Mineração de Dados

Trabalho 10

Arthur do Prado Labaki - 11821BCC017

27-05, 2023

GBC212

Estudo do Conceito

O coeficiente de silhueta é uma métrica utilizada na área de mineração de dados para avaliar a qualidade dos agrupamentos (clusters) obtidos por algoritmos de aprendizado não supervisionado, como o k-means, DBSCAN, entre outros.

O objetivo do coeficiente de silhueta é medir o quão bem cada objeto de dados se encaixa dentro do seu cluster atribuído em comparação com outros clusters. Ele varia de -1 a 1, onde valores próximos de 1 indicam que o objeto está bem ajustado ao seu cluster, valores próximos de 0 indicam que o objeto está próximo da fronteira entre dois clusters e valores próximos de -1 indicam que o objeto está mais próximo de um cluster diferente do que do cluster atribuído.

A fórmula do coeficiente de silhueta é calculada individualmente para cada objeto de dados e envolve o cálculo da distância média entre o objeto e todos os outros objetos dentro do mesmo cluster (distância intra-cluster) e a distância média entre o objeto e todos os objetos de outros clusters (distância inter-cluster). Em seguida, é calculada uma métrica combinando essas distâncias para fornecer um valor de silhueta para cada objeto.

Resolução do Exercício 1)

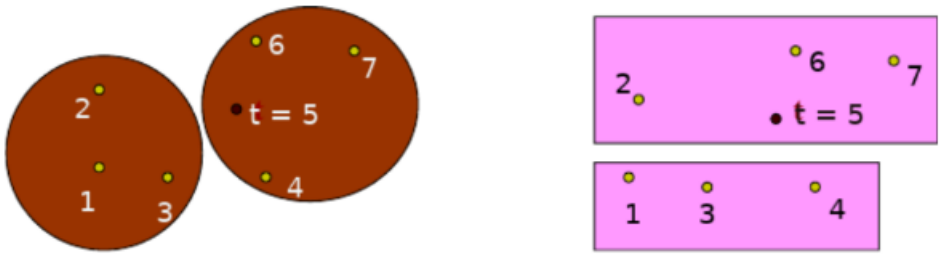
Considerando os dois agrupamentos listados abaixo, faça:

Calcule o coeficiente de silhueta do objeto t com relação a cada um dos agrupamentos e calcule a silhueta global de cada um dos agrupamentos e decida qual e melhor.

Para calcular o coeficiente de silhueta do objeto t, devemos calcular as distancias intra e inter cluster:

Conjunto 1:
Dm(T) intra = (D(T,4) + D(T,6) + D(T,7) / 3) = 0.943 + 1.236 + 1.923 / 3 = 1.367
Dm(T) inter = (D(T,1) + D(T,2) + D(T,3) / 3) = 2.118 + 2.039 + 1.280 / 3 = 1.812
Coeficiente de silhueta = (1.812 - 1.367) / max(1.812, 1.367) = 0.245

Conjunto 2:
Dm(T) intra = (D(T,2) + D(T,6) + D(T,7) / 3) = 2.039 + 1.236 + 1.923 / 3 = 1.732



	X	Y
1	1,5	1,4
2	1,5	2,5
3	2,5	1,3
4	4	1,3
t = 5	3,5	2,1
6	3,8	3,3
7	5,2	3

Figura 1: Base de dados

Dm(T) inter = (D(T,1) + D(T,4) + D(T,3) / 3) = 2.118 + 0.943 + 1.280 / 3 = 1.447
Coeficiente de silhueta = (1.447 - 1.732) / max(1.447, 1.732) = -0.164

Para calcular a silhueta global, temos que fazer os mesmos passos anteriores, so que com todos os pontos, então:

Conjunto 1:
Ponto 1 - Coeficiente de silhueta = 0.0716
Ponto 2 - Coeficiente de silhueta = -0.1432
Ponto 3 - Coeficiente de silhueta = 0.2891
Ponto 4 - Coeficiente de silhueta = 0.3837
Ponto 5 - Coeficiente de silhueta = 0.2453
Ponto 6 - Coeficiente de silhueta = 0.2974
Ponto 7 - Coeficiente de silhueta = 0.5207
Silhueta global = (0.0716 + (-0.1432) + 0.2891 + 0.3837 + 0.3749 + 0.2974 + 0.5207) / 7 = 0.2129

Conjunto 2:

Ponto 1 - Coeficiente de silhueta = 0.5531
Ponto 2 - Coeficiente de silhueta = 0.4729
Ponto 3 - Coeficiente de silhueta = 0.3317
Ponto 4 - Coeficiente de silhueta = 0.6696
Ponto 5 - Coeficiente de silhueta = -0.164
Ponto 6 - Coeficiente de silhueta = 0.1084
Ponto 7 - Coeficiente de silhueta = 0.5324
Silhueta global = (0.5531 + 0.4729 + 0.3317 + 0.6696 + 0.5486 + 0.3084 + 0.5324) / 7 = 0.1887

Logo o melhor agrupamento é o primeiro, pois tem o maior coeficiente de silhueta (mais próximo de 1).