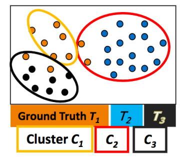
Entropy-Based Measures (II): Normalized Mutual Information (NMI)

- **Mutual information**:
  - Quantifies the amount of shared info between the clustering C and partitioning T $\quad I(C,T)=\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij}\log(\frac{p_{ij}}{p_{C_i}\cdot p_{T_j}})$
  - Measures the dependency between the observed joint probability p_{ij} of C and T, and the expected joint probability $p_{Ci} \cdot p_{Tj}$ under the independence assumption
  - When C and T are independent, $p_{ij} = p_{Ci} \cdot p_{Tj}$, $I(C, T) = 0$. However, there is no upper bound on the mutual information
- **Normalized mutual information** (NMI)



$$NMI(\mathcal{C},\mathcal{T})=\sqrt{\frac{I(\mathcal{C},\mathcal{T})}{H(\mathcal{C})}\cdot\frac{I(\mathcal{C},\mathcal{T})}{H(\mathcal{T})}}=\frac{I(\mathcal{C},\mathcal{T})}{\sqrt{H(\mathcal{C})\cdot H(\mathcal{T})}}$$

  - Value range of NMI: [0,1]. Value close to 1 indicates a good clustering

Entropy of clustering C: $\quad H(\mathcal{C})=-\sum_{i=1}^{r}p_{C_i}\log p_{C_i}$

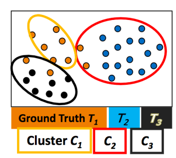
Entropy of partitioning T: $\quad H(\mathcal{T})=-\sum_{j=1}^{k}p_{T_j}\log p_{T_j}$

Pairwise Measures: Four Possibilities for Truth Assignment

- **Four possibilities** based on the agreement between cluster label and partition label
  - *TP*: true positive—Two points $\mathbf{x}_i$ and $\mathbf{x}_j$ 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 $\mathbf{x}_i$



  - *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: $\quad N=\binom{n}{2}$ Total # of pairs of points

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

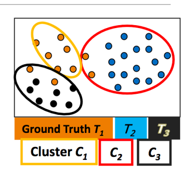
$$FP=\sum_{i=1}^{r}\binom{n_i}{2}-TP\qquad TN=N-(TP+FN+FP)=\frac{1}{2}(n^2-\sum_{i=1}^{r}n_i^2-\sum_{j=1}^{k}m_j^2+\sum_{i=1}^{r}\sum_{j=1}^{k}n_{ij}^2)$$

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)



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