Existem vários métodos de avaliação de clusterização que ajudam a medir a qualidade dos agrupamentos produzidos por algoritmos de clusterização. A escolha do método de avaliação adequado depende do tipo de dados e do contexto do problema.

Matriz de Contingência

A Matriz de Contingência é uma métrica usada para avaliar a qualidade da clusterização quando você possui rótulos verdadeiros (verdadeiros agrupamentos) e os agrupamentos encontrados pelo algoritmo de clusterização. Ela é uma tabela que mostra a contagem de pontos que pertencem a cada combinação de cluster encontrado e rótulo verdadeiro.
A Matriz de Contingência é particularmente útil quando você deseja avaliar a correspondência entre os agrupamentos encontrados e os agrupamentos verdadeiros de forma detalhada. Ela fornece informações sobre quantos pontos foram corretamente atribuídos a seus clusters verdadeiros e quantos foram erroneamente atribuídos.
A Matriz de Contingência é uma tabela bidimensional (ou matriz) com linhas representando os rótulos verdadeiros e colunas representando os clusters encontrados. Os elementos da matriz são as contagens de pontos que pertencem a cada combinação de cluster e rótulo verdadeiro.

Por exemplo, a matriz de contingência pode ser assim:

	Cluster 1	Cluster 2	...	Cluster N
Classe A	20	2	...	4
Classe B	8	15	...	3
...	...	...	...	...
Classe M	0	6	...	8

- A linha "Classe A" representa os pontos que pertencem à classe "A" de acordo com os rótulos verdadeiros.
- A coluna "Cluster 1" representa os pontos atribuídos ao Cluster 1 pelo algoritmo de clusterização.
- O valor 20 na interseção entre "Classe A" e "Cluster 1" indica que 20 pontos da Classe A foram atribuídos ao Cluster 1.
- Da mesma forma, os outros valores na matriz representam as contagens de pontos para outras combinações de rótulos verdadeiros e clusters encontrados.

A Matriz de Contingência é uma ferramenta valiosa para entender como a clusterização se compara aos rótulos verdadeiros e é frequentemente usada em conjunto com outras métricas de avaliação, como a precisão e a pureza, para fornecer uma visão mais completa da qualidade da clusterização.

Silhueta

O Índice de Silhueta (Silhouette Score) [1] é uma métrica comum e amplamente utilizada para avaliar a qualidade da clusterização em conjuntos de dados. Ele fornece uma medida da coesão e separação dos clusters, ajudando a determinar o quão bem os pontos de dados estão agrupados. A ideia por trás do Índice de Silhueta é calcular o quão semelhante cada ponto de dados em um cluster é em relação aos pontos em seu próprio cluster em comparação com o cluster mais próximo vizinho (ou seja, o cluster a que ele não pertence). O índice varia de -1 a 1:

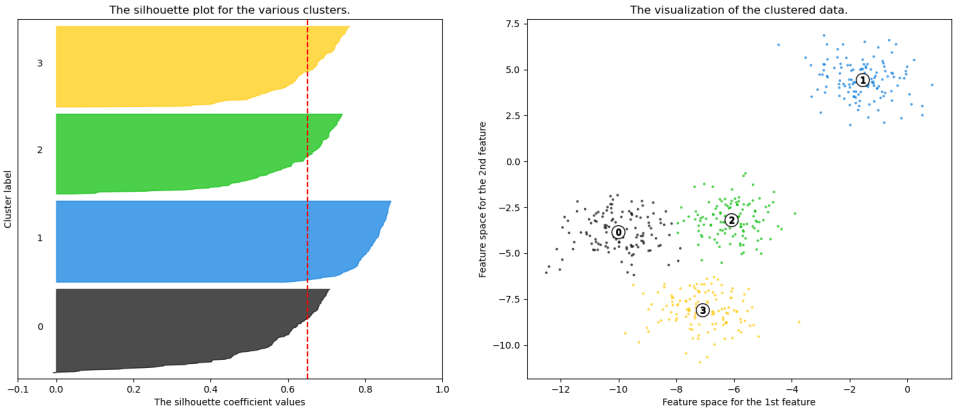
- Um valor próximo de +1 indica que o ponto está bem dentro do seu próprio cluster e longe dos outros clusters, o que é desejável.
- Um valor próximo de 0 indica que o ponto está próximo ou exatamente na fronteira entre dois clusters.
- Um valor próximo de -1 indica que o ponto está mais próximo de um cluster diferente do que do seu próprio cluster, o que indica que a clusterização pode ser inadequada.

A fórmula do Índice de Silhueta para um ponto *i* é a seguinte:

$$Silhouette(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

Onde:

- a(i) é a média das distâncias entre o ponto *i* e todos os outros pontos no mesmo cluster (coesão).
- b(i) é a menor média das distâncias entre o ponto *i* e todos os pontos em qualquer cluster diferente do seu próprio (separação).



Para calcular o Índice de Silhueta para um conjunto de dados, você pode seguir estas etapas:

1. Execute um algoritmo de clusterização (por exemplo, K-Means) no conjunto de dados para criar os clusters.

2. Para cada ponto de dados, calcule a(i) como a média das distâncias para os outros pontos no mesmo cluster e b(i) como a menor média das distâncias para os pontos em outros clusters.
3. Calcule o Índice de Silhueta para cada ponto de dados usando a fórmula acima.

O Índice de Silhueta é uma métrica útil para avaliar a qualidade dos clusters formados em um conjunto de dados, considerando tanto a coesão interna quanto a separação entre os clusters, geralmente, o valor médio da silhueta de todos os pontos é apresentado para resumir os resultados.

NMI (Normalized Mutual Information)

O Normalized Mutual Information (NMI) [2] é uma métrica de avaliação de clusterização que mede a dependência estatística entre os agrupamentos encontrados por um algoritmo de clusterização e os agrupamentos verdadeiros (rótulos conhecidos, quando disponíveis). O NMI é uma métrica de entropia normalizada que fornece uma medida da qualidade da clusterização em termos de informações mútuas entre os agrupamentos encontrados e os agrupamentos verdadeiros.

O NMI é calculado usando a entropia de Shannon e a entropia conjunta das classes e clusters. A fórmula para o cálculo do NMI é:

$$NMI(X,\ Y)\ =\ \frac{I(X;Y)}{\sqrt{H(X)H(Y)}}$$

Onde:

- *NMI(X,Y)* é o Normalized Mutual Information entre os rótulos verdadeiros (X) e os clusters encontrados (Y).
- *I(X;Y)* é a informação mútua entre X e Y, calculada como I(X;Y) = H(X) + H(Y) − H(X,Y).
- H(X) é a entropia de Shannon dos rótulos verdadeiros.
- H(Y) é a entropia de Shannon dos clusters encontrados.
- H(X,Y) é a entropia conjunta entre os rótulos verdadeiros e os clusters encontrados.

O valor do NMI varia de 0 a 1, onde valores mais altos indicam uma melhor correspondência entre os agrupamentos encontrados e os agrupamentos verdadeiros. Um NMI de 1 indica uma correspondência perfeita, enquanto valores mais baixos indicam menor correspondência.
O NMI é uma métrica útil para avaliar a qualidade da clusterização, especialmente quando você tem rótulos verdadeiros disponíveis para comparação. Ele leva em consideração tanto a coesão interna dos clusters quanto a separação entre os clusters, proporcionando uma medida abrangente da qualidade da clusterização.

Rand Index

O Índice de Rand, ou Rand Index (RI) [3], é uma medida de validação e avaliação de clusterização amplamente utilizada na área de aprendizado de máquina e análise de agrupamento. Ele é usado para avaliar a qualidade de como objetos, pontos de dados ou elementos foram agrupados ou classificados em clusters. O RI é especialmente útil quando você tem um conjunto de dados onde as categorias ou rótulos verdadeiros dos elementos são conhecidos, e você deseja avaliar quão bem seus clusters correspondem aos resultados esperados.

O funcionamento do Índice de Rand pode ser explicado da seguinte maneira:

Definição de Termos Básicos:

Verdadeiros Positivos (TP): São os pares de elementos que estão no mesmo cluster tanto no agrupamento verdadeiro quanto no agrupamento calculado.
Falsos Positivos (FP): São os pares de elementos que estão no mesmo cluster no agrupamento calculado, mas em clusters diferentes no agrupamento verdadeiro.
Verdadeiros Negativos (TN): São os pares de elementos que estão em clusters diferentes tanto no agrupamento verdadeiro quanto no agrupamento calculado.
Falsos Negativos (FN): São os pares de elementos que estão em clusters diferentes no agrupamento calculado, mas no mesmo cluster no agrupamento verdadeiro.

Cálculo do Índice de Rand: O RI é calculado usando a seguinte fórmula:

$$RI = \frac{TP + FP + FN + TN}{TP + TN}$$

Interpretação:

- Um RI próximo a 1 indica uma clusterização de alta qualidade, onde os agrupamentos correspondem muito bem aos agrupamentos verdadeiros.
- Um RI próximo a 0 indica que os agrupamentos são essencialmente aleatórios e não têm relação com os agrupamentos verdadeiros.

É importante observar que o Índice de Rand é sensível ao número de clusters e à forma dos clusters, o que significa que dois agrupamentos idênticos podem ter pontuações de RI diferentes se o número de clusters for diferente. Portanto, é geralmente aconselhável usar outras métricas juntamente com o RI para uma avaliação mais completa da qualidade da clusterização, como o Índice de Silhueta, a Entropia Normalizada ou a Pureza.

Alinhamento de clusters

O alinhamento de clusters, também conhecido como "cluster alignment," é uma técnica usada em análise de agrupamento (clusterização) que visa comparar ou alinhar os clusters gerados por um algoritmo de agrupamento com as classes verdadeiras e conhecidas dos pontos de dados. Com isso pode-se gerar comparações a fim de criar métricas para avaliação de clusters. O método de alinhamento mais simples faz com que cada cluster gerado se alinhe com o “cluster de referência” (classe) que tem maior sobreposição.
O primeiro passo do algoritmo é, para cada cluster, verificar qual a classe dominante nesse cluster. É necessário que se garanta que apenas um cluster esteja alinhado a apenas um único cluster de referência, caso cada cluster de referência esteja alinhado com mais de um cluster é necessário que haja competição para determinar qual o único cluster que irá permanecer alinhado, isso é necessário para que se evite a boa avaliação de um método de agrupamento que produz apenas um cluster ou que assinala um cluster para cada ponto.
Com o alinhamento feito pode-se calcular as métricas de avaliação mais conhecidas, Recall, F1-score, Accuracy, etc.

Exemplo de alinhamento de clusters:

Cluster ⇓	**R2**	**R1**	**R3**	**⇐ True Class**
C1	3	1	2	
C2	0	0	1	
C3	7	1	8	
C4	2	0	1	

Referência

[1] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score", 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, NSW, Australia, 2020, pp. 747-748, doi: 10.1109/DSAA49011.2020.00096.

[2] Aaron F. McDaid, Derek Greene, Neil Hurley, “Normalized Mutual Information to evaluate overlapping community finding algorithms”, 2011. https://arxiv.org/abs/1110.2515

[3] W. M. Rand (1971). "Objective criteria for the evaluation of clustering methods". Journal of the American Statistical Association. American Statistical Association. 66 (336): 846–850.