Agrupamento espectral e Isomap

Piero Conti Kauffmann

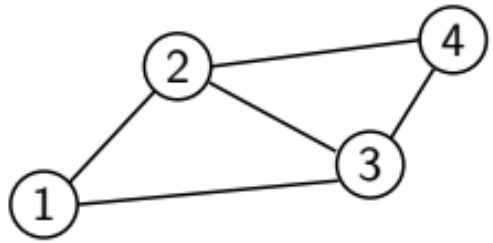
IME-USP

26 de Junho de 2019

Grafos valorados

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

- Definimos os elementos w_{ij} da matriz de adjâcencia ponderada do grafo W_{nxn} como
  - $w_{ij} = 0$ se não existir uma arresta $v_i - v_j$
  - $w_{ij} > 0$ caso contrário
- Definimos a matriz de grau do grafo G como $D_{nxn} = diag(d_1, ..., d_n)$. Onde $d_i = \sum_{j=1}^{n} w_{ij}$



Para o grafo da figura, assumindo $w_{ij} \in \{0, 1\}$ e $w_{ii} = 0$

$$W = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \text{ e } D = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

Grafos de similaridades e dissimilaridades

O peso w_{ij} é utilizado para representar algum tipo de relação entre os vértices v_i e v_j

- Grafos de similaridade
  - w_{ij} representa a similaridade entre os dois vértices
- Grafos de dissimilaridades (ou grafos de distância)
  - w_{ij} representa a dissimilaridade (ou distância) entre os dois vértices

Os dois modelos que iremos ver utilizam desses dois conceitos para agrupar pontos no espaço

Construção de um grafo de similaridades a partir de pontos no $\mathbb{R}^p$

Método da ϵ-vizinhança

- Todos os pares de pontos no conjunto de dados cujas distâncias sejam menores que ϵ são conectados

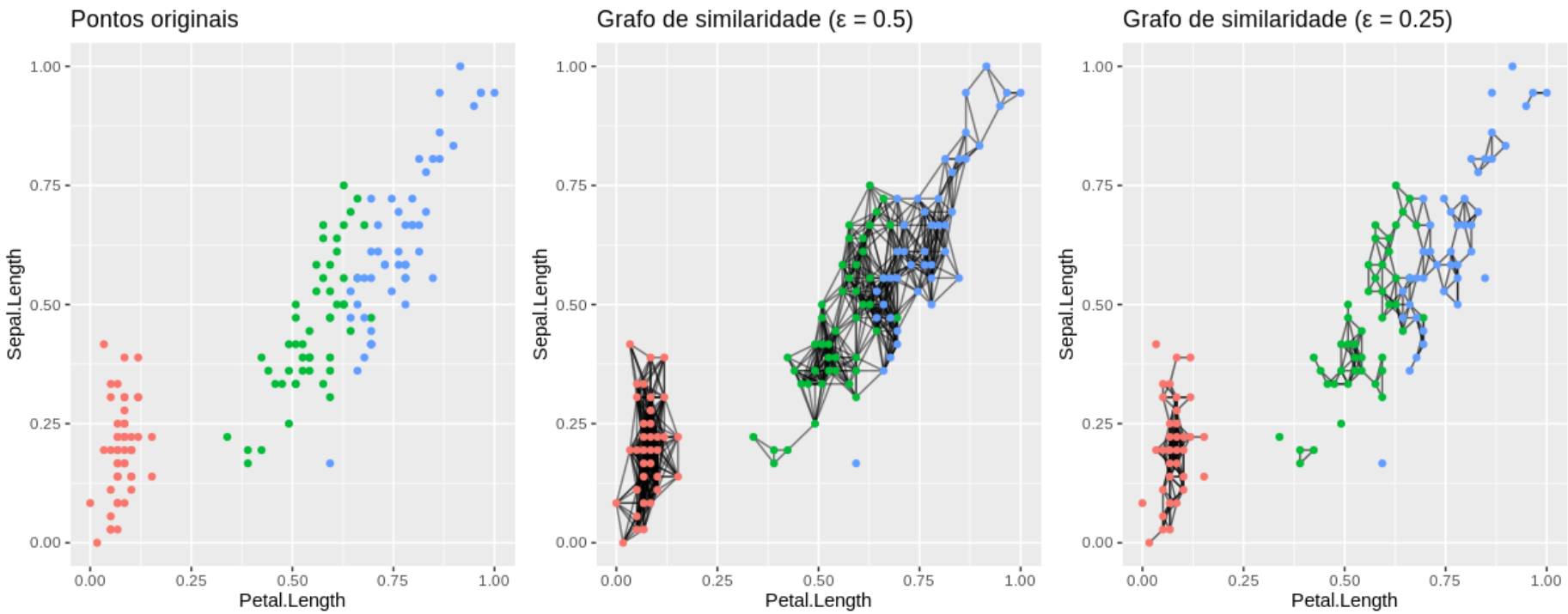
Método dos k-vizinhos mais próximos

- 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 $s : \mathbb{R}^p \times \mathbb{R}^p \to [0, 1]$
  - Exemplo: $s(x_i, x_j) = exp(- \parallel x_i - x_j \parallel^2 /(2\sigma^2))$

Método da ϵ-vizinhança - *Iris*



Laplaciano de um grafo

O Laplaciano de um grafo (L) é uma matriz definida em termos de W e D que carrega propriedades importantes sobre o grafo.

As definições de Laplacianos mais utilizadas são:

- Não normalizado: $L = D - W$
- Normalizado: $L_n = D^{-1/2} L D^{-1/2}$
- *Random-walk*: $L_{rw} = I - D^{-1} W = I - P$
  - Nesta definição, assume-se que $w_{ii} = 1$, para $i = 1, ..., n$
  - A matriz $P = D^{-1} W$ pode ser vista como a matriz de transição de um passeio aleatório no grafo G, onde $p_{ij} = \frac{w_{ij}}{d_i}$

Propriedades principais do Laplaciano L_{rw}

- 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.

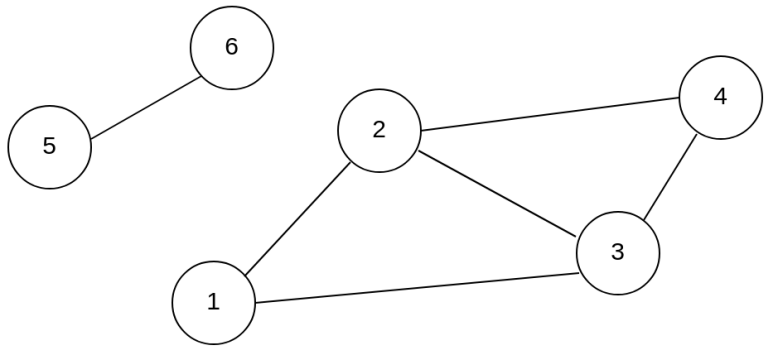


Figura: $\lambda_1 = \lambda_2 = 0$

- Se u é um autovetor da matriz de transição P do passeio aleatório $\Leftrightarrow$ u é autovetor de L_{rw}
  - Isso nos permite interpretar os autovetores de L_{rw} em termos de distribuições

Problema de agrupamento em termos de um passeio aleatório no grafo para

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

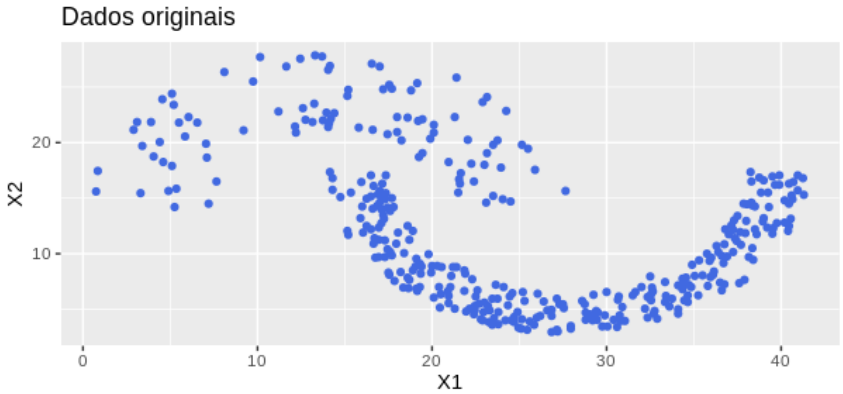
Algoritmo de agrupamento espectral

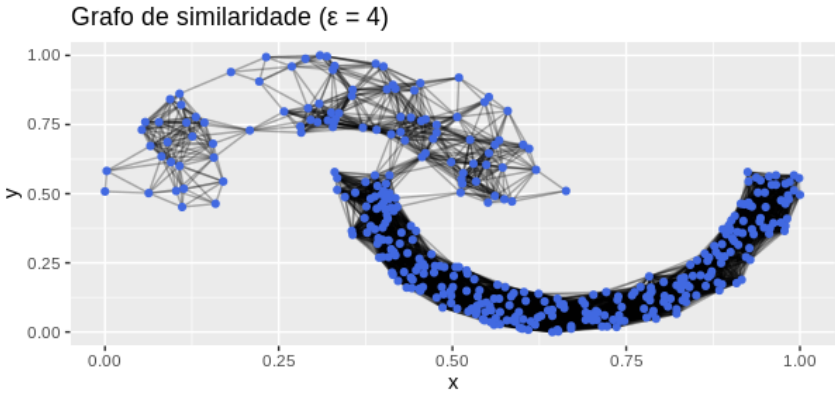
Algoritmo de agrupamento espectral para pontos no $\mathbb{R}^p$

1. Criar um grafo de similaridades entre os pontos do conjunto de dados
   - Usando o métodos da ϵ-vizinhança, por exemplo
2. Calcular o Laplaciano $L_{rw} = I - D^{-1}W = I - P$
3. Extrair os k-primeiros autovetores de L_{rw}, excluindo-se o primeiro autovetor
4. Aplicar o método das K-médias na matriz de dados dos k-autovetores extraidos

Exemplos

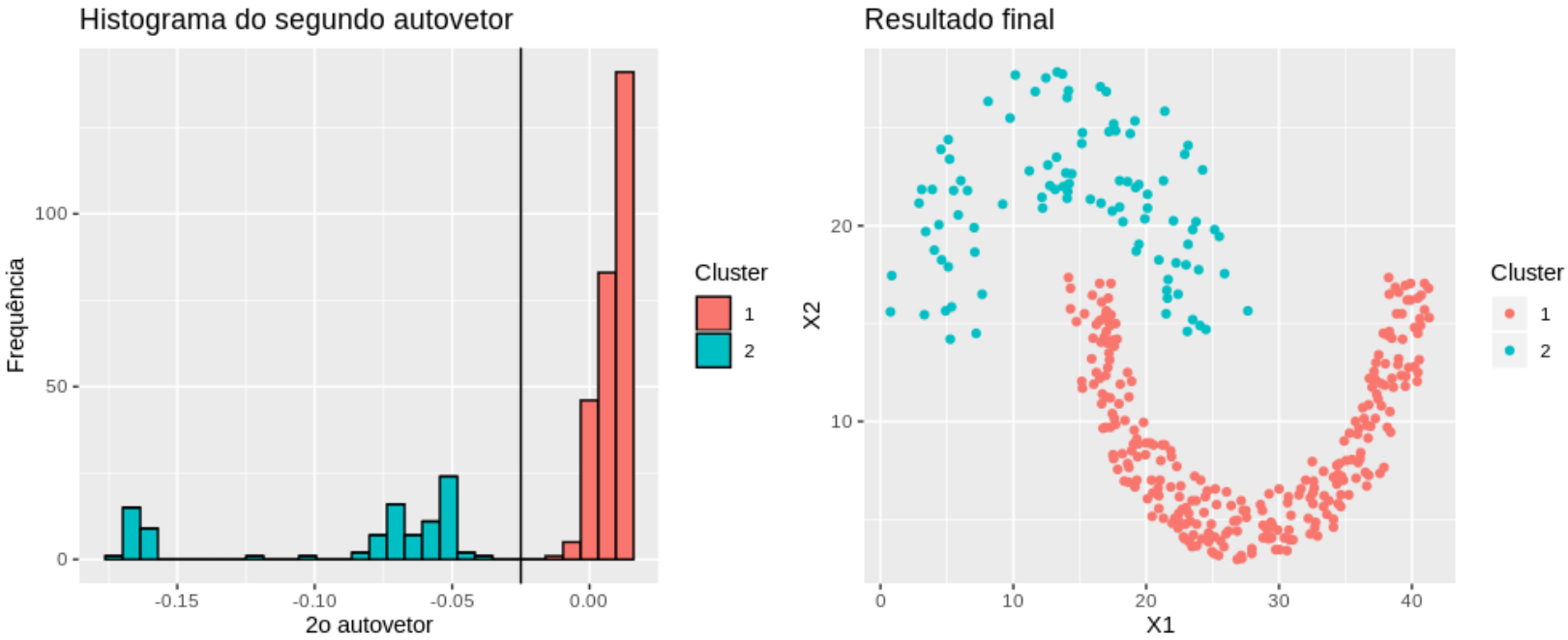
1. Do conjunto de dados de 373 pontos no $\mathbb{R}^2$, utilizando o método da ϵ-vizinhança, obtemos um grafo de similaridades





2. Obtemos os 2 primeiros autovetores de L_{rw}. Após descartar o primeiro autovetor, ficamos com o segundo autovetor

3. Aplicamos o método das K-médias (com $K = 2$) nos valores do segundo autovetor



ISOMAP

- Alguns métodos alternativos exploram o conceito de grafos de dissimilaridade (distâncias)

- O algoritmo ISOMAP, assim como o modelo de agrupamento espectral, cria um grafo a partir de pontos espalhados no R^p (usando algum dos três métodos que vimos acima).
- A diferença é que os pesos das arestas w_{ij} são definidos como a distância euclidiana entre os dois pontos x_i, x_j. Isto é,

$$w_{ij} = \| x_i - x_j \|^2$$

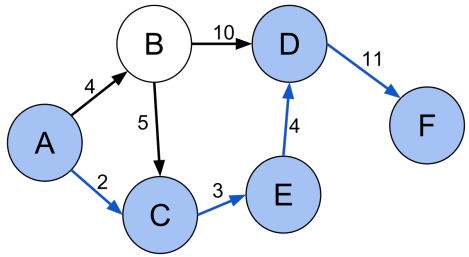
Distância geodésica em um grafo

Caminho possíveis entre dois vértices

Seja C_{ij} o conjunto de caminhos possíveis no grafo partindo-se do vértice v_i ao vértice v_j.

A distância geodésica entre dois vértices v_i e v_j é definida como a distância percorrida do menor caminho possível unindo v_i e v_j, isto é

$$d(i,j) = min_{K \in C_{ij}} \left(\sum_{(k_1 - k_2) \in K} w_{k_1 k_2} \right)$$

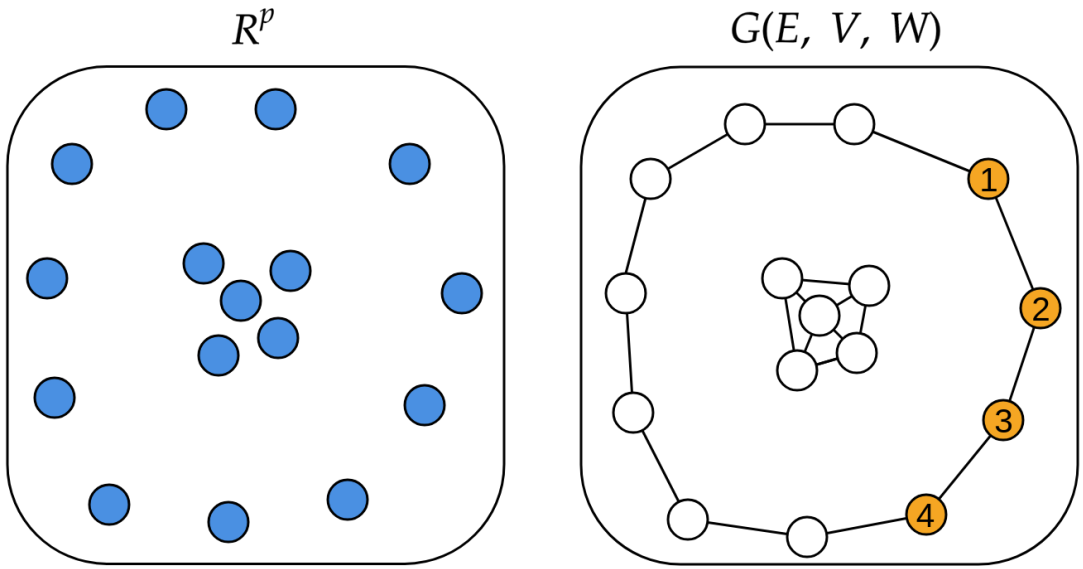


ISOMAP

Algoritmo

1. Criar um grafo de distâncias entre os pontos do conjunto de dados
   - Usando o métodos da ϵ-vizinhança, por exemplo
   - Os pesos w_{ij} dos vértices que foram conectados assumem o valor da distância euclidiana entre estes dois pontos, $d(x_i, x_j)$
2. Obter a matriz de distâncias geodésicas D_{nxn}
3. Extrair coordenadas principais de D_{nxn} usando Escalonamento Multidimensional
4. Aplicar o método das K-médias nos pontos projetados nas coordenadas principais

Exemplo

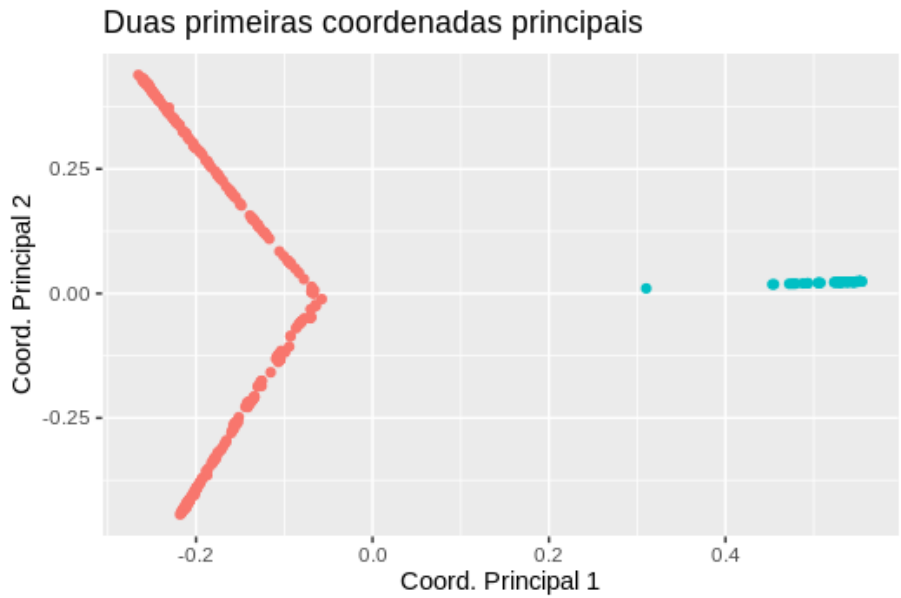


A distância geodésica entre os vértices 1 e 4 é

$$w_{12} + w_{23} + w_{34} = d(x_1, x_2) + d(x_2, x_3) + d(x_3, x_4)$$

Exemplo


Dados originais
Y
X
4
0
-4
-5.0
-2.5
0.0
2.5
5.0


Duas primeiras coordenadas principais
Coord. Principal 2
Coord. Principal 1
0.25
0.00
-0.25
-0.2
0.0
0.2
0.4

Referências

1. Luxburg, U. (2007), *A Tutorial on Spectral Clustering*
2. Lovàsz, L. (1993). *Random walks on graphs: A Survey*
3. Tenenbaum, J. (2000). *A Global Geometric Framework for Nonlinear Dimensionality Reduction*