

External Measures I: Matching-Based Measures

Matching-Based Measures (I): Purity vs. Maximum Matching

❑ **Purity:** Quantifies the extent that cluster C_i contains points only from one (ground truth) partition:

$$purity_i = \frac{1}{n_i} \max_{j=1}^k \{n_{ij}\}$$

❑ Total purity of clustering C :

$$purity = \sum_{i=1}^r \frac{n_i}{n} purity_i = \frac{1}{n} \sum_{i=1}^r \max_{j=1}^k \{n_{ij}\}$$

❑ Perfect clustering if $purity = 1$ and $r = k$ (the number of clusters obtained is the same as that in the ground truth)

❑ Ex. 1 (green or orange): $purity_1 = 30/50$; $purity_2 = 20/25$; $purity_3 = 25/25$; $purity = (30 + 20 + 25)/100 = 0.75$

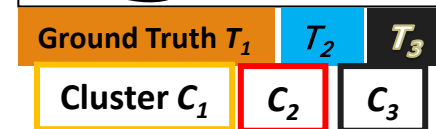
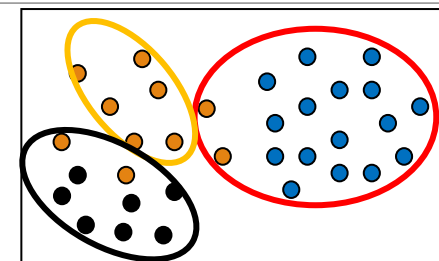
❑ Two clusters may share the same majority partition

❑ **Maximum matching:** Only one cluster can match one partition

❑ Match: Pairwise matching, weight $w(e_{ij}) = n_{ij}$ $w(M) = \sum_{e \in M} w(e)$

❑ Maximum weight matching: $match = \arg \max_M \left\{ \frac{w(M)}{n} \right\}$

❑ Ex2. (green) $match = purity = 0.75$; (orange) $match = 0.65 > 0.6$



$C \backslash T$	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

$C \backslash T$	T_1	T_2	T_3	Sum
C_1	0	30	20	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	50	25	100



Matching-Based Measures (II): F-Measure

- Precision:** The fraction of points in C_i from the majority partition T_{j_i} (i.e., the same as purity), where j_i is the partition that contains the maximum # of points from C_i

- Ex. For the green table

$$prec_1 = 30/50; prec_2 = 20/25; prec_3 = 25/25$$

- Recall:** The fraction of point in partition T_{j_i} shared in common with cluster C_i , where $m_{j_i} = |T_{j_i}|$

- Ex. For the green table

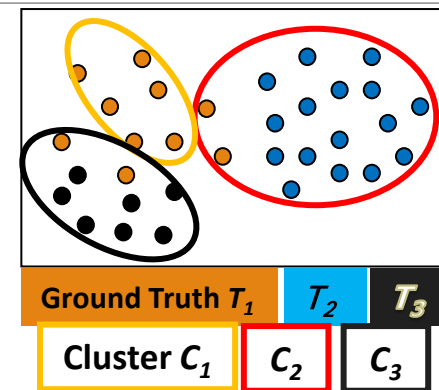
$$recall_1 = 30/35; recall_2 = 20/40; recall_3 = 25/25$$

- F-measure** for C_i : The harmonic means of $prec_i$ and $recall_i$: $F_i = \frac{2n_{ij_i}}{n_i + m_{j_i}}$

- F-measure** for clustering C : average of all clusters: $F = \frac{1}{r} \sum_{i=1}^r F_i$

- Ex. For the green table

$$F_1 = 60/85; F_2 = 40/65; F_3 = 1; F = 0.774$$



$C \backslash T$	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100