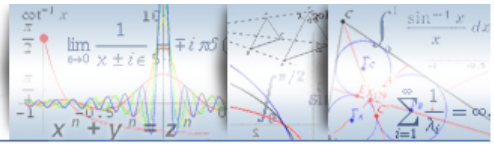
DEPARTAMENTO DE MATEMÁTICA
FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE COIMBRA



Clustering

Diogo Alexandre Mousinho dos Reis

Trabalho realizado no âmbito do **Programação Linear**
Disciplina da Licenciatura em Matemática

Orientador: Doutor João Soares

Conteúdo

1 Introdução

No presente trabalho será tratado e desenvolvido o tema Clustering.
Cluestering é um método bastante popular em análise de dados. O objetivo é particionar um conjunto de N objetos em subconjuntos denomidados de clusters. Assim sendo, objetos que pertençam ao mesmo cluster apresentam uma maior relação do que com objetos que pertençam a outro cluster. ([1], p.142,143)
Inicialmente, na secção 2, abordamos e descrevemos matemáticamente o problema.
De seguida, na secção 3, introduzimos o tópico da relaxação lagragiana que nos permite resolver probelmas desta natureza com restrições bastante complicadas, transportando-as assim para a função objetivo.
Na secção 4, é realizada a formalização de uma heurística apresentada por Cornuéjols, que resulta numa aplicação da relaxação lagragiana anteriormente descrita. São ainda descritos o motivo pelo qual o modelo introduzido na secção 2 apresenta inconsistência quando tentamos trabalhar com valores elevados de N, e duas proposições que nos levam a concluir que estamos perante um problema de otimização linear convexa.
Por fim, na secção 5, é apresentado um exemplo prático que implementa o modelo descrito na secção 2 para um pequeno valor de N com recurso a AMPL. Este apresenta tabelas e imagens que servem de suporte à interpretação e exemplificam de forma clara e sucinta o funcionamento do modelo desenvolvido.

Nota: Na pasta que contém o ficheiro pdf estão ainda os ficheiros relativos aos algoritmos utilizados na secção 5.

2 Clustering

Suponhamos que pretendemos particionar N objetos em $K < N$ clusters com p_{ij} a medida de semelhança entre os objetos i, j.

Este problema, relacionado com o famoso K-Median problem, pode ser resolvido matemáticamente quando transformado num problema binário.

Queremos dividir N objetos em K clusters $C_1, C_2, ..., C_k$. A ideia chave do problema é designar um elemento j_l em cada cluster C_l. Designemos j_l o centróide de cada cluster. Em cada centróide a medida de semelhança é dada por:

$$\sum_{i \in C_l} p_i, j_l$$

Portanto a medida de semelhança dos cluster's $C_1, C_2, ..., C_k$ é dada por:

$$\sum_{l=i}^{K} \sum_{i \in C_l} p_i, j_l$$

O centróide j_l representa os elementos que estão atribuídos ao cluster C_l. Assim sendo, cada cluster contém apenas os elementos associados ao seu centróide, e os clusters são completamente determinados pela sua escolha. ([1], p.142, 143)

2.1 Modelo de Programação Linear Binária para Clustering

Variáveis

Para $j = 1, ..., N$

$$y_j = \begin{cases} 1, & \text{se j é centróide} \\ 0, & \text{caso contrário} \end{cases} \tag{1}$$

Para $i, j = 1, ..., N$

$$x_{i_j} = \begin{cases} 1, & \text{se i é representado por j} \\ 0, & \text{caso contrário} \end{cases} \tag{2}$$

Função Objetivo

$$\sum_{j=i}^{N} \sum_{i=1}^{N} p_{i_j} x_{i_j}$$

.

Restrições

$$\sum_{j=i}^{N} y_j = K$$

Como $y_j = 1$ se for centróide e existem K clusters então escolhemos K centróides.

$$\sum_{j=i}^{N} x_{i_j} = 1 i, j = 1, ..., N$$

Cada objeto deve ser representado apenas por um centróide.

$$x_{i_j} \leq y_j i, j = 1, ..., N$$

Deve-se ao facto de que i é representado por j só se j é um centróide.

$$x_{i_j}, y_j \in \{0, 1\} i, j = 1, ..., N$$

Garante que as variáveis são binárias.

Assim ficamos com o problema:

$$\begin{aligned}
\max \quad & \sum_{j=1}^{N} \sum_{i=1}^{N} \rho_{ij} x_{ij} \\
\text{s.a} \quad & \sum_{j=1}^{N} y_j = K \\
& \sum_{j=1}^{N} x_{ij} = 1 \ , \qquad i, j = 1, \ldots, N \\
& x_{ij} \leq y_j \ , \qquad i, j = 1, \ldots, N \\
& x_{ij}, y_j \in \{0, 1\} \ , \quad i, j = 1, \ldots, N
\end{aligned} \tag{3}$$

([1], p.143)

3 Relaxação de Lagrange

A ideia presente na aplicação da relaxação lagragiana a problemas lineares é a seguinte: deslocar um conjunto de restrições "difíceis" para a função objetivo.

Consideremos o seguinte problema:

$$\begin{aligned} &\min\ c^T x \\ &\text{s.a. } Ax = b \\ &\qquad x \in \chi \end{aligned} \tag{4}$$

onde o conjunto de restrições $Ax = b,\ x \in \chi$ é "díficil"mas as restrições $x \in \chi$ são "fáceis".

Uma relaxação do problema descrito em (4) pode ser obtida da seguinte forma: assuma-se que temos um vetor u com uma dimensão adequada e consideremos o seguinte problema já sem as restrições "difíceis".

$$\begin{aligned} L(u) := \min\ & c^T x + u^T(b - Ax) \\ \text{s.a.}\quad & x \in \chi \end{aligned} \tag{5}$$

Este problema (5) é a relaxação de lagrange do problema dado. ([1], p.146)

Propriedade 1. *([1], p.147)*

Consideremos o seguinte problema de optimização

$$\begin{aligned} &\min\ c^T x \\ &s.a.\ Ax = b \\ &\qquad Dx \geq d \\ &\qquad x \in \chi \end{aligned} \tag{6}$$

e a sua relaxação lagragiana

$$\begin{aligned} L(u,v) := \min\ & c^T x + u^T(b - Ax) + v^T(d - Dx) \\ s.a.\quad & x \in \chi \end{aligned} \tag{7}$$

para vetores $u \geq 0$, $v \geq 0$. Então as seguintes proposições são satsifeitas:

(a) O valor ótimo $L(u,v)$ da relaxação (7) é menor ou igual ao valor ótimo de (6).

(b) Se a solução ótima x^* da relaxação (7) satisfaz $Ax^* = b$, $Dx^* \geq d$, e $v^T(Dx^* - d) = 0$ então $x*$ também é solução ótima de (6).

O dual lagragiano de (7) é:

$$\begin{aligned} &max\ L(u,v) \\ &s.a.\quad v \geq 0 \end{aligned} \tag{8}$$

4 Heurística baseada na Relaxação de Lagrange aplicada ao Clustering

Vamos agora descrever uma aplicação da relaxação de lagrange ao problema de clustering introduzido em 2.1. Esta heurística foi apresentada por Cornuéjols em 1977.

Tínhamos

$$\begin{aligned} \text{Z} := \max \quad & \sum_{j=1}^{N}\sum_{i=1}^{N} \rho_{ij} x_{ij} \\ \text{s.a} \quad & \sum_{j=1}^{N} y_j = K \\ & \sum_{j=1}^{N} x_{ij} = 1 \ , \qquad i,j = 1,\dots,N \\ & x_{ij} \leq y_j \ , \qquad i,j = 1,\dots,N \\ & x_{ij}, y_j \in \{0,1\} \ , \quad i,j = 1,\dots,N \end{aligned} \tag{9}$$

O modelo apresentado pode ser resolvido de diversas formas, tais como: excel, gurobi ou simplex mas apenas para valores pequenos de N. Uma das principais inconsistencias que o modelo apresenta é que envolve N^2+N variáveis binárias e N^2+N+1 restrições, e, como tal, por exemplo, para valores de $N = 100$ o problema torna-se de resolução praticamente impossível por qualquer um dos métodos mencionados.

Consideremos a seguinte relaxação do problema descrito em 4: dado um vetor de multiplicadores $u = [u_1 \ ... \ u_N]^T$ seja:

$$\begin{aligned} \text{L}(u) := \max & \sum_{i=i}^{N}\sum_{j=1}^{N} p_{i_j} x_{i_j} + \sum_{i=1}^{N} u_i (1 - \sum_{j=1}^{N} x_{i_j}) \\ \text{s.a.} \quad & \sum_{j=i}^{N} y_j = K \\ & x_{i_j} \leq y_j, \quad i,j = 1,\dots,N \\ & x_{i_j}, y_j \in \{0,1\}, \quad i,j = 1,\dots,N \end{aligned} \tag{10}$$

Este novo modelo, transportou as restrições consideradas "difíceis" $\sum_{j=i}^{N} x_{i_j} = 1, \ i,j = 1,...,N$ para a função objetivo através dos multiplicadores u e manteve as restrições designadas "fáceis". ([1], p.148)

Este modelo de relaxação de lagrange satisfaz as seguintes propriedades:

Propriedade 2. *([1], p.149)*

Para um dado u, o problema obtido após a relaxação torna-se de fácil resolução. Vejamos

$$\begin{aligned} L(u) &:= \max \sum_{i=i}^{N} \sum_{j=1}^{N} (p_{i_j} - u_i) x_{i_j} + \sum_{i=1}^{N} u_i \\ s.a. \quad & \sum_{j=i}^{N} y_j = K \\ & x_{i_j} \leq y_j, \quad i, j = 1, \ldots, N \\ & x_{i_j}, y_j \in \{0, 1\}, \quad i, j = 1, \ldots, N \end{aligned} \tag{11}$$

Os limites superior e inferior de x_{i_j} são y_j e 0, respectivamente. Dado y vemos que x_{i_j} toma os valores do seu limite superior ou infeiror dependo do sinal do coeficiente $p_{i_j} - u_i$ de x_{i_j}.

Assim sendo, o problema pode ser reescrito da seguinte forma:

$$\begin{aligned} L(u) &:= \max \sum_{j=i}^{N} C_j y_j + \sum_{i=1}^{N} u_i \\ s.a. \quad & \sum_{j=i}^{N} y_j = K \\ & y_j \in \{0, 1\}, \quad j = 1, \ldots, N \end{aligned} \tag{12}$$

para $C_j := \sum_{i=i}^{N} max\ (0, p_{i_j} - u_i)$.

Por fim, é possível observar que a solução para este último modelo 12 é computável: ordenemos os valores de C_j por ordem decrescente, ou seja, $C_{j_1} \geq C_{j_2} \geq \ ... \ \geq C_{j_N}$. A solução ótima do problema é obtida tomando $\overline{y}_{j_1} = \ ... \ = \overline{y}_{j_k} = 1$ e os restantes $\overline{y}_{y_{j_S}} = 0$. Obtemos $L(u) = \sum_{t=i}^{K} C_{j_t} + \sum_{i=i}^{N} u_i$.

Propriedade 3. *([1], p.149)*

Partindo da solução ótima $\overline{y}$ de L(u) do problema anterior é possível obter uma solução heurística $(\overline{x}, \overline{y})$ para o problema inicial e uma avaliação da sua qualidade.

A partir de cada u obtemos o limite superior $L(u) \geq Z$ e a heurística permite-nos obter uma solução viável $\overline{x}$ para o problema inicial. Assumamos que $\overline{y}$ é solução do problema anterior. Para cada $i = 1, ..., N$ atribuimos i ao centróide com maior semelhança aos K centróides tal que $\overline{y}_j = 1$. Isto é: seja $j(i) = \arg\max_{j:\overline{y}j=1} \rho_{ij}$ e seja $\overline{x}$:

$$\overline{x}_{ij} = \begin{cases} 1, & \textit{se } j = j(i) \\ 0, & \textit{caso contrário} \end{cases} \tag{13}$$

Observe-se que $\sum_{i,j} p_{ij}\overline{x}_{ij} \leq Z \leq L(u)$. *Portanto se* $\sum_{i,j} p_{ij}\overline{x}_{ij}$ *e* $L(u)$ *estão próximos um do outro então devem ambos estar próximos do valor ótimo* Z.

Para obtermos o "melhor" limite superior L(u) juntamente com a solução da heurística é então necessário resolver min $L(u)$. *Temos então um problema de otimização linear convexa.*

5 Exemplo de Aplicação

Suponhamos que pretendemos dividir 5 casas, definidas pelas suas coordenadas (x, y), de modo a formarmos dois bairros, ou seja, queremoss dividir as casas em 2 clusters procurando o valor que minimiza as distâncias de cada casa ao seu cluster.
O modelo do problema que pretendemos resolver, com recurso a AMPL, que devido ao pequeno valor de N é computável, é o seguinte:

$$\begin{aligned} \min & \sum_{j=i}^{5}\sum_{i=1}^{5} p_{i_j} x_{i_j} \\ \text{s.a.} \quad & \sum_{j=i}^{5} y_j = 2 \\ & \sum_{j=i}^{5} x_{i_j} = 1, \quad i,j = 1,\ldots,5 \\ & x_{i_j} \leq y_j, \quad i,j = 1,\ldots,5 \\ & x_{i_j}, y_j \in \{0,1\}, \quad i,j = 1,\ldots,5 \end{aligned} \tag{14}$$

Vejamos agora o seguinte exemplo:

$$\begin{bmatrix} 2.07 & 2.24 \\ 5.97 & 5.41 \\ 10.69 & 2.74 \\ 13.99 & 7.39 \\ 14.76 & 7.73 \end{bmatrix}$$

Tabela 1: Coordenadas X e Y dos pontos gerados

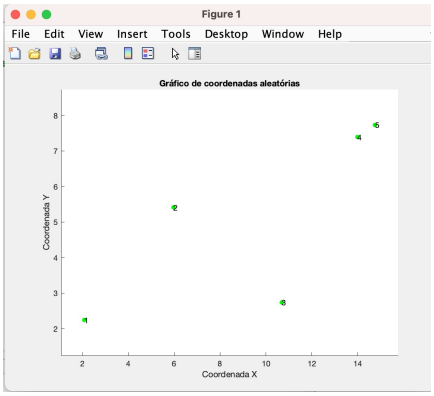


Figura 1: Gráfico das coordenadas dos pontos

$$\begin{bmatrix} 0 & 5.03 & 8.63 & 12.99 & 13.83 \\ 5.03 & 0 & 5.42 & 8.26 & 9.10 \\ 8.63 & 5.42 & 0 & 5.71 & 6.45 \\ 12.99 & 8.26 & 5.71 & 0 & 0.84 \\ 13.83 & 9.10 & 6.45 & 0.84 & 0 \end{bmatrix}$$

Tabela 2: Matriz de distâncias euclidianas

Uma vez que o centróide j_l representa os elementos que estão atribuídos ao cluster C_l conseguimos observar que, à partida, a divisão em clusters será a seguinte:

Clusters			
C_1	1	2	3
C_2	4	5	

Tabela 3

```
ampl: include linear.run
CPLEX 22.1.1.0: optimal integer solution; objective 11.29
8 MIP simplex iterations
0 branch-and-bound nodes
x [*,*]
:   1   2   3   4   5    :=
1   0   1   0   0   0
2   0   1   0   0   0
3   0   1   0   0   0
4   0   0   0   1   0
5   0   0   0   1   0
;

y [*] :=
1  0
2  1
3  0
4  1
5  0
;

obj = 11.29
```

Figura 2: Display dos resultados

Analisando os resultados, conseguimos observar que a divisão em clusters é respeitada, tendo sido fixados como centróides os pontos 2 *e* 4. É ainda possível confirmar se o valor obtido para a solução ótima, após 8 iterações do simplex (11.29), é consistente. Averiguemos: a soma das distâncias dos pontos 1 *e* 3 ao centro do cluster no ponto 2 é dada por $5.03 + 5.42 = 10.45$ e a distância do ponto 5 ao centro do cluster a que foi atribuido, neste caso, o ponto 4 é 0.84. Assim, somando as distâncias vemos o valor que minimiza a soma das distâncias é dado por $5.03 + 5.42 + 0.84 = 11.29$, que é coerente com o valor obtido utilizando o código AMPL.

Note-se que o ponto 2 é escolhido para centróide pois é o ponto que minimiza as distâncias a 1 e 3, por exemplo, se 1 fosse o centróide a soma das distâncias dos pontos 2 *e* 3 ao centróide seria dada por $5.03 + 8.63 = 13.66$, respectivamente, se 3 fosse o centróide a soma das distâncias dos pontos 1 *e* 2 ao centróide seria dada por $8.63 + 5.42 = 14.05$, ambas distâncias superiores a 10.45. No cluster C_2 seria indiferente a escolha para centróide uma vez que como são apenas dois pontos só existe uma distância. No entanto, no caso de existirem mais pontos a escolha para centróide é realizada de forma análoga à utilizada em C_1.

6 Conclusão

Com a realização deste trabalho, para além de ter aplicado as minhas capacidades de análise, investigação e tratamento de informação em tópicos relacionados com a disciplina de de programação linear, tive a oportunidade de desenvolver competências na linguagem AMPL. O tema do trabalho tem elevada importância na àrea de ciência de dados, na qual irei prosseguir estudos, o que contribuiu para um maior interesse e satisfação durante a realização do mesmo.

Referências

[1] Gérard Cornuéjols, Javier Peña, and Reha Tütüncü. Optimization Methods in Finance. Cambridge University Press, 2 edition, 2018.