



**Data Science
Academy**

www.datascienceacademy.com.br

Business Analytics

**Métricas de Clusterização
Definição e Interpretação**



Abaixo você encontra as principais métricas da análise de cluster e como deve ser a interpretação.

Métrica: Número de observações

O número de observações em cada cluster na partição final.

Interpretação:

Examine o número de observações em cada cluster ao interpretar as medidas de variabilidade, como a distância média e a soma dos quadrados dentro do cluster. A variabilidade de um cluster pode ser afetada por ter um número menor ou maior de observações. Por exemplo, a soma dos quadrados dentro do cluster se torna maior à medida que mais observações são adicionadas.

Examine os clusters que possuem significativamente menos observações que outros clusters. Clusters que têm poucas observações podem conter discrepâncias ou observações incomuns com características únicas.

Métrica: Soma de quadrados dentro do cluster (Within cluster sum of squares)

A soma dos desvios ao quadrado de cada observação e do centróide do cluster.

Interpretação:

A soma dos quadrados dentro do cluster é uma medida da variabilidade das observações dentro de cada cluster. Em geral, um cluster que possui uma pequena soma de quadrados é mais compacto do que um cluster que possui uma grande soma de quadrados. Clusters com valores mais altos exibem maior variabilidade das observações dentro do cluster.

No entanto, semelhante à soma dos quadrados e dos quadrados médios na ANOVA, a soma dos quadrados dentro do cluster é influenciada pelo número de observações. À medida que o número de observações aumenta, a soma dos quadrados se torna maior. Portanto, a soma dos quadrados dentro do cluster geralmente não é diretamente comparável entre os clusters com diferentes números de observações. Para comparar a variabilidade dentro do cluster de diferentes clusters, use a distância média do centróide.

**Métrica: Distância média do centróide**

A média das distâncias das observações ao centróide de cada cluster.

Interpretação:

A distância média das observações ao centróide do cluster é uma medida da variabilidade das observações dentro de cada cluster. Em geral, um cluster que possui uma distância média menor é mais compacto do que um cluster que tem uma distância média maior. Clusters com valores mais altos exibem maior variabilidade das observações dentro do cluster.

Métrica: Distância máxima do centróide

O máximo das distâncias das observações ao centróide de cada cluster.

Interpretação:

A distância máxima das observações ao centróide do cluster é uma medida da variabilidade das observações dentro de cada cluster. Um valor máximo mais alto, especialmente em relação à distância média, indica uma observação no cluster que fica mais distante do centróide do cluster.

Métrica: Centróide do cluster

O meio de um cluster. Um centróide é um vetor que contém um número para cada variável, e que cada número é a média de uma variável para as observações nesse cluster. O centróide pode ser considerado a média multidimensional do cluster.

Interpretação:

Use o centróide do cluster como uma medida geral da localização do cluster e para ajudar a interpretar cada cluster. Cada centróide pode ser visto como representando a "observação média" dentro de um cluster em todas as variáveis da análise.



Métrica: Distâncias entre centróides de cluster

As distâncias entre os centróides do cluster medem a distância entre os centróides dos clusters na partição final.

Interpretação:

Embora os valores da distância não sejam muito informativos, você pode comparar as distâncias para ver quão diferentes os clusters são. Uma distância maior geralmente indica uma diferença maior entre os clusters.

Referências:

Cluster Analysis

<https://www-users.cs.umn.edu/~kumar001/dmbook/ch8.pdf>