K-MEANS E CLUSTERIZAÇÃO HIERÁRQUICA

Shirlei Dalila Machado Rezende

https://github.com/dalilamachado/clusterizacao

PERGUNTAS TEÓRICAS

Escreva em tópicos as etapas do algoritmo de K-médias até sua convergência.

1. Inicia-se k centróides em pontos aleatórios, pois o K-means não dá o valor de K.
2. Para cada ponto escolhido, deve-se encontrar o centróide mais próximo
3. Calcula-se o baricentro dos pontos para cada centróide
4. Mover o centróide na direção do baricentro
5. Repetir a partir de 2.

O algoritmo converge quando o movimento for menor que um valor pré-definido ou quando o número de iterações pré-especificado for atingido.

O algoritmo de K-médias converge até encontrar os centróides que melhor descrevem os clusters encontrados (até o deslocamento entre as interações dos centróides ser mínimo). Lembrando que o centróide é o baricentro do cluster em questão e não representa, em via de regra, um dado existente na base. Refaça o algoritmo apresentado na questão 1 a fim de garantir que o cluster seja representado pelo dado mais próximo ao seu baricentro em todas as iterações do algoritmo.
Obs: nesse novo algoritmo, o dado escolhido será chamado medóide.

O medoide é o ponto mais central localizado no cluster e precisa, necessariamente pertencer ao conjunto.

Passo a passo:
Selecionar k objetos para virar medoides.
calcular a matriz de dissimilaridade,
Atribuir cada objeto ao medoido mais proximo -
para cada cluster, observar se existe algum objeto que diminiu o coeficiente de dissimilidade e se existir, usar como novo medoide.

PERGUNTAS TEÓRICAS

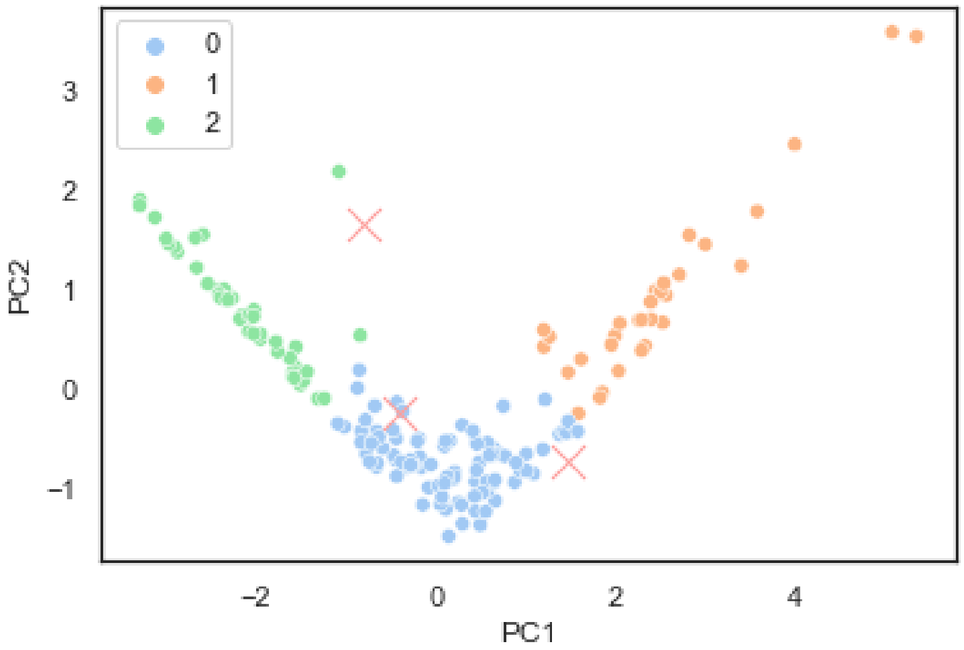
O algoritmo de K-médias é sensível a outliers nos dados. Explique.

A base do K-means é o calculo de médias. O outlier desloca o centroide para fora dos pontos "corretos", pois a média é calculada de forma enviezada, de acordo com com o outlier e nao de acordo com o cluster.
Portanto, tratar outlier é importante para que o K-means funcione corretamente.

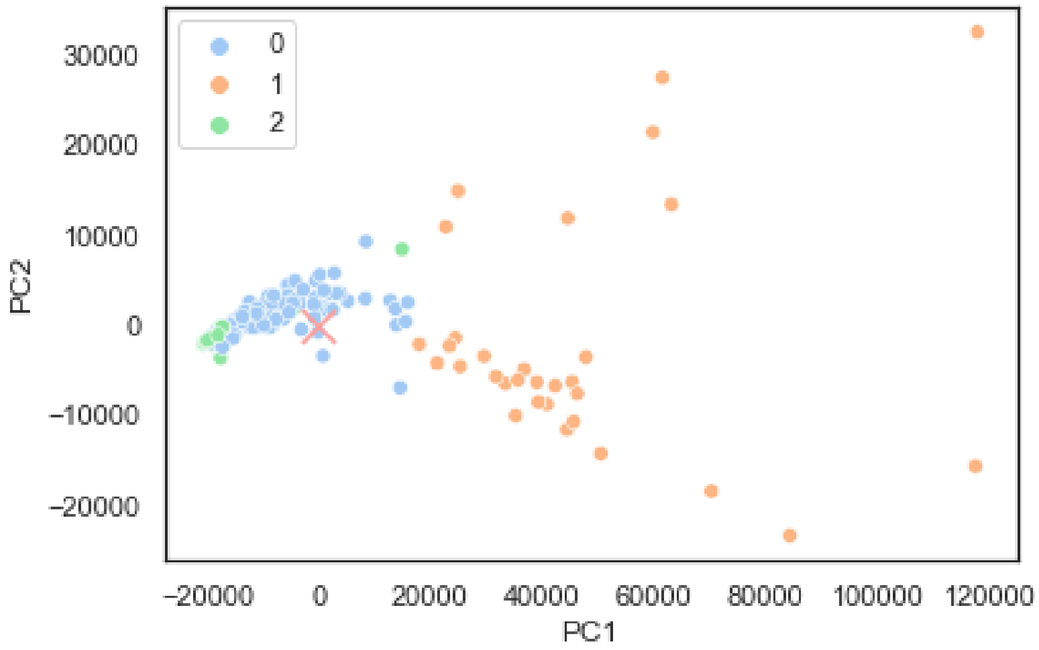
Por que o algoritmo de DBScan é mais robusto à presença de outliers?

Porque é necessário que a vizinhança de cada ponto do cluster tenha um número mínimo de pontos. Isso se mostra eficiente na detecção e tratamento de outliers.

K-MEANS



Modelo com dados normalizados



Média geral(.mean)

K-MEANS

Variável: Maior taxa de mortalidade infantil por país

Cluster 0: países em desenvolvimento

Cluster 1: países desenvolvidos

Cluster 2: países subdesenvolvidos

Cluster	País	%
0	Myanmar	3.2%
1	Saudi Arabia*	8.8%
2	Haiti	5.0%

País que melhor representa cada cluster

também é o país que melhor representa o conjunto de dados completo, com 3,3%

Neste contexto, a Arábia Saudita se enquadrou como país desenvolvido devido aos números elevados das outras variáveis.

CLUSTERIZAÇÃO HIERÁRQUICA

Variável: Maior taxa de mortalidade infantil por país

Cluster 0: países desenvolvidos

Cluster 1: países subdesenvolvidos

Cluster 2: países em desenvolvimento

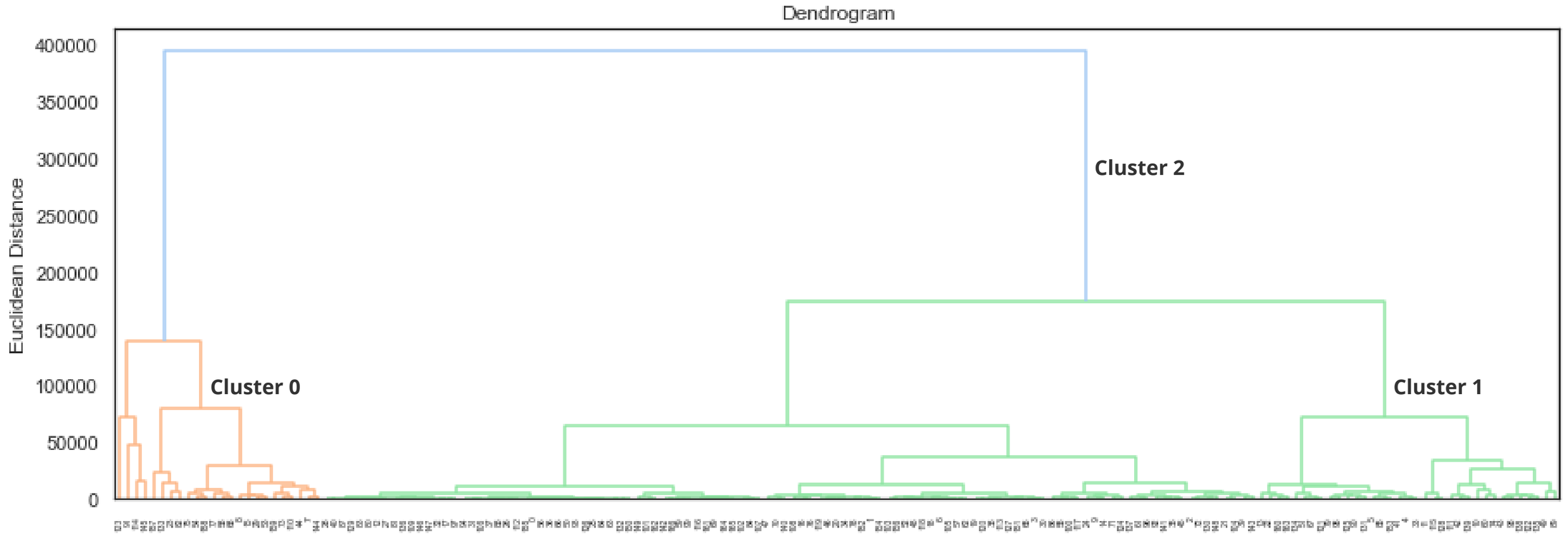
Cluster	País	%
0	Saudi Arabia*	7.3%
1	Haiti	4.9%
2	Myanmar	3.3%

País que melhor representa cada cluster

também é o país que melhor representa o conjunto de dados completo, com 3,3%

Neste contexto, a Arábia Saudita se enquadrou como país desenvolvido devido aos números elevados das outras variáveis.

Dendograma



A clusterização hierárquica mostra 3 clusters principais, mas com diferenças de taxas percentuais para os mesmos países e também diferença na quantidade de elementos em cada conjunto.

K-means

Cluster	Elementos
0 (em desenvolvimento)	**92**
1 (subdesenvolvidos)	43
2 (desenvolvidos)	32

H-cluster

Cluster	Elementos
2 (em desenvolvimento)	**85**
1 (subdesenvolvidos)	44
0 (desenvolvidos)	38