

Statystyka II Mat. Egzamin komputerowy 2.09.08.

1. Wczytaj z archiwum `MacierzePot17_3kol.zip` potencjały kontaktowe (PK), czyli 12 symetrycznych macierzy 20×20 opisujących oddziaływania 210 par aminokwasów w białkach. PK są zapisane w plikach trójkolumnowych tak, że w pierwszej kolumnie znajduje się numer wiersza, w drugiej – numer kolumny, a w trzeciej – wartość elementu macierzy PK o tych współrzędnych. Napisz ogólną funkcję `three2Vec`, która przekształca plik trójkolumnowy na wektor złożony z elementów macierzy PK należących do jej dolnej części bez przekątnej. Ze zbioru wektorów $\mathbf{x}_1, \dots, \mathbf{x}_{12}$ otrzymanych z 12tu 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?
2. 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)`
3. Dla modelu regresji z poprzedniego zadania:
 - 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[,4]`, rozkład QR macierzy `X=cbind(1,X[, -4])` oraz funkcję `pf`.
4. Niech $p=(p_1, \dots, p_k)$ będzie wektorem prawdopodobieństw. Zaimplementuj funkcję rysującą wykres konturowy 2-wymiarowej gęstości $f(x,y)$ tak, że warstwie ograniczają obszary o prawdopodobieństwach p . Na przykład, jeśli $p = c(1,2,3,4)/5$, to pierwsza warstwa jest zadana przez zbiór $\{(x, y) : f(x, y) = z, P(f(x, y) \geq z) = 1/5\}$ dla pewnego z . Dla danych z zadania 2. policz gęstość (`Life.Exp Murder`) za pomocą `kde2d` i narysuj warstwie dla danego wyżej p .
5. 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`.