SAMPOERNA UNIVERSITY

GENERAL COURSE

COMP3315: Artificial Intelligence

Unsupervised Learning – Clustering Evaluation

By: Gilang Raka Rayuda Dewa

Why Evaluation Matters

SAMPOERNA UNIVERSITY

- Clustering is unsupervised — there are no labels available.
- Ground truth labels not always there.
- So, how do we know if clusters make sense?

  That is where we need clustering evaluation

Initial Dataset

SAMPOERNA UNIVERSITY

Internal Evaluation:

- Based only on the data and the clustering results
- Examples: Silhouette Score, Davies-Bouldin Index

External Evaluation:

- Requires ground truth labels (rare in real scenarios)
- Examples: Adjusted Rand Index, Mutual Information

Silhouette Score

Measures how similar an object is to its own cluster vs other clusters

Ranges from -1 to 1

+1: Well clustered

0: On the boundary

-1: Misclassified

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

- a(i) = average Euclidean distance to all other points in the same cluster
- b(i) = average Euclidean distance to all points in the nearest other cluster

Silhouette Score: Example

K-Means with two clusters
Cluster A: [1, 2], [1, 4], [1, 0]
Cluster B: [10, 2], [10, 4], [10, 0]

We want to check if [1,2] is suitable for cluster A

Step 1: a(i): [1,2] to cluster A

- ED([1,2], [1,4]) = 2
- ED([1,2], [1,0]) = 2
- a(i) = (2 + 2) / 2 = **2**

Step 2: b(i): [1,2] to cluster B

- ED([1,2], [10,2]) = 9
- ED([1,2], [10,4]) ≈ 9.22
- ED([1,2], [10,0]) ≈ 9.22
- b(i) = (9 + 9.22 + 9.22) / 3 ≈ 9.15

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

$$= \frac{9.15 - 2}{\max(9.15, 2)}$$

$$= 0.781$$

Davies-Bouldin Index

Measures average similarity between each cluster and its most similar one. The lower metric shows better clustering.

$$DB = \sum \max_{i \neq j}(R_{ij}) = \sum \max_{i \neq j}\left(\frac{S_i + S_j}{M_{ij}}\right)$$

S_i Average ED between each point in cluster i and its centroid (intra-cluster distance).

M_{ij}: ED between the centroids of clusters i and j (inter-cluster separation

Davies-Bouldin Index: Example

SAMPOERNA UNIVERSITY

K-Means with three clusters
Cluster A: **[1, 2]**, [1, 4]
Cluster B: **[10, 2]**, [10, 4]
Cluster C: **[5, 2]**, [5, 4]]

$$DB = \sum \max_{i \neq j}(R_{ij}) = \sum \max_{i \neq j}\left(\frac{S_i + S_j}{M_{ij}}\right)$$

For Cluster A:
- ED([1,2], [1,2]) = 0
- ED([1,4], [1,2]) = 2
- S_A = (0 + 2)/2 = **1**

For Cluster B:
- ED([10,2], [10,2]) = 0
- ED([10,4], [10,2]) = 2
- S_B = (0 + 2)/2 = **1**

For Cluster C:
- ED([5,2], [5,2]) = 0
- ED([5,4], [5,2]) = 2
- S_C = (0 + 2)/2 = **1**

$$M_{AB} = M_{BA} = \text{ED}([1,2], [10,2]) = 9$$
$$M_{AC} = M_{CA} = \text{ED}([1,2], [5,2]) = 4$$
$$M_{BC} = M_{CB} = \text{ED}([10,2], [5,2]) = 5$$

Davies-Bouldin Index: Example Cont'd (1)

SAMPOERNA UNIVERSITY

$$DB = \sum \max_{i \neq j}(R_{ij}) = \sum \max_{i \neq j}\left(\frac{S_i + S_j}{M_{ij}}\right)$$

Finding R_{ij}

- $R_{AB} = \frac{S_A + S_B}{M_{AB}} = \frac{1+1}{9} = 0.222$
- $R_{AC} = \frac{S_A + S_C}{M_{AC}} = \frac{1+1}{4} = 0.5$

$\max_{i \neq j} R_{ij} = 0.5$

- $R_{BA} = \frac{S_B + S_A}{M_{BA}} = \frac{1+1}{9} = 0.222$
- $R_{BC} = \frac{S_B + S_C}{M_{BC}} = \frac{1+1}{5} = 0.4$

$\max_{i \neq j} R_{ij} = 0.4$

- $R_{CA} = \frac{S_C + S_A}{M_{AC}} = \frac{1+1}{4} = 0.5$
- $R_{CB} = \frac{S_C + S_B}{M_{BC}} = \frac{1+1}{5} = 0.4$

$\max_{i \neq j} R_{ij} = 0.5$

Davies-Bouldin Index: Example Cont'd (2)

$$DB = \sum \max_{i \neq j}(R_{ij}) = \sum \max_{i \neq j}\left(\frac{S_i + S_j}{M_{ij}}\right)$$

$\max_{i \neq j} R_{ij}$ = 0.5

$\max_{i \neq j} R_{ij}$ = 0.4

$\max_{i \neq j} R_{ij}$ = 0.5

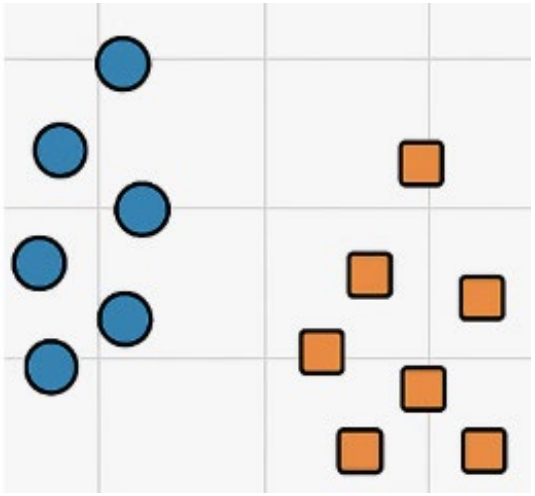
Find average, then DB = (0,5+0.4+0.5)/3 = 0.4667

Closer to 0: The clusters are well-separated and tight, meaning that intra-cluster distances are low, and inter-cluster distances are high.

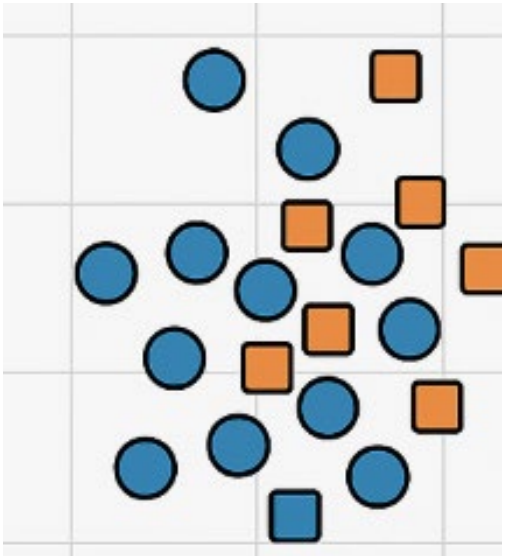
Higher DBI: The clusters are more overlapping or less well-separated, meaning that the average distance within clusters is too high or the distance between clusters is too small.

Visual Interpretations

SAMPOERNA UNIVERSITY



Silhouette Score = 0.82
DB Index = 0.32



Silhouette Score = 0.32
DB Index = 0.82

Example 1

The Sorting Hat wants to use clustering to assign Johan Iswara (6,4) to a house.

Gryffindor:

- Harry (9,3)
- Hermione (8,2)
- Neville (7,1)

Hufflepuff:

- Cedric (2,9)
- Hannah (3,8)
- Ernie (1,9)

Calculate the Silhouette Index. Based on the result, decide which house fits Johan Iswara best.

Example 2

Harry Potter is analyzing the behavior of two groups of students at Hogwarts. Each group's behavior has been tracked during a series of magical events:

Group 1 (Harry's Group): Harry (1, 2), Hermione (2, 3), Ron (3, 3)

Group 2 (Draco's Group): Draco (7, 8), Pansy (8, 7), Blaise (9, 8)

Harry suspects his friend, Ron, might belong to Draco's group.

Can you help Harry confirm this by calculating the Silhouette Index? Additionally, Professor McGonagall asked Harry to verify their BDI score.

The centroid of each group is calculated as the average of the points within that group.