Statystyka II Mat. Egzamin komputerowy 12.09.06

1. Porównaj metody klasyfikacji `lda`, `qda`, `glm` (regresja logistyczna) w eksperymencie kroswalidacji sześciokrotnej na danych `X=na.omit(biopsy[,-1])` z biblioteki `MASS`. Cecha przewidywana `y` przyjmuje 2 wartości: „benign" i „malignant". Jako miarę efektywności predykcji przyjmij współczynnik Goodmana-Kruskala między `y` oraz `y_pred`.

2. Zbuduj model analizy wariancji dla danych `quine` z biblioteki `MASS`. W modelu tym zmienną objaśnianą jest `Days` - liczba opuszczonych dni nauki przez australijskie dzieci. Dobierz transformację zmiennej `Days` korzystając z funkcji `boxcox` i `logtrans` oraz wykonaj automatyczną selekcję zmiennych.

3. Wczytaj z archiwum *SMs2.zip* 26 macierzy podstawień (MP) opisujących podobieństwo biochemiczne aminokwasów. MP są macierzami o wymiarze 20 x 20.

4. Przekształć macierze MP na wektory złożone z 400 elementów, następnie ze zbioru wektorów $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{26}$ zbuduj macierz danych X=[$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{26}$]. Czy w zbiorze X są obserwacje odstające?

5. Wykonaj klasteryzację *k-średnich* danych X. Liczbę klastrów dobierz korzystając z wykresu separowalności. Narysuj dane X na płaszczyźnie dwóch pierwszych składowych głównych i przedstaw wynik klasteryzacji kolorując obserwacje.

6. 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.