External Measures III: Pairwise Measures

Pairwise Measures: Four Possibilities for Truth Assignment

❑ **Four possibilities** based on the agreement between cluster label and partition label

    ❑ *TP*: true positive—Two points x_i and x_j belong to the same partition T, and they also in the same cluster C

$$TP = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j \text{ and } \hat{y}_i = \hat{y}_j\}|$$

    where y_i: the true partition label, and $\hat{y}_i$: the cluster label for point x_i

    ❑ *FN*: false negative: $\quad FN = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j \text{ and } \hat{y}_i \neq \hat{y}_j\}|$

    ❑ *FP: false positive* $\quad FP = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i \neq y_j \text{ and } \hat{y}_i = \hat{y}_j\}|$

    ❑ *TN*: true negative $\quad TN = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i \neq y_j \text{ and } \hat{y}_i \neq \hat{y}_j\}|$
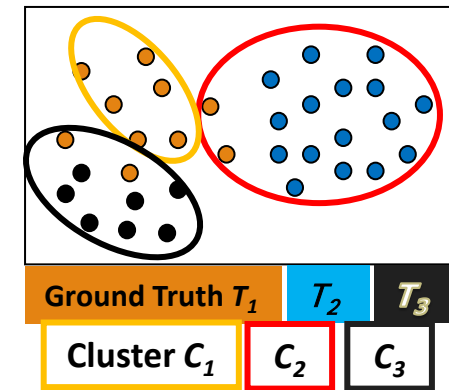
❑ Calculate the four measures:

Ground Truth T_1	T_2	T_3
Cluster C_1	C_2	C_3

$$N = \binom{n}{2}$$ Total # of pairs of points

$$TP = \sum_{i=1}^{r}\sum_{j=1}^{k}\binom{n_{ij}}{2} = \frac{1}{2}((\sum_{i=1}^{r}\sum_{j=1}^{k}n_{ij}^{2}) - n) \quad FN = \sum_{j=1}^{k}\binom{m_j}{2} - TP$$

$$FP = \sum_{i=1}^{r}\binom{n_i}{2} - TP \quad TN = N - (TP + FN + FP) = \frac{1}{2}(n^2 - \sum_{i=1}^{r}n_i^2 - \sum_{j=1}^{k}m_j^2 + \sum_{i=1}^{r}\sum_{j=1}^{k}n_{ij}^2)$$

2

Pairwise Measures: Jaccard Coefficient and Rand Statistic

- ❑ **Jaccard coefficient:** Fraction of true positive point pairs, but after ignoring the true negatives (thus asymmetric)
  - ❑ *Jaccard = TP/(TP + FN + FP)* [i.e., denominator ignores *TN*]
  - ❑ Perfect clustering: *Jaccard = 1*
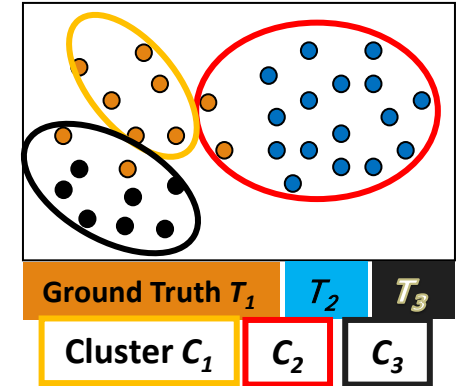- ❑ **Rand Statistic**:
  - ❑ *Rand = (TP + TN)/N*
  - ❑ Symmetric; perfect clustering: *Rand = 1*
- ❑ **Fowlkes-Mallow Measure**:
  - ❑ Geometric mean of precision and recall

$$FM = \sqrt{prec \times recall} = \frac{TP}{\sqrt{(TP + FN)(TP + FP)}}$$

- ❑ Using the above formulas, one can calculate all the measures for the green table (leave as an exercise)

Ground Truth T_1	T_2	T_3

Cluster C_1	C_2	C_3

C\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