AVALIAÇÃO DE QUALIDADE DOS AGRUPAMENTOS

Cristiane Neri Nobre

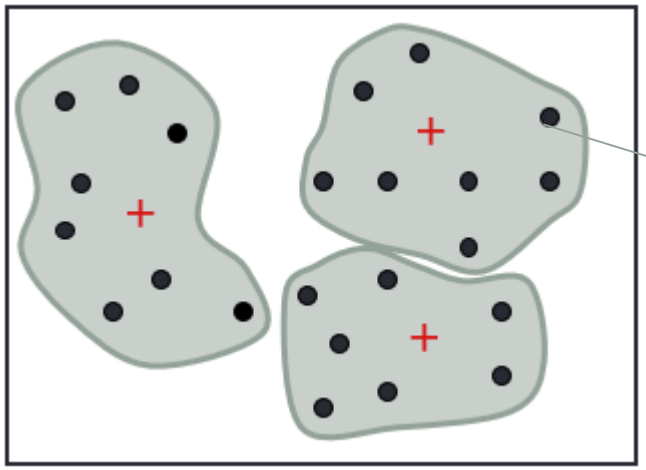
Métricas de avaliação

Silhouette Index

O Silhouette Index é uma medida de avaliação que avalia a coesão e a separação dos clusters, e baseia-se na diferença entre a distância média dos pontos pertencentes ao cluster mais próximo para os pontos de um grupo.

Métricas de avaliação - Silhouette Index

Para cada ponto da base de dados, x_i, é calculado o valor do silhouette index, S_i, de acordo com a Equação:



$$S_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)}$$

Onde:

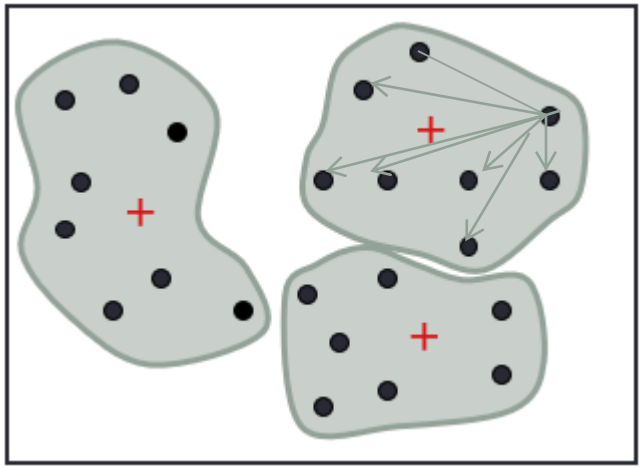
$\mu_{in}(x_i)$ é a distância média de x_i para os **pontos do seu próprio cluster**, e

$\mu_{out}^{min}(x_i)$ é a distância média de x_i para os **pontos dos clusters mais próximo**.

Métricas de avaliação

Onde:

- $\mu_{in}(x_i)$ é a distância média de x_i para os pontos do seu próprio cluster, e

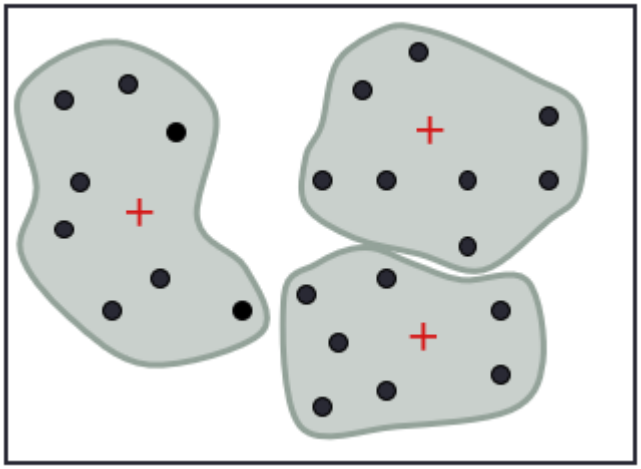


Métricas de avaliação

Onde:

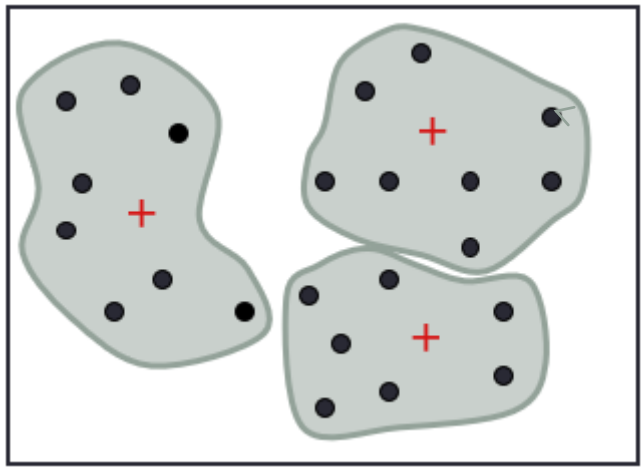
- $\mu_{out}^{min}(x_i)$ é a distância média de x_i para os pontos dos clusters mais próximo.

- Como saber qual o cluster mais próximo?



Métricas de avaliação

Assim, faz-se esta conta para cada instância da base de dados



$$S_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)}$$

Métricas de avaliação

Silhouette Index

- O valor S_i de um ponto encontra-se no intervalo $[-1, 1]$.
- Um valor próximo a 1 indica que x_i está mais próximo dos pontos do seu **próprio cluster** e distante dos clusters vizinhos.
- Um valor próximo de 0 indica que x_i está próximo da **fronteira de dois clusters**.
- Finalmente, um valor próximo a -1 indica que x_i está **próximo dos pontos que pertencem a outro cluster**.

Métricas de avaliação

Silhouette Index

- O valor S_i de um ponto encontra-se no intervalo $[-1, 1]$.

- Um valor próximo a 1 indica que x_i está mais próximo dos pontos do seu **próprio cluster** e distante dos clusters vizinhos.

- Um valor próximo de 0 indica que x_i está próximo da **fronteira de dois clusters**.

- Finalmente, um valor próximo a -1 indica que x_i está **próximo dos pontos que pertencem a outro cluster**

Métricas de avaliação

Silhouette Index

O sillhouette index é definido como o valor médio S_i entre todos os pontos, dado pela Equação

$$SilhouetteIndex = \frac{1}{n}\sum_{i=e}^{n} S_i$$

Métricas de avaliação

Silhouette Index

Segundo Rousseeuw (1987), para cada grupo é apresentado um valor do silhouette index, baseado na comparação da sua coesão (análise intra-cluster) e separação (com os demais grupos).

Este índice mostra quais objetos se encontram adequadamente agrupados no cluster e quais estão localizados indevidamente no cluster.

Todo o agrupamento é exibido pela combinação dos silhouettes em um único valor, permitindo uma validação da qualidade relativa dos clusters e uma visão geral da distribuição dos dados.

Métricas de avaliação

Silhouette Index

Tabela 1 – Análise do *Silhouette Index*

Valor	Significado
0.71 - 1.0	Uma estrutura forte foi encontrada
0.51 - 0.70	Uma estrutura razoável foi encontrada
0.26 - 0.50	A estrutura é fraca e pode ser artificial
< 0.25	Nenhuma estrutura substancial foi encontrada

Fonte: Adaptado de Rousseeuw (1987)

Métricas de avaliação

Investigue outras métricas!

Ex: Davies-Bouldin Index

Veja:
https://www.youtube.com/watch?v=438C3vGxYTE

Neste vídeo, aprendemos:

Sobre como avaliar a qualidade dos algoritmos de agrupamento?

No próximo vídeo, veremos **algumas considerações sobre o K-means e algoritmos de agrupamento**