

Clustering Evaluation

Dr, Nimisha Roy
Georgia Tech

Clustering Evaluation

- Clustering evaluation aims at quantifying the goodness or quality of the clustering.
- Two main categories of measures:
 - **External measures**: employ external ground-truth
 - **Internal measures**: derive goodness from the data itself

Outline

- External measures for clustering evaluation
 - Matching-based measures
 - Entropy-based measures
 - Pairwise measures
- Internal measures for clustering evaluation
 - Graph-based measures
 - Davies-Bouldin Index
 - Silhouette Coefficient

External Measures

External measures assume that the correct or ground-truth clustering is known *a priori*, which is used to evaluate a given clustering.

Let $\mathbf{D} = \{\mathbf{x}_i\}_{i=1}^n$ be a dataset consisting of n points in a d -dimensional space, partitioned into k clusters. Let $y_i \in \{1, 2, \dots, k\}$ denote the ground-truth cluster membership or label information for each point.

The ground-truth clustering is given as $\mathcal{T} = \{T_1, T_2, \dots, T_k\}$, where the cluster T_j consists of all the points with label j , i.e., $T_j = \{\mathbf{x}_i \in \mathbf{D} | y_i = j\}$. We refer to \mathcal{T} as the ground-truth *partitioning*, and to each T_j as a *partition*.

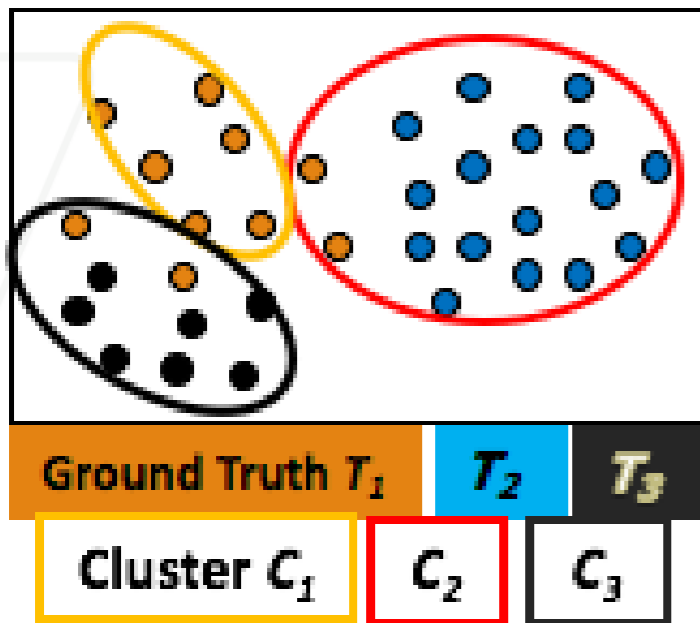
Let $\mathcal{C} = \{C_1, \dots, C_r\}$ denote a clustering of the same dataset into r clusters, obtained via some clustering algorithm, and let $\hat{y}_i \in \{1, 2, \dots, r\}$ denote the cluster label for \mathbf{x}_i .

So **k** is the number of ground truth partitions (T) and **r** is the number of clusters (C) obtained by algorithm

n_{ij} = Number of data points in cluster **i** which are also in ground truth partition **j**

Matching-Based Measures (I): Purity

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



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

$$purity_3 = \frac{1}{n_3} \max(n_{31}, n_{32}, n_{33})$$

$$= \frac{1}{9} \max(2, 0, 7) = \frac{7}{9}$$

The Total purity of clustering C is the weighted sum of the cluster-wise purity:

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

What is purity value for a perfect clustering?

Purity = 1

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

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

Example:

$$purity_1 = 30/50;$$

$$purity_2 = 20/25;$$

$$purity_3 = 25/25;$$

$$purity = (30 + 20 + 25)/100 = 0.75$$

$C \setminus T$	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

Two clusters may be matched to the same partition

C1 is more paired with T3
C2 is more paired with T2

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

$$\text{purity} = (30 + 20 + 25)/100 = 0.75$$

C1 is more paired with T2
C2 is more paired with T2

C\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

$$\text{purity} = (30 + 20 + 25)/100 = 0.75$$

Maximum weight matching: Only one cluster can match one partition

Ex. If C1 is more paired with T2 **THEN** C2 and C3 cannot be paired with T2

C\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

C1 is more paired with T2 = $\frac{30+5+25}{100} = 0.6$

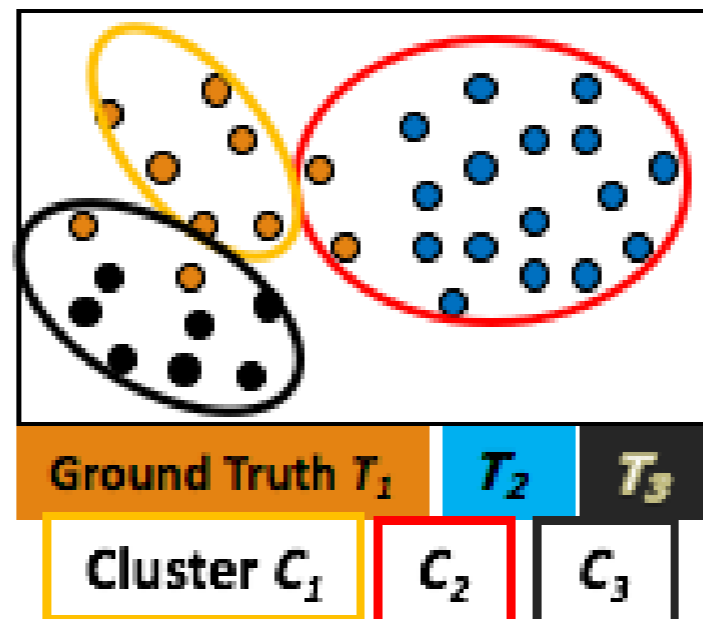
C1 is more paired with T3 = $\frac{20+20+25}{100} = 0.65$

MAX

Purity = 0.65

Matching-Based Measures (II): Maximum Matching

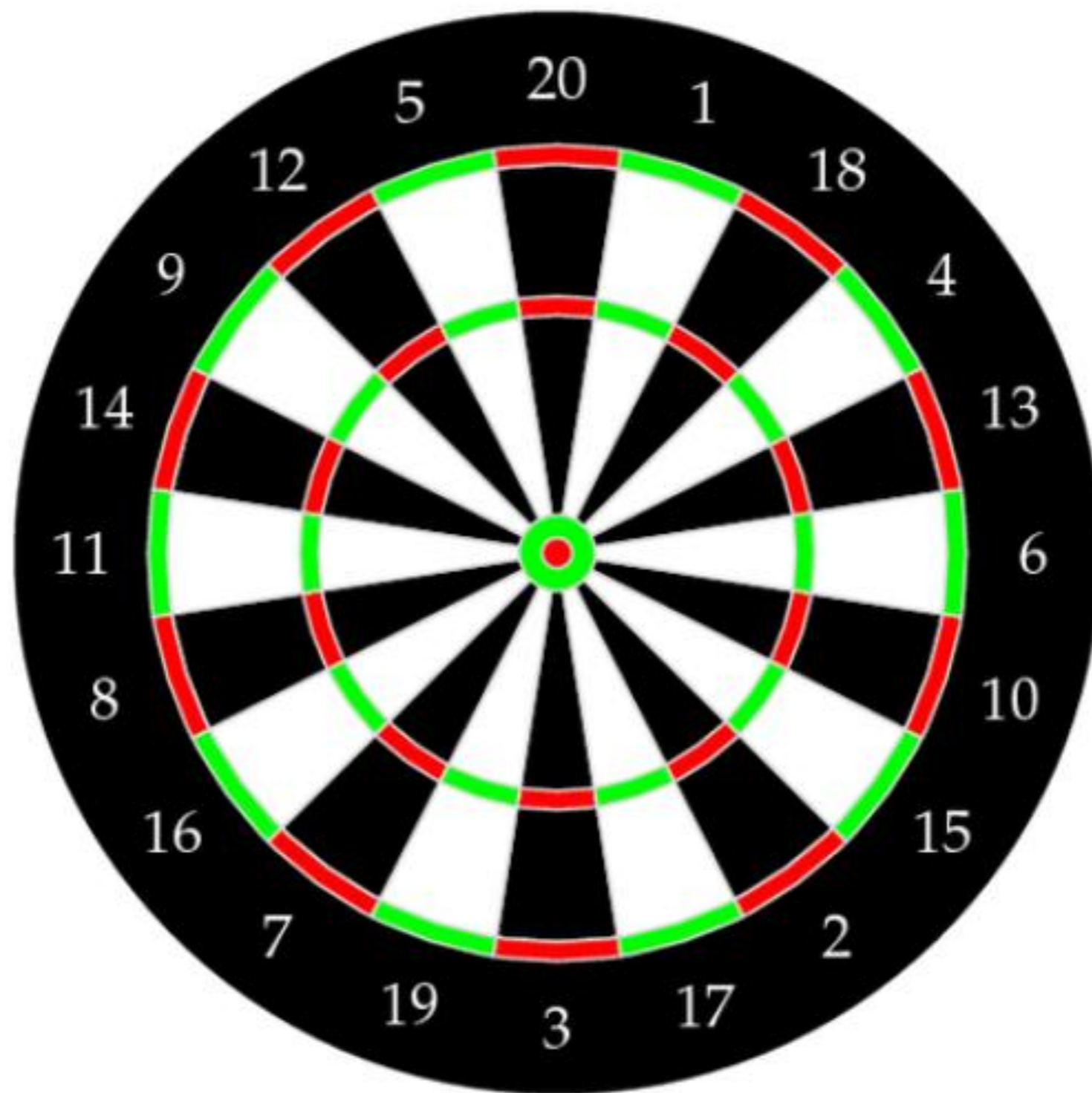
- **Drawback of purity:** two clusters may be matched to the same partition.
- **Maximum matching:** the maximum purity under the one-to-one matching constraint.
 - Examine all possible pairwise matching between C and T and choose the best (the maximum)



$C \setminus T$	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

Example:

Maximum matching = $0.65 > 0.6$



In a general context: Precision, Recall and Accuracy

Correct prediction

Wrong prediction

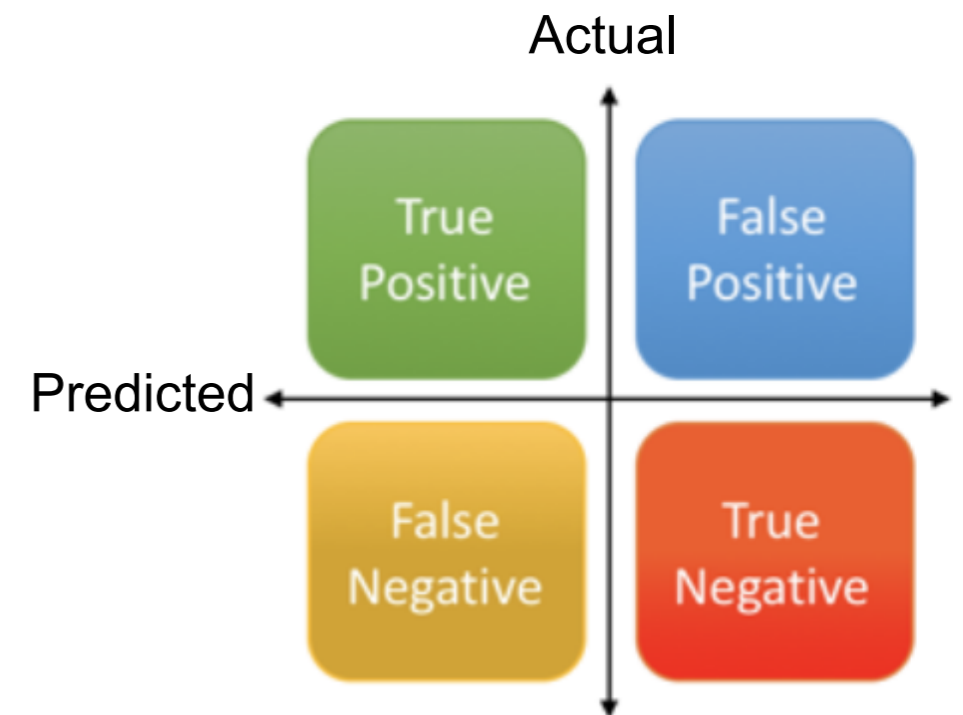
Number of predicted “positive” labeled data = True Positive + False Positive

Number of predicted “negative” labeled data = True Negative + False Negative

$$\text{Precision} = \frac{\textit{True Positive}}{\textit{Predicted Results}} \text{ or } \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\textit{True Positive}}{\textit{Actual Results}} \text{ or } \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total}}$$



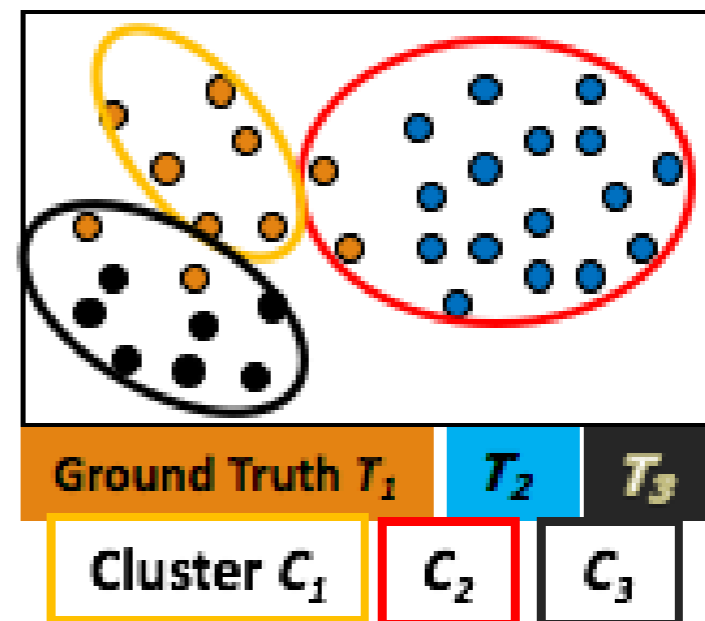
False positive is also called false alarm

Matching-Based Measures (II): F-Measure

- **Precision:** which measure *quality*, is the same as purity:
 - How precisely does each cluster represent the ground truth?
 - Of all predicted positives, how many are actually positive?
- **Recall:** measures completeness
 - How completely does each cluster recover the ground truth?
 - Of all actual positives, how many were correctly predicted?

$$recall_i = \frac{n_{ij_i}}{|T_{j_i}|} = \frac{n_{ij_i}}{m_{j_i}} \quad Recall_1 = \frac{6}{10} \quad prec_i = \frac{1}{n_i} \max_{j=1}^k \{n_{ij}\} = \frac{n_{ij_i}}{n_i} \quad Prec_1 = \frac{6}{6}$$

- C_i = i -th **cluster** produced by the algorithm
- T_j^* = **ground truth partition** that best matches cluster C_i
- n_{ij}^* = number of shared data points between C_i and T_j^*
- m_j^* = total number of points in that ground truth partition T_j^*



Precision and Recall

(Precision here is same as the purity)

Precision:

$$\text{prec}_1 = 30/50;$$

$$\text{prec}_2 = 20/25;$$

$$\text{prec}_3 = 25/25$$

Recall:

$$\text{recall}_1 = 30/35;$$

$$\text{recall}_2 = 20/40;$$

$$\text{recall}_3 = 25/25$$

$C \setminus T$	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

Matching-Based Measures (II): F-Measure

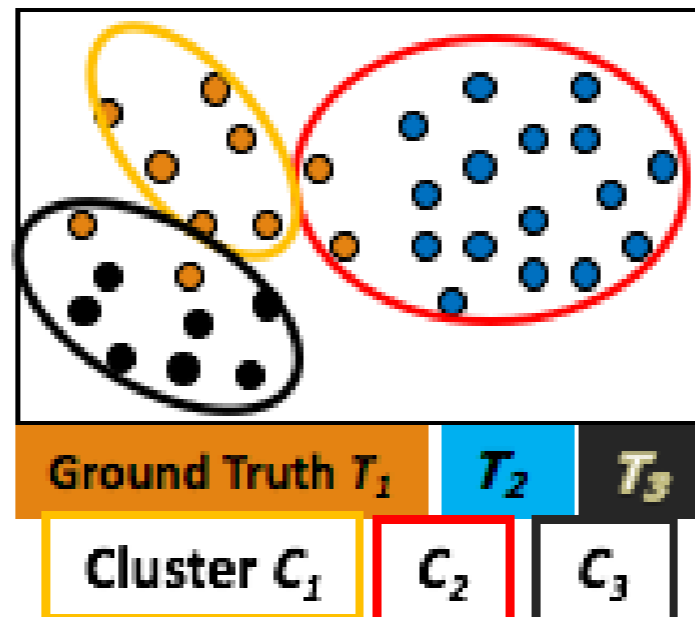
- **F-Measure**: the harmonic mean of precision and recall
 - Take into account both *precision* and *completeness*

$$F_i = \frac{2}{\frac{1}{prec_i} + \frac{1}{recall_i}} = \frac{2 \cdot prec_i \cdot recall_i}{prec_i + recall_i} = \frac{2 n_{ij_i}}{n_i + m_{j_i}}$$

The F-measure for the clustering \mathcal{C} is the mean of clusterwise F-measure values:

$$F = \frac{1}{r} \sum_{i=1}^r F_i$$

F1 = 60/85;
 F2 = 40/65;
 F3 = 1;
 F = 0.774



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

Entropy-Based Measures

- Expectation
- Information
- Entropy
- We want entropy of a cluster to be high or low?
- Conditional Entropy

Entropy-Based Measures (I): Conditional Entropy

Conditional entropy measures how mixed the true labels are inside clusters.

Pure clusters \rightarrow entropy near 0.

Mixed clusters \rightarrow high entropy.

$H(T|C)$: How uncertain are the true labels T once we know the cluster assignment C ?

Interpretation:

- Low value \rightarrow clusters match true classes well
- High value \rightarrow clusters mix many different classes

Entropy-Based Measures (I): Conditional Entropy

Amount of information orderliness in different partitions

- The entropy for clustering C and partition T is

$$H(C) = - \sum_{i=1}^r p_{C_i} \log p_{C_i}$$

$$H(T) = - \sum_{j=1}^k p_{T_j} \log p_{T_j}$$

where $p_{C_i} = \frac{n_i}{n}$ and $p_{T_j} = \frac{m_j}{n}$

← i.e., The probability of ground truth T_j

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

Entropy-Based Measures (I): Conditional Entropy

- **Conditional Entropy:** The cluster-specific entropy, namely the conditional entropy of \mathcal{T} with respect to cluster C_i :

$$H(\mathcal{T} | C_i) = - \sum_{j=1}^k \left(\frac{n_{ij}}{n_i} \right) \log \left(\frac{n_{ij}}{n_i} \right)$$

n_{ij} ← Ground truth (T)
Cluster (C)

How ground truth is distributed within each cluster

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

Entropy-Based Measures (I): Conditional Entropy

- The conditional entropy of \mathcal{T} given clustering \mathcal{C} is defined as the **weighted average of the entropy inside each cluster**.

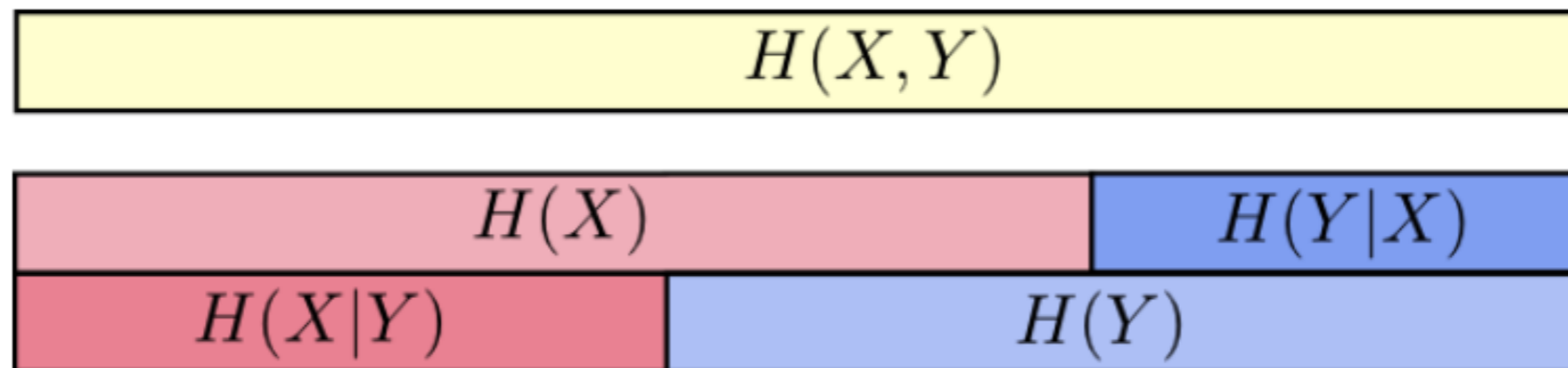
$$\begin{aligned} H(\mathcal{T}|\mathcal{C}) &= \sum_{i=1}^r \frac{n_i}{n} H(\mathcal{T}|C_i) = - \sum_{i=1}^r \sum_{j=1}^k p_{ij} \log \left(\frac{p_{ij}}{p_{C_i}} \right) \\ &= H(\mathcal{C}, \mathcal{T}) - H(\mathcal{C}) \end{aligned}$$

$\frac{n_{ij}}{n}$
 $\frac{n_i}{n}$

The more clusters members are split into different partitions, the higher the conditional entropy (not a desirable condition and the max value is $\log k$)

$H(\mathcal{T}|\mathcal{C}) = 0$ if and only if \mathcal{T} is completely determined by \mathcal{C} , corresponding to the ideal clustering. If \mathcal{C} and \mathcal{T} are independent of each other, then $H(\mathcal{T}|\mathcal{C}) = H(\mathcal{T})$.

$$\begin{aligned}
H(\mathcal{T}|\mathcal{C}) &= - \sum_{i=1}^r \sum_{j=1}^k p_{ij} \log \frac{p_{ij}}{p_{c_i}} \\
&= - \sum_{i=1}^r \sum_{j=1}^k p_{ij} (\log p_{ij} - \log p_{c_i}) = - \sum_{i=1}^r \sum_{j=1}^k p_{ij} (\log p_{ij}) + \sum_{i=1}^r (\log p_{c_i} \sum_{j=1}^k p_{ij}) \\
&\quad - \sum_{i=1}^r \sum_{j=1}^k p_{ij} \log p_{ij} + \sum_{i=1}^r (p_{c_i} \log p_{c_i}) = H(\mathcal{T}, \mathcal{C}) - H(\mathcal{C})
\end{aligned}$$

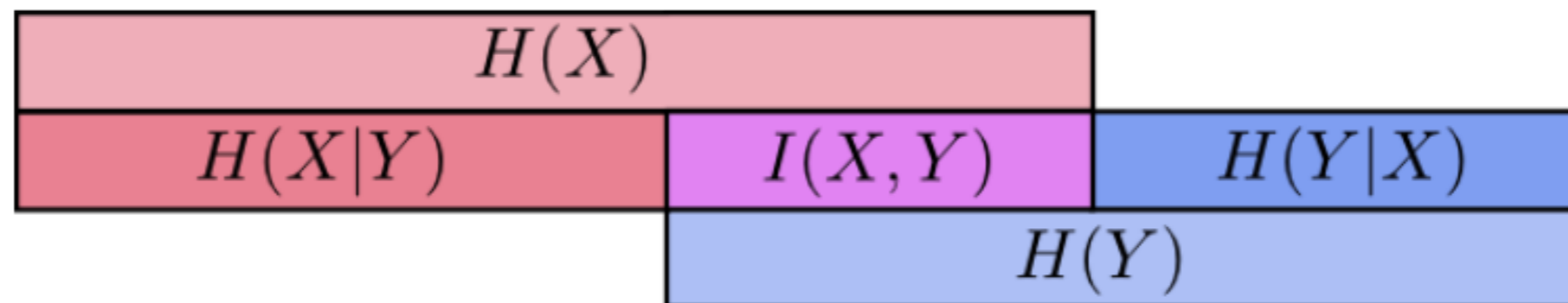


Entropy-Based Measures (I): Mutual Information

The *mutual information* tries to quantify the amount of shared information between the clustering \mathcal{C} and partitioning \mathcal{T} , and it is defined as

$$I(\mathcal{C}, \mathcal{T}) = \sum_{i=1}^r \sum_{j=1}^k p_{ij} \log \left(\frac{p_{ij}}{p_{C_i} \cdot p_{T_j}} \right) = H(\mathcal{T}) - H(\mathcal{T}|\mathcal{C})$$

When \mathcal{C} and \mathcal{T} are independent then $p_{ij} = p_{C_i} \cdot p_{T_j}$, and thus $I(\mathcal{C}, \mathcal{T}) = 0$. However, there is no upper bound on the mutual information.

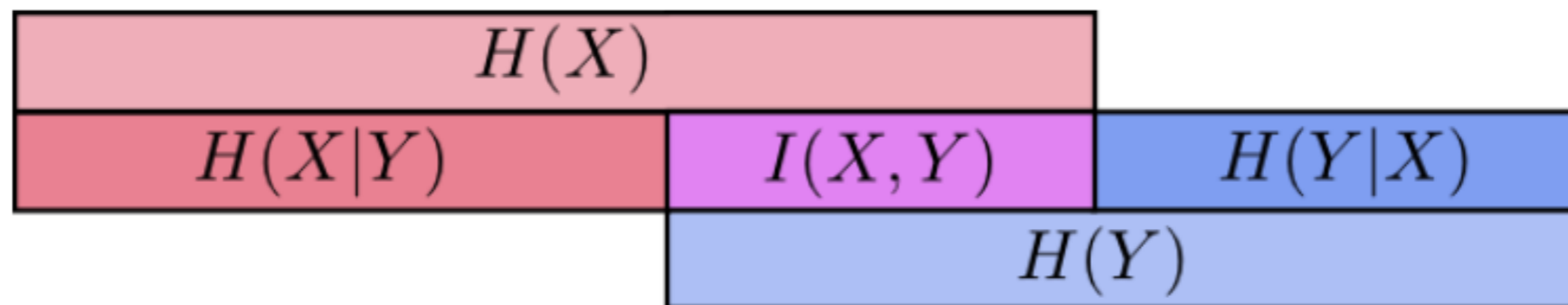


Entropy-Based Measures (I): Mutual Information

The *normalized mutual information* (NMI) is defined as the geometric mean:

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

The NMI value lies in the range $[0, 1]$. Values close to 1 indicate a good clustering.

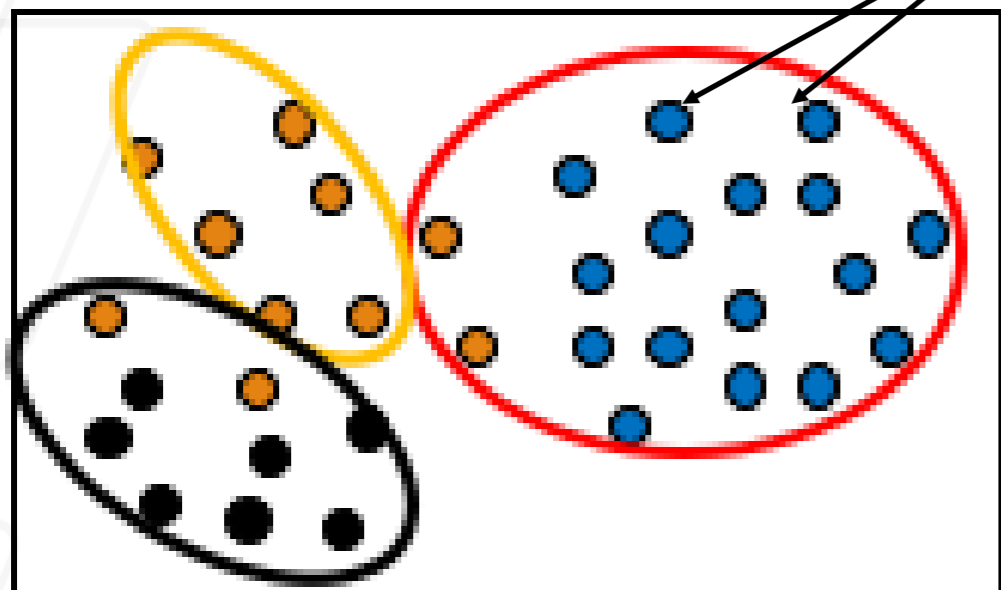


Pairwise Measures

These metrics evaluate clustering based on how well it groups **pairs of points** that should (or should not) be together.

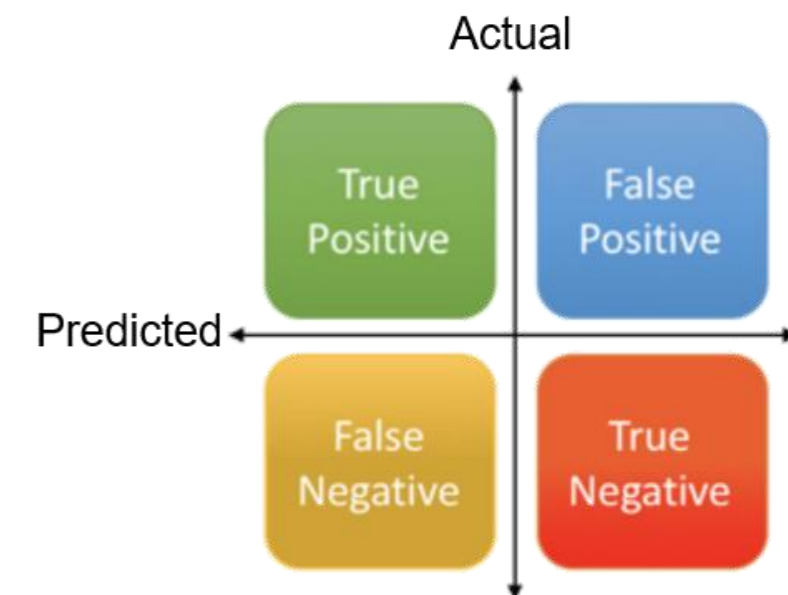
Given clustering \mathcal{C} and ground-truth partitioning \mathcal{T} , let $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{D}$ be any two points, with $i \neq j$. Let y_i denote the true partition label and let \hat{y}_i denote the cluster label for point \mathbf{x}_i .

Pairwise Measures



Same partition

Same cluster

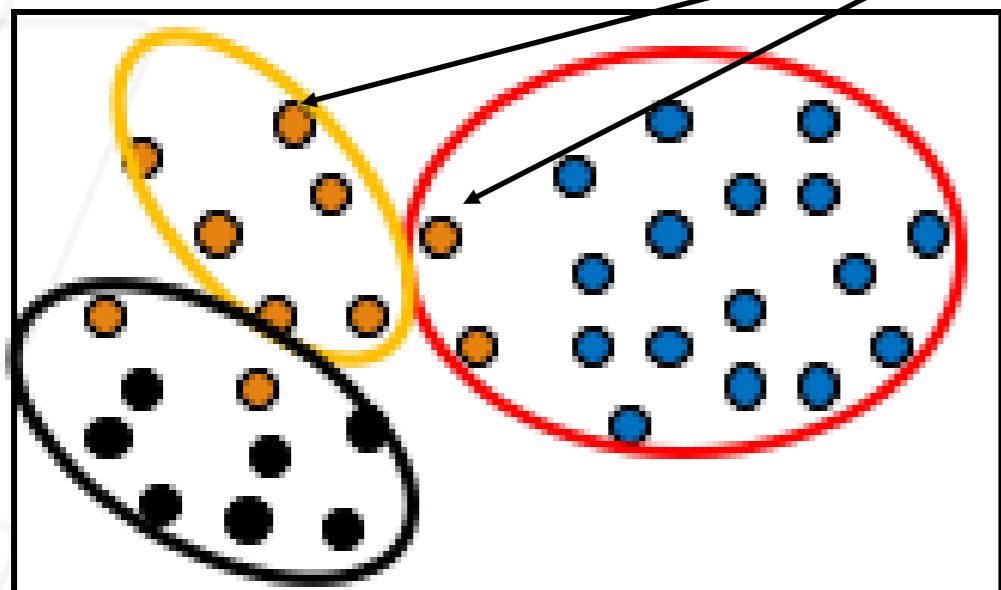


$C \setminus T$	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

True Positives: \mathbf{x}_i and \mathbf{x}_j belong to the same partition in \mathcal{T} , and they are also in the same cluster in \mathcal{C} . The number of true positive pairs is given as

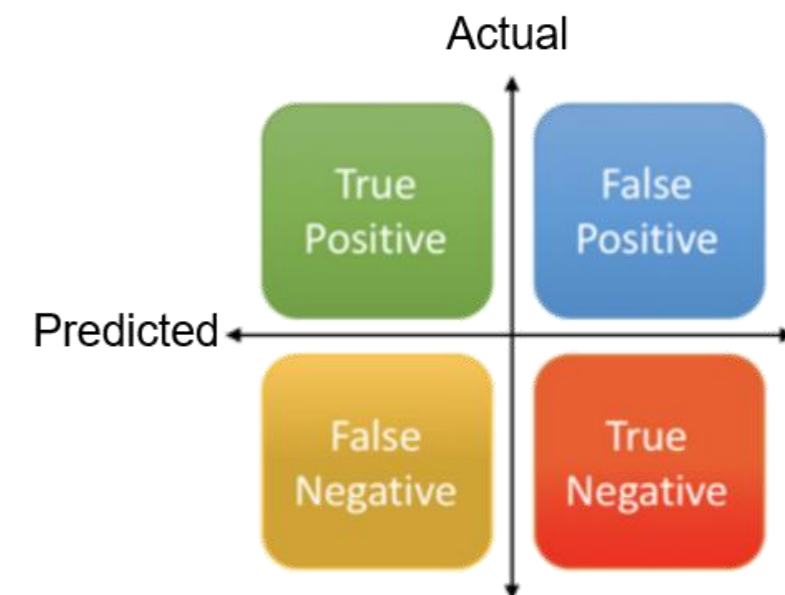
$$TP = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j \text{ and } \hat{y}_i = \hat{y}_j\}|$$

Pairwise Measures



Same partition

Different cluster

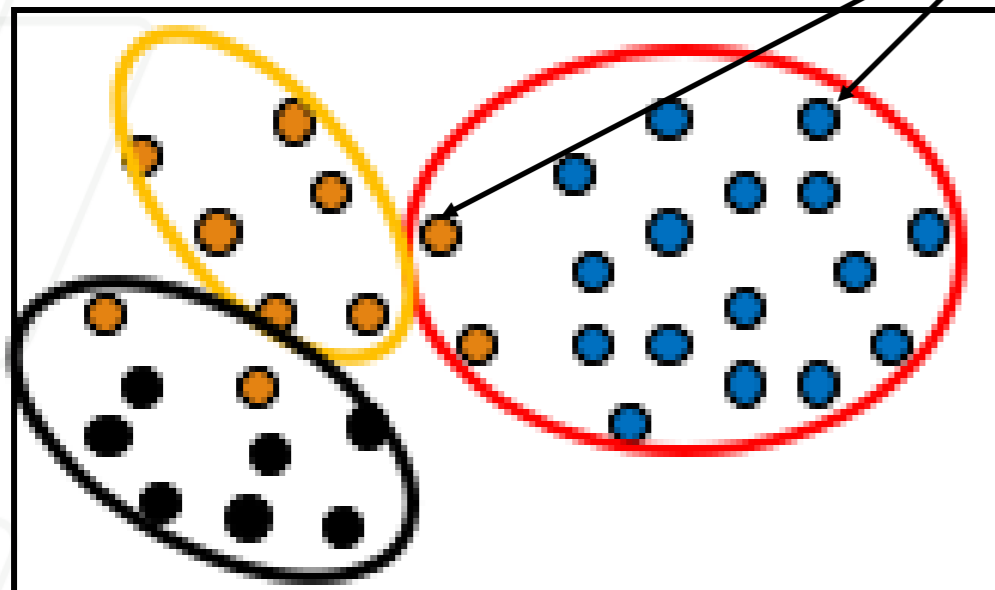


$C \setminus T$	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

False Negatives: \mathbf{x}_i and \mathbf{x}_j belong to the same partition in \mathcal{T} , but they do not belong to the same cluster in \mathcal{C} . The number of all false negative pairs is given as

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

Pairwise Measures



Ground Truth T_1

T_2

T_3

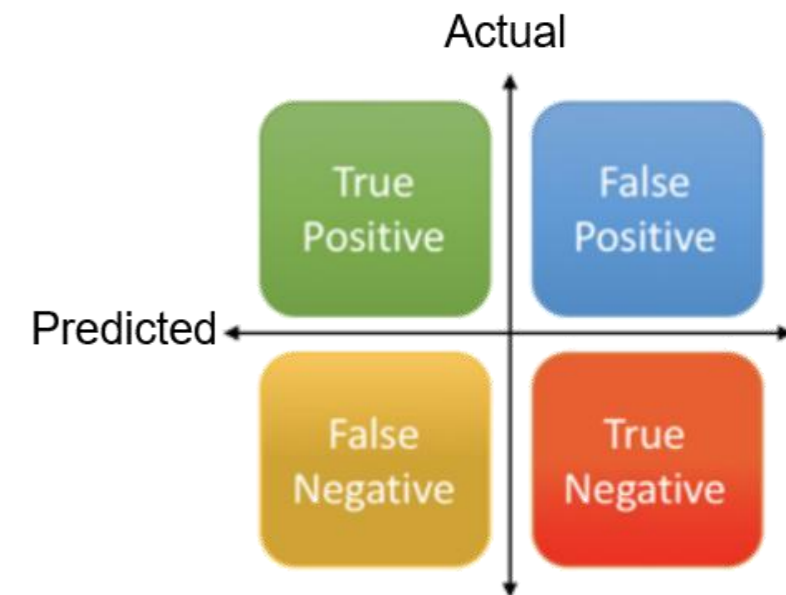
Cluster C_1

C_2

C_3

Different partition

Same cluster

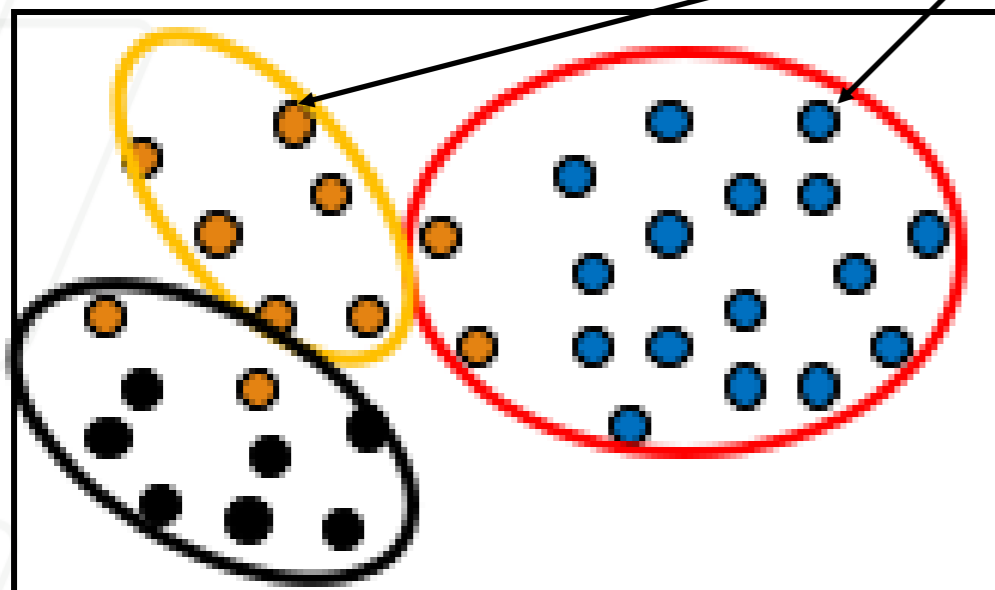


$C \setminus T$	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

False Positives: \mathbf{x}_i and \mathbf{x}_j do not belong to the same partition in \mathcal{T} , but they do belong to the same cluster in \mathcal{C} . The number of false positive pairs is given as

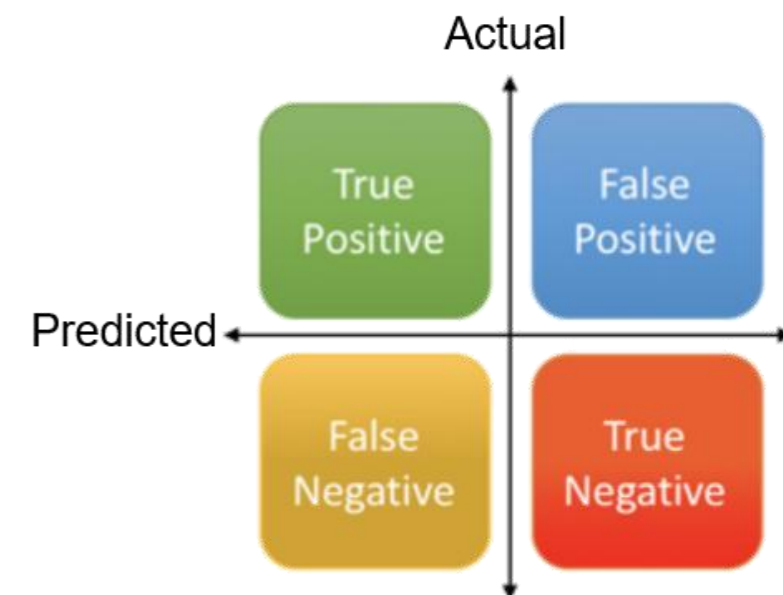
$$FP = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i \neq y_j \text{ and } \hat{y}_i = \hat{y}_j\}|$$

Pairwise Measures



Different partition

Different cluster

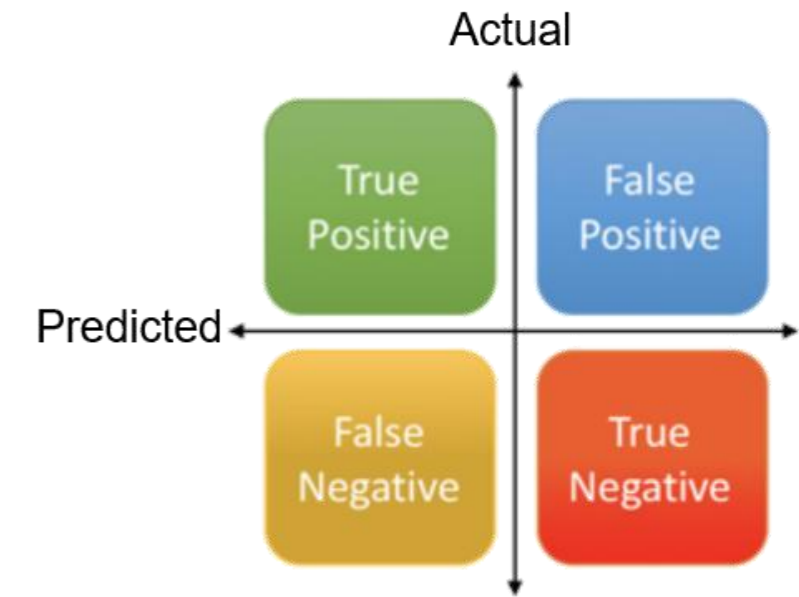


$C \setminus T$	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

True Negatives: \mathbf{x}_i and \mathbf{x}_j neither belong to the same partition in \mathcal{T} , nor do they belong to the same cluster in \mathcal{C} . The number of such true negative pairs is given as

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

Pairwise Measures



Because there are $N = \binom{n}{2} = \frac{n(n-1)}{2}$ pairs of points, we have the following identity:

$$N = TP + FN + FP + TN$$

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

$$FP = \sum_{i=1}^r \binom{n_i}{2} - TP$$

$$FN = \sum_{j=1}^k \binom{m_j}{2} - TP$$

$$TN = N - (TP + FN + FP)$$

Pairwise Measures: Jaccard Coefficient

Jaccard Coefficient: measures the fraction of true positive point pairs, but after ignoring the true negative:

$$Jaccard = \frac{TP}{TP + FN + FP}$$

Measures **similarity** between clustering and ground truth — focuses on positive pairs only (ignores TN).

A **perfect clustering** gives Jaccard = 1.

It answers: “Of all pairs that should or were grouped together, what fraction are correctly grouped?”

Pairwise Measures: Rand

Rand Statistic: measures the fraction of true positives and true negatives over all point pairs:

$$Rand = \frac{TP + TN}{N}$$

- Measures the **fraction of agreements** (both same-cluster and different-cluster pairs) between clustering and ground truth.
- Includes true negatives (pairs correctly not grouped together).
- Conceptually similar to **accuracy** in classification.
- **Perfect clustering** → Rand = 1.
- **Downside:** Because TN pairs dominate (most pairs are in different clusters), Rand Index can appear high even if clustering is mediocre.

Pairwise Measures: FM

Fowlkes-Mallows Measure: Define the overall *pairwise precision* and *pairwise recall* values for a clustering \mathcal{C} , as follows:

$$prec = TP / TP + FP$$

$$recall = TP / TP + FN$$

The Fowlkes–Mallows (FM) measure is defined as the geometric mean of the pairwise precision and recall

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

- It combines **pairwise precision** and **pairwise recall** — just like the F1 score did for classification.
- A **perfect clustering** gives $FM = 1$.

Jaccard and FM are preferred when number of TNs is large, since they avoid the dominance of negatives. Rand Index is simple and interpretable but can be inflated when most pairs are different.

Summary

- External measures for clustering evaluation
 - Matching-based measures
 - Entropy-based measures
 - Pairwise measures
- Internal measures for clustering evaluation
 - Graph-based measures
 - Davies-Bouldin Index
 - Silhouette Coefficient

Quick Knowledge Check

1. **What does purity measure?** A. Compactness of clusters. B. Homogeneity with respect to ground truth. C. Separation between clusters
2. **What problem does maximum matching fix?** A. Duplicate cluster-partition matches. B. Small cluster bias. C. Missing cluster labels
3. **Why is the harmonic mean used in F-measure?** A. To emphasize the larger value. B. To balance precision and recall fairly. C. To simplify computation
4. **What does conditional entropy $H(T | C)$ represent?** A. Uncertainty of true labels within clusters. B. Compactness of clusters. C. Distance between centroids
5. **Higher conditional entropy implies:** A. Better clustering. B. Poorer clustering. C. No change in clustering quality
6. **Why doesn't mutual information have a universal upper bound in clustering evaluation?** A. It depends on the number and size of clusters, which can vary. B. It ignores entropy in computation
7. **Jaccard coefficient ignores:** A. True negatives. B. False positives. C. False negatives
8. **Rand Index is most similar to:** A. Accuracy. B. Precision. C. Recall
9. **Why can Rand Index appear high even for poor clustering?** A. TN pairs dominate. B. FP pairs dominate. C. FN pairs dominate