

## Statystyka II Mat. Egzamin komputerowy 13.06.08

1. Porównaj metody klasyfikacji *lda*, *qda*, *glm* (regresja logistyczna) w eksperymencie krosvalidacji 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. Wczytaj z archiwum *MacierzePot1.zip* 12 potencjałów kontaktowych (PK). PK są to symetryczne macierze 20 x 20 opisujące oddziaływania 210 par aminokwasów w białkach. Napisz ogólną funkcję `lowerVec`, która przekształca macierz na wektor złożony z elementów należących do jej dolnej części bez przekątnej. Ze zbioru wektorów  $\mathbf{x}_1, \dots, \mathbf{x}_{12}$  otrzymanych z 12 PK, zbuduj macierz danych  $X=[\mathbf{x}_1, \dots, \mathbf{x}_{12}]$  (190 obserwacji 12-wymiarowych). Za pomocą funkcji `boxplot.stat` sprawdź, czy w zbiorze *X* są obserwacje odstające?
3. Wykonaj klasteryzację hierarchiczną `hclust` metodą "median" danych z poprzedniego zadania. Dendrogram zapisz do pliku `.eps`. Na podstawie wykresu separowalności i sylwetki 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=75` ?
4. Wykorzystując rozkład  $X=QR$ , bez użycia pętli i bez mnożenia macierzy, napisz ogólną funkcję obliczającą przekątną macierzy  $(X^T X)^{-1}$ . Argumentem funkcji ma być tylko macierz cech obserwowalnych (predyktorów) *X*. Odwracanie macierzy trójkątnej wykonaj za pomocą funkcji `backsolve`.
5. 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)`