normalnej regresji liniowej, w którym $\mathbf{E}(\log(\mathbf{y})) = \mathbf{X\beta}$ oraz streszczenie modelu poissonowskiej regresji logliniowej, w którym $\log(\mathbf{E}(\mathbf{y})) = \mathbf{X\beta}$ na danych Pima=rbind(Pima.te,Pima.tr) z biblioteki MASS, gdzie **y**=npreg+1. Następnie policz p-wartości testu ilorazu wiarygodności (dla modelu normalnego – dokładną, dla modelu logliniowego – przybliżoną) dla hipotezy <=> modelu mniejszego **y** ~ glu+age+type przeciwko alternatywie <=> modelu pełnego **y** ~ . .

3. 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 dokładności klasyfikacji przyjmij wspólną informację zawartą w `y` oraz `y_pred`.

4. Wczytaj z katalogu
http://www.mimuw.edu.pl/~pokar/StatystykaII/DANE/DaneEgzaminacyjne/
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). Za pomocą funkcji `boxplot.stat` sprawdź, czy w zbiorze X są obserwacje odstające?

5. Wykonaj klasteryzację metodą k-means danych z poprzedniego zadania. 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`=90 ?