06

Método para definir o número de clusters

Além do método do cotovelo e do método matemático, ainda existe mais uma forma de determinar o número ideal de cluster: o método da silhueta.

Você pode ler esse texto no Medium (https://medium.com/@jyotiyadav99111/selecting-optimal-number-of-clusters-in-kmeans-algorithm-silhouette-score-c0d9ebb11308) falando sobre o método. Como o texto está em inglês, segue a tradução do parágrafo que diz respeito ao assunto:

> **Coeficiente de silhueta**: Essa é uma medida melhor para decidir o número de clusters a serem formulados a partir dos dados. É calculado para cada instância e a fórmula é assim:
>
> $$S = \frac{(x - y)}{\max(x, y)}$$
>
> onde **y** é a distância média intra cluster: distância média para as outras instâncias no mesmo cluster.
>
> E **x** representa a distância média mais próxima do cluster, ou seja, a distância média às instâncias do próximo cluster mais próximo.
>
> O coeficiente varia entre -1 e 1. Um valor próximo a 1 implica que a instância está próxima ao cluster e faz parte do cluster correto. Por outro lado, um valor próximo de -1 significa que o valor está atribuído ao cluster errado.

Sendo assim, considerando o gráfico abaixo de coeficiente de silhueta, o número de cluster deveria ser 5, onde ocorre o maior valor.