05

A estrutura dos clusters

Transcrição

Outra forma de validação é comparar os valores que obtemos do nosso dataset com outros de um banco de dados aleatório. A ideia é que possamos garantir que nosso conjunto de dados apresente melhores métricas, e por conseguinte, uma estrutura capaz de ser clusterizada.

Geraremos um conjunto de dados aleatório, com valores entre 0 e 1. Utilizaremos a mesma quantidade de variáveis, isto é, 16, e a mesma quantidade de instâncias.

O novo conjunto de dados que criaremos será chamado de `random_data`. Evocaremos a função que irá criar esse banco.

```
import numpy as np
random_data = np.random.rand(8950,16)
s, dbs, calinski = clustering_algoritithm(5, random_data)
print(s, dbs, calinski)
print(s2,dbs2, calinski2)
```

Temos uma diferença grande entre nosso banco de dados original e o conjunto aleatório, o que é bom para nossa validação.

Para o conjunto aleatório o valor de silhoutte foi de 0,03 e para o nosso 0,36. Para baouldin teremos 3,5 e 1.07. Já para o calinski teremos 300 e 3431.

Com base nisso, sabemos que estamos no caminho correndo e prosseguiremos com a análise dos nossos clusters.