

Statystyka II Mat. Egzamin komputerowy 21.06.06

1. Porównaj metody klasyfikacji *lda*, *qda*, *multinom* (dyskryminacja logistyczna dla wielu klas z `library(nnet)`) oraz *nnet* w eksperymencie krosvalidacji sześciokrotnej na danych `crabs[1:150,]`. Wykorzystaj wszystkie obserwacje. Cecha przewidywana przyjmuje 3 wartości BM, BF, OM. Jako miarę efektywności predykcji przyjmij wspólną informację.
2. Wczytaj z archiwum *MacierzePot.zip* 11 potencjałów kontaktowych (PK) oznaczonych numerami parzystymi. PK są to symetryczne macierze 20 x 20 opisujące oddziaływania 210 par aminokwasów w białkach.
3. Napisz ogólną funkcję `upperVec`, która przekształca macierz na wektor złożony z elementów należących do jej górnej części (ewentualnie dolnej) oraz przekątnej. Ze zbioru wektorów $\mathbf{x}_1, \dots, \mathbf{x}_{11}$ otrzymanych z jedenastu PK, zbuduj macierz danych $X = [\mathbf{x}_1, \dots, \mathbf{x}_{11}]$. Czy w zbiorze X -- 210 obserwacji 11-wymiarowych są obserwacje odstające?
4. Napisz odpowiednik, uogólnienie funkcji `lm` dla modelu $y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i$ w którym $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 W)$, gdzie W jest znaną macierzą dodatnio określoną.
5. Wykonaj klasteryzację *complete linkage* na danych z zadania 3. Dendrogram zapisz do pliku *.ps*. Narysuj również oddziaływania między aminokwasami na płaszczyźnie dwóch pierwszych składowych głównych i policz udział ich wariancji w całkowitej wariancji danych. Ile jest klastrow? Odpowiedź uzasadnij wykresem separowalności.
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)
```