

task_12.1

Machine Learning (WiSe 2025/2026)

Author: Suvansh Shukla

Matriculation No. 256245

Assignment 12 Task 1

Part A

We evaluate cluster in two major ways. Internal and External evaluation methods. Internal evaluation methods use the inherent properties of data to determine quality of a cluster, e.g. compactness and separation. Meanwhile, External cluster evaluation methods compare clustering results to known class labels or external benchmarks.

External cluster evaluation methods are used when priori information about the dataset is known, whereas Internal evaluation methods don't require such priori information. Most real world problems do not have priori information in the dataset, hence they use Internal evaluation methods.

Part B

Given:

A(1, 2), B(2, 1), C(3, 4), D(4, 1), E(4, 4) F(5, 3)

First we calculate the cluster centers for both clusters in Clustering A.

$$C_i = \frac{\sum_{i=1}^N (x_i, y_i)}{N}$$

Centroid for class 1 cluster: (1.50, 1.50)

Centroid for class 2 cluster: (4.00, 3.00)

Now we calculate the distance of each point to their cluster centers:

Cluster 1:

$$D(C_1, A) = \sqrt{(1 - 1.5)^2 + (2 - 1.5)^2} = 0.7071$$

$$D(C_1, B) = \sqrt{(2 - 1.5)^2 + (1 - 1.5)^2} = 0.7071$$

Cluster 2:

$$D(C_1, A) = \sqrt{(3 - 4)^2 + (4 - 3)^2} = 1.4142$$

$$D(C_1, B) = \sqrt{(4 - 4)^2 + (1 - 3)^2} = 2.0000$$

$$D(C_1, C) = \sqrt{(4 - 4)^2 + (4 - 3)^2} = 1.0000$$

$$D(C_1, D) = \sqrt{(5 - 4)^2 + (3 - 3)^2} = 1.0000$$

So $q = 0.7071 + 0.7071 + 1.4142 + 2.0000 + 1.0000 + 1.0000 = 6.8284$ for the Clustering A

Calculating cluster centers for Clustering B:

Centroid for cluster 1: (2.00, 2.33)

Centroid for cluster 2: (4.33, 2.67)

Now we calculate the distance of each point to their cluster centers:

Cluster 1:

$$D(C_1, A) = \sqrt{(1 - 2)^2 + (2 - 2.33)^2} = 1.0530$$

$$D(C_1, B) = \sqrt{(2 - 2)^2 + (1 - 2.33)^2} = 1.3300$$

$$D(C_1, C) = \sqrt{(3 - 2)^2 + (4 - 2.33)^2} = 1.9465$$

Cluster 2:

$$D(C_1, A) = \sqrt{(4 - 4.33)^2 + (1 - 2.67)^2} = 1.7023$$

$$D(C_1, B) = \sqrt{(4 - 4.33)^2 + (4 - 2.67)^2} = 1.3703$$

$$D(C_1, C) = \sqrt{(5 - 4.33)^2 + (3 - 2.67)^2} = 0.7469$$

So $q = 1.0530 + 1.3300 + 1.9465 + 1.7023 + 1.3703 + 0.7469 = 8.149$ for clustering B

Part C

Formula for Rand's index:

$$R = \frac{a + B}{\binom{n}{2}}$$

Total number of pairs is calculated as:

$$\frac{1}{2}n(n - 1)$$

For Clustering A we have 2 items in the first cluster and 4 items in the second cluster.

For Clustering B we have 3 items in the first cluster and 3 items in the second cluster.

We count the number of agreement-pairs in both clusterings: (A,B), (D,E), (D,F), (E,F)

Value of a = 4

We count the number of disagreement-pairs in both clusterings: (A,D), (A,E), (A,F), (B,D), (B,E), (B,F)

Value of b = 6

Total number of pairs = 15

Rand index = $10/15 = 0.6667$

Note how we do not count disagreement-pairs (C,D), (C,E), (C,F) because these pairs are only present in Clustering B

And agreement-pairs (A,C), (B,C) are only present in Clustering A.