Agrupamento espectral

Piero Kauffmann (8940810)

26 de Junho de 2019

1 Introdução e definições

Este capitulo introduz algumas definições sobre grafos valorados não-direcionados, que serão essenciais para a formulação do algoritmo de agrupamento espectral. As importantes matrizes de adjacência e de grau de um grafo são definidas abaixo

Definição 1.1 *Matriz de adjacências*

Seja $G = (V, E, W)$ um grafo valorado com vértices $V = \{v_1, ..., v_n\}$

Definimos a matriz de adjacências W_{nxn} como a matriz cujos elementos w_{ij} satisfazem

$$w_{ij} = \begin{cases} 0, & \text{se } i \neq j \text{ e } \{i,j\} \notin E \\ s_{ij}, & \text{se } i = j \text{ ou } \{i,j\} \in E \end{cases}$$

Onde $s_{ij} > 0$

Definição 1.2 *Grau do grafo*

Seja $G = (V, E, W)$ um grafo valorado com vértices $V = \{v_1, ..., v_n\}$

A matriz $D_{nxn} = diag(d_1, ..., d_n)$, é a matriz de grau do grafo G se

$$d_i = \sum_{j=1}^{n} w_{ij}, \text{ para } i = 1, ..., n$$

A matriz de adjacências (W) de um grafo carrega significados muito importantes. O principal deles é a informação sobre as arrestas do grafo, que podem ser inferidas pelos valores não-nulos da matriz. É possível codificar informações de similaridade ou dissimilaridade entre os vértices do grafo por meio dos pesos s_{ij}. Por razões de simplicidade, neste trabalho usaremos os pesos unitários ($s_{ij} \in \{0, 1\}$) para medir similaridade.

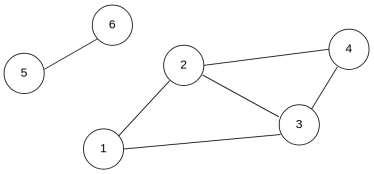


Figura 1: Exemplo de um grafo cujo Laplaciano L_{rw} possui dois autovalores nulos

Definição 1.3 *Laplaciano Random Walk*

Seja $G = (V, E)$ um grafo valorado com vértices $V = \{v_1, ..., v_n\}$ e matriz de adjacências ponderada W. Definimos

$$L_{rw} = I_n - D^{-1}W$$

Verificamos da definição do Laplaciano que $L_{rw} = I - P$, onde $P = D^{-1}W$ é a matriz de transições de um passeio aleatório no grafo.

O Laplaciano carrega ingormações e propriedades importantes sobre o grafo. Algumas delas são listadas abaixo:

- L_{rw} é simétrica e semi positiva definida
- O número de autovalores iguais a zero de L_{rw} é igual ao número de conjuntos de vertices disjuntos não conexos.
- Se u é um autovetor da matriz de transição P do passeio aleatório $\Leftrightarrow$ u é autovetor de L_{rw}

A segunda propriedade é importante pois está fortemente relacionada a presença de grupos aparentes no grafo. Um fato observável é que autovalores próximos a zero, indicam a presença de grupos quase não conexos também. A terceira propriedade será útil para compreender a construção do modelo em termos de passeios aleatórios no grafo.

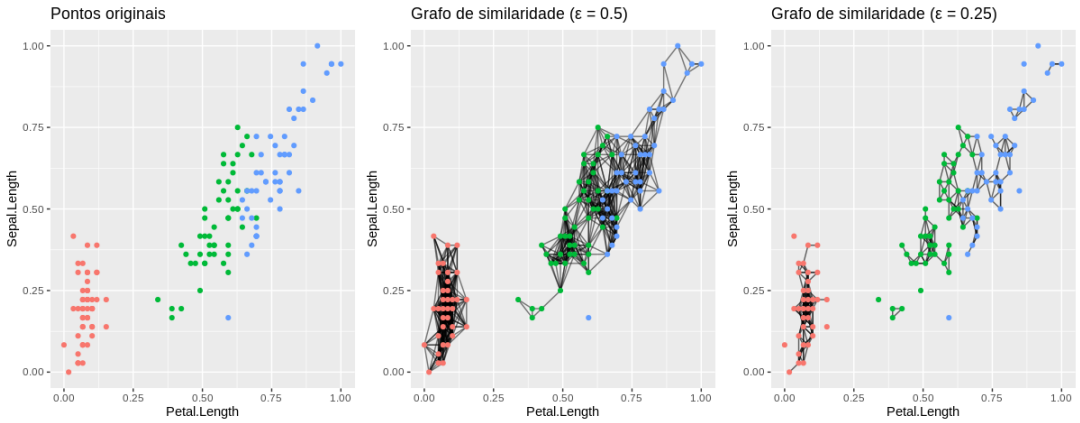


Figura 2: Método da ϵ-vizinhança aplicado ao conjunto de dados *iris* representados em duas variáveis, as cores representam as três espécies de plantas.

2 Construção do grafo de similaridade a partir de pontos no $\mathbb{R}^p$

Na literatura, existem três métodos mais utilizados para construir grafos de similaridade a partir de uma amostra $x_i, ..., x_n \in \mathbb{R}^p$.

- Método da ϵ-vizinhança
  - Todos os pares de pontos no conjunto de dados cujas distâncias sejam menores que ϵ são conectados
- Método dos k-vizinhos mais próximos
  - Cada ponto no conjunto de dados é conectado aos seus k vizinhos mais próximos
- Método da função de similaridade
  - Todos os pontos são conectados entre si, mas as similaridades entre cada par de pontos w_{ij} é obtida segundo uma função *Kernel* de similaridade $s : \mathbb{R}^p \times \mathbb{R}^p \to [0, 1]$
  - Uma escolha comum é $s(x_i, x_j) = exp(- \parallel x_i - x_j \parallel^2 /(2\sigma^2))$

A Figura 2 é um exemplo da aplicação do método da ϵ-vizinhança para a construção do grafo de similaridades. O parâmetro ϵ controla o nível de detalhamento das conexões formadas. Um valor baixo para ϵ (quadro da direita) pode resultar em um grafo com muitos sub grupos não conexos entre si, enquanto um valor muito elevado para o parâmetro pode resultar em um grafo que associa pontos muito distantes entre sí.

3 Intuição e justificativa do algoritmo

Sejam $A_1, ..., A_k \subset V$ conjuntos disjuntos de vértices do grafo G.

Definimos distribuição π^A como

$$\pi_i^A = \begin{cases} 1/|A|, & \text{se } i \in A \\ 0, & \text{caso contrário} \end{cases}$$

Procuramos π^A que minimize a probabilidade do passeio aleatório sair ou entrar no conjunto A , $C_A = P(A|A^C) + P(A^C|A)$

Pode ser provado que os k-primeiros autovetores de L_{rw} são uma solução aproximada para o problema, relaxando-se algumas suposições sobre π^A.

Uma interpretação possível está ligada com a distribuição estacionária π do passeio aleatório no grafo. É conhecido da teoria de matrizes estocásticas que os autovalores (com exceção do primeiro) estão relacionados com a velocidade de convergência a distribuição estacionária partindo-se de uma distribuição inicial. Fazendo uma ponte com a matriz L_{rw}, uma interpretação é que os autovetores são representações associadas a distribuições no grafo e seus autovalores associados indicam a velocidade de convergência desta distribuição à distribuição estacionária π. Autovalores mais baixos indicam tempos maiores para a distribuição da cadeia convergir.

O exemplo na Figura 3 demonstra o efeito da estrutura do grafo nos autovalores. O tempo de convergência da cadeia é maior quando partimos de uma distribuição π^A em um grafo muito particionado, pois o passeio aleatório demora para explorar a região A^C. Um exemplo extremo é quando o grafo possui dois subgrupos não conexos, o que indicaria que o tempo de convergência é infinito, isto é, $\lambda_2 = 0$.

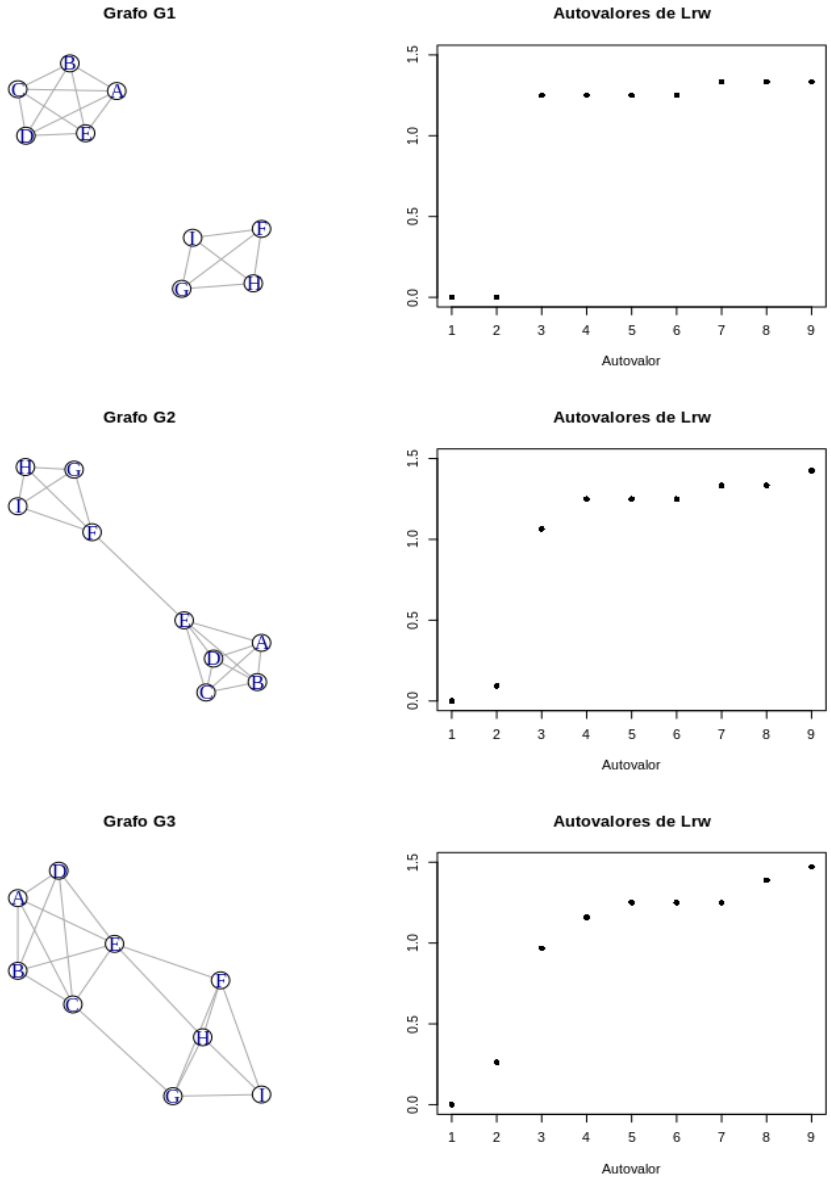


Figura 3: De cima para baixo, o comportamento do segundo autovalor varia de acordo com a conectividade entre os dois grupos de vértices {A, B, C, D, E} e {F, G, H, I}

.

4 Algoritmo

A. Criar um grafo de similaridades entre os pontos do conjunto de dados

   - Usando o métodos da ϵ-vizinhança, por exemplo

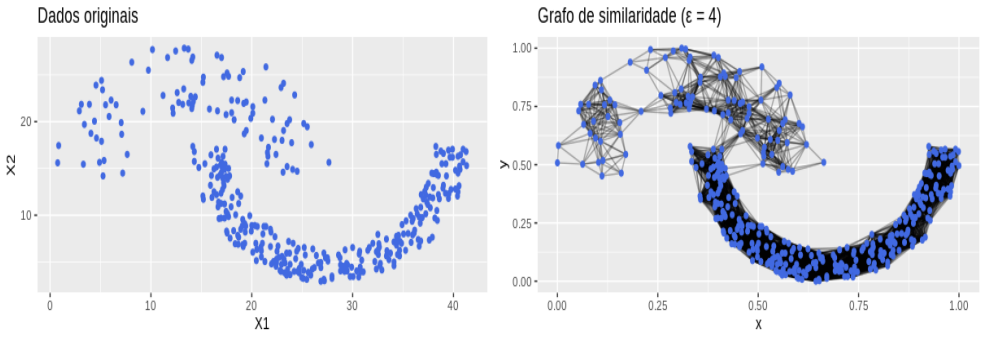
B. Calcular o Laplaciano $L_{rw} = I - D^{-1}W = I - P$

C. Extrair os k-primeiros autovetores de L_{rw}, excluindo-se o primeiro autovetor

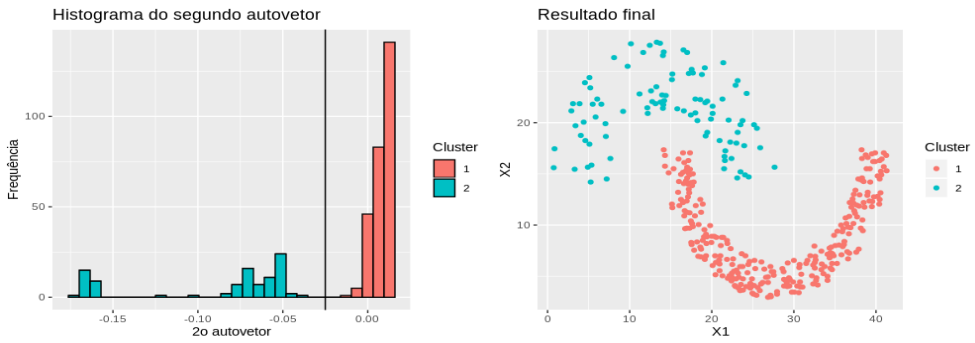
D. Aplicar o método das K-médias na matriz de dados dos k-autovetores extraidos

4.1 Exemplo de aplicação do algoritmo

Utilizando o conjunto de dados criado por Jain & Law (2006) o algoritmo foi aplicado. Os dados originais foram transformados em um grafo de similaridades usando o método da ϵ-vizinhança, com $\epsilon = 4$.



Após a criação do grafo de similaridade, foi extraido o segundo autovetor do grafo para aplicação do método das K-médias (K = 2)



Os resultados mostraram a separação dos dois grupos mais evidentes utilizando o segundo autovetor.

5 Referências

LUXBURG, U. (2007); "**A Tutorial on Spectral Clustering**".

LOVÀSZ, L. (1993); "**Random walks on graphs: A Survey**".

JAIN, A. & LAW, M. (2006); "**Data Clustering: A User's Dilemma**".