Clustering Metrics

Dr. Kalidas Y., IIT Tirupati

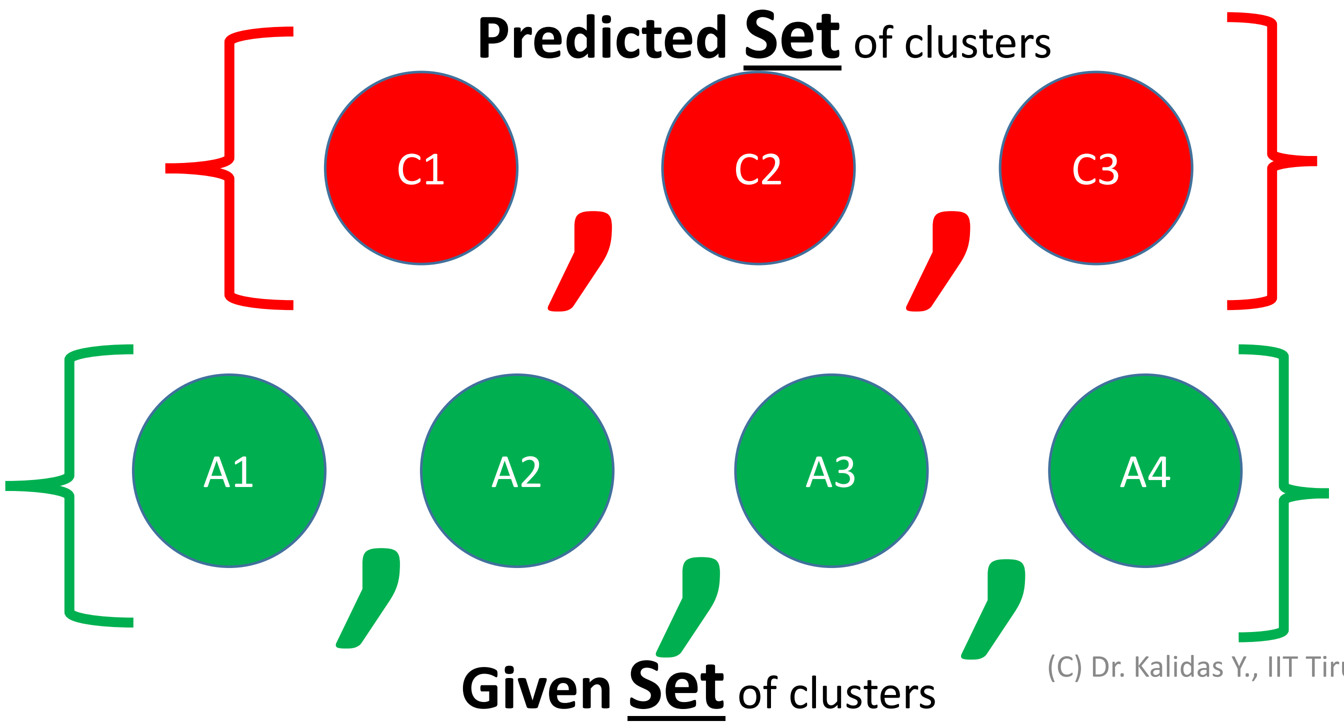
By the end of this lecture, you will be able to understand how clustering output needs to be evaluated

When are what type of clustering algorithms are useful?

Data shape Algorithm	Concentric Circles (Well separated)	Concentric Circles (*Thin wire connection*)	Blobs (Well separated)	Blobs (*Thin wire connection*)	Speed	Handles Complex shapes	Predicts for new points	Applications
K Means	No	No	Yes	Yes	Yes	No	Yes	Everywhere
Agglomerative	Yes	No	Yes	No	Yes	No		Life science – dendrograms and phylogenetic studies
DBSCAN	Yes	Yes	Yes	Yes	Yes	Yes	No	Used in computer vision domain

Quality of clustering (given ground truth)

- How much similar are the two clusterings?
- Assume all clusters are all mutually exclusive i.e. no common elements



Predicted Predicted cluster ids are different

Actual Actual cluster ids are different

Challenge… How do you compare *two sets of sets of clusters?*

- Predicted sets
  1. {1,2,3}
  2. {4,5}
  3. {6,7,8,9}
- Actual sets (Ground truth)
  1. {1,2}
  2. {3,4,5}
  3. {6}
  4. {7,8}
  5. {9}

There is *NO one-to-one CORRESPONDANCE* between *any of the predicted sets to any of the ground truth sets*

115) key phrase… “Adjusted Random Index (ARI)”

ARI – Higher the better indicative of the quality of clustering

- Predicted sets
  1. {1,2,3}
  2. {4,5}
  3. {6,7,8,9}
- Actual sets (Ground truth)
  1. {1,2}
  2. {3,4,5}
  3. {6}
  4. {7,8}
  5. {9}

- Consider *pairs of points*
- Consider Ground truth sets
- Let GSS = Pairs of points (xi,xj) occurring in a set
  - example.. (1,2), (4,5), (7,8) etc. [non-self]
  - example.. (1,1), (3,3) etc. [self]
  - Consider only unique pairs
    - For example, you can exclude (3,1) if (1,3) is already considered
    - Sort all points in row order and take (i,j) pairs
- Let GDS = Pairs of points (xi,xj) occurring in different sets
  - example… (1,3), (2,4) etc.
- Consider Predicted sets and compute, PSS and PDS respectively for same set and predicted set pairs of points
- Now, define, Random Index (RI) = $\frac{|GSS \cap PSS| + |\mathrm{GDS} \cap PDS|}{\mathrm{N} * \frac{(\mathrm{N}-1)}{2}}$
- Where N is the number of points
- Adjusted Random Index (ARI) is a metric, where $ARI = \frac{RI - Avg(RI)}{Max(RI) - Min(RI)}$

This is computed by generating random clusters and computing average, minimum and maximum random index scores among those

116) key phrase… "Mutual Information"

- Consider Predicted Sets – $\mathrm{C} = C_1, \ldots, C_k$ where each C_i is a set of points.
- Consider Ground Truth Sets – $\mathrm{G} = G_1, \ldots, G_m$ where each G_j is a set of points.
- Let N be the number of points
- $MI(C, G) = \sum_{i=1}^{i=k} \sum_{j=1}^{j=m} \frac{|C_i \cap G_j|}{N} \times \log\left(\frac{N \times |C_i \cap G_j|}{|C_i| \times |G_j|}\right)$

- The higher the similarity between the two sets, the higher this MI score is.
- When this score is lower then that clustering is not matching 'well' with ground truth

- If $\left|C_i \cap G_j\right|$ is less, while $|C_i|$ and $|G_j|$ are large, then MI score becomes less.

117) key phrase… "homogeneity score"
118) key phrase… "completeness score"

- Consider Predicted Sets – C $= C_1, \ldots, C_k$ where each C_i is a set of points.
- Consider Ground Truth Sets – G $= G_1, \ldots, G_m$ where each G_j is a set of points.
- Let N be the number of points
- Conditional 'Entropy of G given C': $H(G|C) = -\sum_{i=1}^{i=k} \sum_{j=1}^{j=m} \frac{|C_i \cap G_j|}{N} \times \log\left(\frac{|C_i \cap G_j|}{|C_i|}\right)$
- 'Entropy of G': $H(G) = -\sum_{j=1}^{j=m} \frac{|G_j|}{N} \times \log\left(\frac{|G_j|}{N}\right)$
- 'Entropy of C': $H(C) = -\sum_{i=1}^{j=k} \frac{|C_i|}{N} \times \log\left(\frac{|C_i|}{N}\right)$
- Homogeneity score: $h = 1 - \frac{H(G|C)}{H(G)}$ (higher better)
- Completeness score: $c = 1 - \frac{H(C|G)}{H(C)}$ (higher better)
- v measure: $v = 2 * \frac{h \times c}{(h+c)}$ (higher better)