**PHƯƠNG PHÁP ĐÁNH GIÁ MÔ HÌNH**

1. **Silhouette Coefficient:**Silhouette Coefficient là một số liệu đo lường mức độ phù hợp của mỗi điểm dữ liệu với cụm được chỉ định của nó. Nó kết hợp thông tin về cả sự gắn kết (mức độ gần của một điểm dữ liệu với các điểm khác trong cụm riêng của nó) và sự phân tách (khoảng cách của một điểm dữ liệu với các điểm trong các cụm khác) của điểm dữ liệu.

Hệ số nằm trong khoảng từ -1 đến 1, trong đó giá trị gần 1 cho biết điểm dữ liệu được phân cụm tốt, giá trị gần 0 cho thấy các cụm chồng chéo và giá trị gần -1 cho biết điểm dữ liệu được phân loại sai.

Silhouette Coefficient cao hơn cho thấy các điểm dữ liệu được phân cụm tốt, với sự tách biệt rõ ràng giữa các cụm và sự gắn kết chặt chẽ trong mỗi cụm. Ngược lại, Silhouette Coefficient thấp hơn cho thấy rằng việc phân cụm có thể kém chính xác hơn, với các cụm hoặc điểm chồng chéo không được gán tốt cho các cụm tương ứng của chúng.

{REF: [Hệ số hình bóng được giải thích bằng một ví dụ thực tế: Đánh giá sự phù hợp của cụm" | bởi Suraj Yadav | Đau vừa (medium.com)](https://medium.com/@Suraj_Yadav/silhouette-coefficient-explained-with-a-practical-example-assessing-cluster-fit-c0bb3fdef719)}

Hệ số Silhouette được tính dựa trên khoảng cách trung bình trong cụm (a) và khoảng cách trung bình đến cụm gần nhất (b) cho mỗi mẫu. Hệ số Silhouette của một mẫu được xác định bằng công thức (b - a) / max(a, b). Để

giải thích rõ hơn, b là khoảng cách giữa một mẫu và cụm gần nhất mà mẫu đó không thuộc về. Lưu ý rằng Hệ số Silhouette chỉ có ý nghĩa khi số lượng nhãn nằm trong khoảng 2 <= n\_labels <= n\_samples - 1.



Ảnh: công thức Silhouette Coefficient

*Trong đó:*

a (i): Khoảng cách trung bình của điểm đó với tất cả các điểm khác trong cùng một cụm.

b (i): Khoảng cách trung bình của điểm đó với tất cả các điểm trong cụm gần nhất với cụm của nó.

s (i) - hệ số hình bóng hoặc điểm thứ i

{REF: [sklearn.metrics.silhouette\_score — scikit-learn 1.4.2 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html#:~:text=The%20Silhouette%20Coefficient%20is%20calculated%20using%20the%20mean,that%20the%20sample%20is%20not%20a%20part%20of.)

[Phương pháp Silhouette - Tốt hơn Phương pháp Elbow để tìm các Cụm tối ưu (ichi.pro)](https://ichi.pro/vi/phuong-phap-silhouette-tot-hon-phuong-phap-elbow-de-tim-cac-cum-toi-uu-61080390822033)}

1. **Calinski-Harabasz Index**

Chỉ số Calinski–Harabasz (CHI), còn được gọi là Tiêu chí Tỷ lệ Phương sai (VRC), là một thước đo để đánh giá các thuật toán phân cụm, được Tadeusz Caliński và Jerzy Harabasz giới thiệu vào năm 1974. Đây là một thước đo đánh giá nội bộ, trong đó việc đánh giá chất lượng phân cụm chỉ dựa trên tập dữ liệu và kết quả phân cụm, không dựa trên các nhãn chân lý gốc bên ngoài.

Cho một tập dữ liệu gồm n điểm: {x1, ..., xn}, và việc phân loại các điểm này vào k cụm: {C1, ..., Ck}, chỉ số Calinski–Harabasz (CH) được định nghĩa là tỷ số giữa sự phân tán giữa các cụm (BCSS) và sự phân tán trong cụm (WCSS), được chuẩn hóa theo số bậc tự do của chúng:



Từ công thức trên, chúng ta có thể kết luận rằng các giá trị lớn của chỉ số Calinski-Harabasz thể hiện sự phân cụm tốt hơn.

BCSS (Between-Cluster Sum of Squares) là tổng trọng số của các khoảng cách Euclid bình phương giữa mỗi tâm cụm (trung bình) và tâm dữ liệu tổng thể (trung bình):



trong đó ni là số điểm trong cụm Ci, ci là tâm của Ci, và c là tâm tổng thể của dữ liệu. BCSS đo lường mức độ phân tách của các cụm (càng cao càng tốt).

WCSS (Within-Cluster Sum of Squares) là tổng các khoảng cách Euclid bình phương giữa các điểm dữ liệu và tâm cụm tương ứng của chúng:



WCSS đo lường độ chặt chẽ hoặc sự liên kết của các cụm (càng nhỏ càng tốt). Việc tối thiểu hóa WCSS là mục tiêu của các thuật toán phân cụm dựa trên tâm, chẳng hạn như k-means.

{Ref: [Calinski–Harabasz index - Wikipedia](https://en.wikipedia.org/wiki/Calinski%E2%80%93Harabasz_index#:~:text=The%20Calinski%E2%80%93Harabasz%20index%20%28CHI%29%2C%20also%20known%20as%20the,Tadeusz%20Cali%C5%84ski%20and%20Jerzy%20Harabasz%20in%201974.%20)

[Calinski-Harabasz Index for K-Means Clustering Evaluation using Python (pyshark.com)](https://pyshark.com/calinski-harabasz-index-for-k-means-clustering-evaluation-using-python/)}