Statystyka II Mat. Egzamin komputerowy 24.06.15.

1. Niech czas życia żarówki C ma rozkład gamma z parametrami: skala=1, kształt=3. (i) Wylosuj 1000 niezależnych obserwacji z tego rozkładu i oszacuj p-stwo, że C jest nie większy niż d=1.5*z_0.75, gdzie z_0.75 jest kwantylem próbkowym rzędu 0.75 tego rozkładu (wskazówka: użyj funkcji quantile). (ii) Policz d numerycznie (bez losowania).

2. Policz streszczenie modelu normalnej regresji liniowej, w którym **E(**log(**y**)) = **Xβ** oraz streszczenie modelu poissonowskiej regresji logliniowej, w którym log(**E**(**y**)) = **Xβ** 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: „**y** ~ glu+age+type" przeciwko alternatywie „**y** ~ ." .

3. Zbuduj model liniowy dla danych prostate z pliku:
http://www.mimuw.edu.pl/~pokar/StatystykaII/DANE/DaneEgzaminacyjne/prostate.data
Cechą objaśnianą jest 9-ta kolumna (lpsa), a predyktory są w kolumnach 1-8. W eksperymencie pięciokrotnej kroswalidacji oszacuj średniokwadratowy błąd predykcji dla tego modelu.

4. Wczytaj z katalogu
http://www.mimuw.edu.pl/~pokar/StatystykaII/DANE/DaneEgzaminacyjne/
archiwum MacierzePot1.zip 12 potencjałów kontaktowych (PK) zawierających energię oddziaływania między aminokwasami w białkach. (PK są to symetryczne macierze 20 x 20). Napisz funkcję upperVec, która przekształca macierz na wektor złożony z elementów należących do jej górnej części. 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 tych danych. Na podstawie wykresu separowalności oszacuj liczbę klastów.

5. Wykorzystując funkcję `eigen`, a nie tak jak zwykle prcomp czy `svd` 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` oraz dla X z poprzedniego zadania ?