Statystyka II Mat. Egzamin komputerowy 12.06.09.

1. Porównaj metody klasyfikacji *lda*, *qda*, *glm* (regresja logistyczna) w eksperymencie kroswalidacji pięciokrotnej na danych `MathAchSchool[,-c(1,6)]` z biblioteki nlme. Cecha przewidywana `y` przyjmuje dwie wartości: „Catholic" i „Public". Jako miarę efektywności predykcji przyjmij wspólną informację zawartą w `y` oraz `y_pred`.

2. Niech dany będzie model liniowy $\mathbf{y} = \boldsymbol{\beta}^T\mathbf{X} + \boldsymbol{\varepsilon}$. Wykorzystując rozkład SVD macierzy **X** (funkcja `svd`), bez użycia pętli i bez mnożenia macierzy, napisz ogólną funkcję obliczającą przekątną macierzy $(X^TX)^{-1}$. Argumentem funkcji ma być tylko macierz cech obserwowalnych (predyktorów) X. Odwracanie macierzy trójkątnej wykonaj za pomocą funkcji `backsolve`.

3. Wczytaj z pliku
   http://www.mimuw.edu.pl/~pokar/StatystykaII/DANE/AAindex
tabele własności (wiersze tabeli) fizykochemicznych aminokwasów. Za pomocą funkcji `boxplot.stat` sprawdź, czy w zbiorze X są obserwacje odstające?. Wykonaj klasteryzację hierarchiczną `hclust` metodą `"centroid"` tych danych. Dendrogram zapisz do pliku .*eps*. 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`=80 ?

4. Dla modelu regresji $\mathbf{y} = \boldsymbol{\beta}^T\mathbf{X} + \boldsymbol{\varepsilon}$:
   - 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`.

5. Napisz funkcję `rnormMV(n,m,S)` generującą prostą próbę `n`-elementową z wielowymiarowego rozkładu normalnego o wektorze średnich **`m`** i dodatnio określonej macierzy kowariancji **`S`**. Dekompozycję **`S`** wykonaj za pomocą rozkładu Choleskiego.

6. Zbuduj model regresji liniowej wielu zmiennych dla długości życia `Life.Exp` wykorzystując pozostałe cechy ze zbioru danych `state.x77`:
   `X=data.frame(state.x77,row.names=state.abb,check.names=T)`
   Proszę nie NIE przeprowadzać transformacji cech `boxcox` i `logtrans`.
   .