Cluster Validation

Q1. Explain the following terms:

(a) Cluster validation

(b) Internal index

(c) External index

[Answer]

(a) Cluster validation refers to any procedures that evaluate the results of clustering in a *quantitative* and *objective* fashion.

(b) Internal index is a quantitative criterion designed based on "common sense" and "a priori knowledge" for clustering validity as the ground truth or the reference information is unavailable. Each time it only evaluates a partition.

(c) External index is a quantitative criterion designed to compare a partition to the ground truth or a reference to evaluate the clustering performance or similarity when the ground truth or a reference is available. It always gets two partitions involved during evaluation.

Q2. Briefly describe an application of internal index in clustering analysis. It is essential to describe the main steps in this application.

[Answer]

Internal index may be applied in finding the "proper" number of clusters in a data set although there is no guarantee. In general, one can run a clustering algorithm, e.g., *K*-means, on different conditions to lead to several candidate partitions of different clusters. Then an internal index, e.g., F-ratio, is used to evaluation each candidate partition. By detecting the "knee" point from the plot of cluster number vs. index, we can decide what the cluster number is optimal in terms of this criterion and so it will be viewed as the "proper" cluster number.

Q3. Briefly describe an application of external index in clustering analysis. It is essential to describe the main steps in this application.

[Answer]
External index may be applied in performance evaluation of a clustering algorithm with well-studied benchmark data sets where the ground truth is available. An external index, e.g., Rand index, needs to deal with problems such as cluster-ID permutation, point-pair corresponding and inconsistence in cluster numbers. Then the index value in [0, 1] would exhibit how similar a partition generated by the clustering algorithm to the ground truth; the higher, the closer, and the index value is one if and only if the partition is identical to the ground truth.

Q4: After running K-means (where *K*=2) twice on a data set of 5 objects on different initialization conditions, we achieve two partitions X and Y as follows:

	1	2	3	4	5
X	i	ii	ii	i	i
Y	q	p	q	p	q

You are asked to compare the above two partitions X to Y to see how similar they are with the Rand index. (a) Calculate the contingency table used in the Rand index and (b) calculate the Rand index based on the contingency table achieved in (a).

[Answer]
(a) As each partition has two clusters, the contingency table is 2 X 2 as follows:

X/Y	p	q
i	2	1
ii	1	1

(b) From the contingency table achieved in (a), we have

$N = 5,\ n_{1.} = 1 + 2 = 3,\ n_{2.} = 1 + 1 = 2,\ n_{.1} = 2 + 1 = 3,$ and $n_{.2} = 1 + 1 = 2.$

According to the formulae given in the lecture note, we can calculate

$$a = \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}\left(n_{ij}-1\right) = \frac{1}{2}[2(2-1)+1(1-1)+1(1-1)+1(1-1)] = 1$$

$$b = \frac{1}{2}\left[\sum_{j=1}^{2} n_{\cdot j}^{2} - \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}^{2}\right] = \frac{1}{2}[(3^2+2^2)-(2^2+1^2+1^2+1^2)] = 3$$

$$c = \frac{1}{2}\left[\sum_{j=1}^{2} n_{i\cdot}^{2} - \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}^{2}\right] = \frac{1}{2}[(3^2+2^2)-(2^2+1^2+1^2+1^2)] = 3$$

$$d = \frac{1}{2}\left[N^2 + \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}^{2} - \left(\sum_{j=1}^{2} n_{i\cdot}^{2} + \sum_{j=1}^{2} n_{\cdot j}^{2}\right)\right]$$

$$= \frac{1}{2}[5^2 + (2^2+1^2+1^2+1^2) - ((3^2+2^2)+(3^2+2^2))] = 3$$

Hence, we have $RI = \frac{a+d}{a+b+c+d} = \frac{1+3}{1+3+3+3} = 0.4$