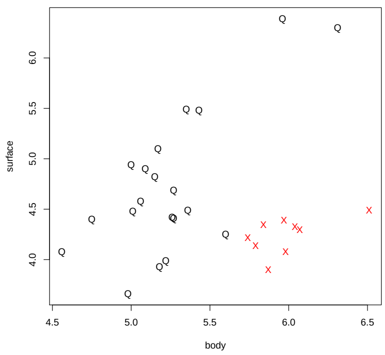
DATA MINING 2014 **Laboratorium 1 (LDA)**

1.1

Dane *earthquake.txt* dotyczą klasyfikacji wstrząsów na podstawie danych sejsmologicznych. Zmienna grupująca **popn** opisuje rodzaj wstrząsu: może to być trzęsienie ziemi (wartość *equake*) lub wybuch nuklearny (wartość *explosn*). Każdy wstrząs jest opisywany przez dwie zmienne objaśniające: **body** (magnituda fali głębokiej) i **surface** (magnituda fali powierzchniowej). Celem analizy jest identyfikacja rodzaju wstrząsu na podstawie zmiennych sejsmologicznych.

a) Wykonać wykres rozproszenia dla zmiennych **body** i **surface**. Obiekty z klasy *equake* oznaczyć literą "Q", a obiekty z klasy *explosn* literą "X".



b) Wyznaczyć:

- macierze kowariancji w klasach,
- macierz kowariancji wewnątrzgrupowej,
- pierwszy wektor kanoniczny (z definicji),
- równanie prostej rozdzielającej dwie klasy (prostopadłej do pierwszego wektora kanonicznego i przechodzącej przez środek odcinka między rzutami średnich w grupach). Otrzymaną prostą nanieść na wykres rozproszenia.

c) Wyznaczyć pierwszy wektor kanoniczny używając funkcji `lda(MASS)`.

d) Przedstawić graficznie obszary klasyfikacji obu klas na podstawie przynależności dla punktów kraty (np. 50×50).

e) Wyznaczyć pierwszą zmienną kanoniczną dla elementów próby oraz próg rozdzielający klasy. Stwierdzić która obserwacja została niepoprawnie zaklasyfikowana. Używając funkcji `predict.lda(MASS)` wyznaczyć tabelę reklasyfikacji.

f) Używając wyznaczonej prostej dyskryminacyjnej stwierdzić jakiego typu będzie wstrząs dla którego wartości zmiennych **body** i **surface** są równe odpowiednio: 6 i 4?

1.2

(**Praca domowa**) Rozważmy sytuację większej (niż 2) liczby klas $g > 2$. Uogólnienie zadania Fishera dla $g = 2$ na przypadek większej liczby klas:

- Znajdź kierunek $a \in R^p$ maksymalizujący wyrażenie:

$$\frac{a'Ba}{a'Wa}, \tag{1}$$

  gdzie

$$B = \frac{1}{g-1}\sum_{k=1}^{g} n_k(\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})',$$

$$W = \frac{1}{n-g}\sum_{k=1}^{g}(n_k - 1)S_k.$$

- Obserwację x przypisz do klasy j jeżeli

$$|a'x - a'\bar{x}_j| < |a'x - a'\bar{x}_k|, \quad k \neq j.$$

Pokaż że dla $g = 2$ problem maksymalizacji wyrażenia (1) jest równoważny maksymalizacji wyrażenia

$$\frac{(a'\bar{x}_2 - a'\bar{x}_1)^2}{a'Wa}.$$

Wskazówki:

1. Pokaż że $(a'\bar{x}_2 - a'\bar{x}_1)^2 = a'(\bar{x}_2 - \bar{x}_1)(\bar{x}_2 - \bar{x}_1)'a$.
2. Udowodnij:

$$a'(\bar{x}_2 - \bar{x}_1)(\bar{x}_2 - \bar{x}_1)'a = \frac{n_1 + n_2}{n_1 n_2} a' \left[\sum_{k=1}^{2} n_k(\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})'\right] a.$$

1.3

Zbiór danych *wine.data* zawiera dane dotyczące wyników chemicznej analizy win pochodzących z tego samego regionu Włoch, ale od trzech różnych plantatorów (Barolo, Grignolino, Barbera). W zbiorze mamy m.in zmienne:

- **V1**- zmienna grupująca, przyjmuje wartości 1, 2, 3 (numer plantatora),
- **V2**- zawartość alkoholu,
- **V8**- zawartość flawanoidów,
- **V14**- protolina (aminokwas), i inne (dokładny opis na `http://archive.ics.uci.edu/ml/datasets/Wine`).

a) Sporządzić wykres rozproszenia dla zmiennych **V2** i **V8** z zaznaczeniem numerów klas.

b) Utworzyć funkcję dyskryminacyjną za pomocą funkcji `lda(MASS)` przy wykorzystaniu tylko dwóch atrybutów: **V2** i **V8**. Wyznaczyć tabelę reklasyfikacji i obliczyć procent poprawnej klasyfikacji dla próby treningowej.

c) Przedstawić graficznie obszary klasyfikacji każdej z klas na podstawie przynależności dla punktów kraty (np. 50×50).

d) Utworzyć funkcję dyskryminacyjną za pomocą funkcji `lda(MASS)` przy wykorzystaniu trzech atrybutów: **V2**, **V8**, **V14**. Wyznaczyć tabelę reklasyfikacji i obliczyć procent poprawnej klasyfikacji dla próby treningowej. Porównać z wynikiem otrzymanym w punkcie b).

e) Używając funkcji `cloud(lattice)` wykonać trójwymiarowy wykres rozproszenia dla zmiennych **V2**, **V8**, **V14** z zaznaczeniem klas.

1.4

a) Używając dwóch zmiennych objaśniających **V2** i **V8** ze zbiory *wine.data* zbudować klasyfikator dla zmiennej **V1** używając metody wielowymiarowej regresji liniowej.

b) Wyznaczyć tabelę reklasyfikacji i obliczyć procent poprawnej klasyfikacji dla próby treningowej.