Questão 1. O k-means clustering é um método que agrupa n observações em k grupos minimizando as distâncias euclideanas de cada observação à média de seu respectivo grupo. No caso de $k = 2$ podemos representar a função objetiva como

$$f(X|z, \mu_1, \mu_2) = \sum_{i=1}^{n} z_i(x_i - \mu_1)^2 + \sum_{i=1}^{n}(1 - z_i)(x_i - \mu_2)^2,$$

em que μ_j é o vetor de médias do grupo $j \in \{1, 2\}$, e z_i é o indicador de $X_i \in$ grupo 1. Obtenha o algoritmo de otimização com atualizações em blocos que primeiro atualiza Z e em seguida μ_1 e μ_2.

Dica: Note que $z_i \in \{0, 1\}$ é discreto, de modo que otimização dessa variável não pode ser feita utilizando critérios envolvendo derivadas. Em cada atualização, escolha o valor de z_i que minimiza f_n.

Questão 2. Considere o modelo de regressão não linear $Y_i = \beta_0 e^{\beta_1 x_i} + \epsilon$, em que $\epsilon \sim N(0, \sigma^2)$, e seu estimador de mínimos quadrados.

(a) Simule dados de acordo com esse modelo para $n = 200$, $x \sim U[2, 40]$, $\beta_0 = 60$ e $\beta_1 = -0.05$.

(b) Construa o algoritmo pelo método de Gauss-Newton para estimar os parâmetros desse modelo.

(c) Implemete o algoritmo de (b) e compare seu desempenho com o do método de Newton-Raphson apresentado em aula.

Questão 3. Sejam $X = (X_1, \cdot, X_n)$ uma amostra de vetores $X_i = (X_{i,1}, \ldots X_{i,m})^t$ no R^m provenientes de uma mistrura de distribuições normais multivariadas com vetores de médias $\mu_1 = (\mu_{1,1}, \ldots, \mu_{1,m})^t$ e $\mu_2 = (\mu_{2,1}, \ldots, \mu_{2,m})^t$ e matrizes de variâncias e covariâncias identidade. Ou seja, as marginais desta distribuições são independentes com variância 1.

(a) Simule um conjunto de dados de acordo com esta distribuição, para $m = 3$, $n = 40$, $\mu_1 = (0, 0, 0)^t$, $\mu_2 = (3, 3, 3)^t$.

(b) Utilize simulated annealing para estimar μ_1 e μ_2.

(c) Seja Z_i a indicadora de que o elemento i saiu da componente de mistura 1, com média μ_1. Considere os dados completos (X, Z) e construa um algoritmo EM para estimar μ_1 e μ_2. (Este modelo pode ser usado para agrupamento) Como você interpreta os $\pi_i = E(Z_i|X_i, \mu_1^{(n)}, \mu_2^{(n)})$?Utilize seu valor na última iteração para estimar quais valores vieram das componentes 1 e 2.

(d) Utilize a função optim do R para resolver a mesma questão.

(e) Compare os métodos de (b), (c) e (d) para esses dados.

- São robustos em relação ao valor original?
- Qual demora mais para convergir?
- Tem problemas de ótimos locis?

(f) Repita o item (e) para $\mu_2 = (1, 1, 0)^t$. O que mudou? Qual o melhor método?

(g) Os dados do arquivo em anexo representam dados genéticos de populações americanas e europeias. Utilize seu algoritmo EM para classificá-los em 2 grupos. Você pode comparar seu resultado com aquele da função kmeans do R.

Desafio: Realizar o exercício considerando um modelo mais geral para a estrutura de variância dos componentes da mistura.

Questão 4. Escolha um dos problemas trabalhados em aula (Poisson sports Ranking ou Sistema ABO) para otimização utilizando o método de Newton, e a função optim do R. Compare os resultados com os obtidos em aula. Dica: não esqueça de incluir restrições apropriadas para seus parâmetros.

Questão 5. A função optim() do softaware R contem as implementações de diversos métodos de otimização.

(a) O método padrão é o de Nelder-Mead. Descreva como esse método opera, e indique em quais as suas vantagens e desvantagens.

(b) Escolha um dos outros métodos implementados na função optim() e repita o item (a).

(c) Quando e porquê é interssante utilizar o método de Newton-Raphson ao invés desses algoritmos?

Questão 6. Contribuição para pacote do R