The week ahead

- **Quiz 4:** mean is 85% and average completion time 5min 40sec!
- **Assignment 2 Early bird special** → 1 complete programming question by Wed, Sep 23rd
- **Fifth round of project seminars**, available Thursday, Sep 24th
- **HW1 grades are out!** Regrade requests by Fri, Sep 25th
- **Open office hours on Thursday, 7pm to 8pm**
  - https://primetime.bluejeans.com/a2m/live-event/qfsqxjec
- **Quiz 5, Friday, Sep 25th 6am until Sep 26th 11:59am (noon)**
  - Hierarchical clustering, cluster evaluation, density estimation

Coming up soon

- **Touch-point 1**: deliverables due Mon, Sep 28th, live-event Wed, Sep 30th
- **Project proposal due Oct 2nd 11:59pm (midnight)**
- **Assignment 2 due Oct 5th 11:59pm (midnight)**

CS4641B Machine Learning

Lecture 10: Clustering evaluation

Rodrigo Borela ▸ rborelav@gatech.edu

These slides are based on slides from Mohammed Zaki, Chao Zhang, Jiawei Han and Mahdi Roozbahani

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

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

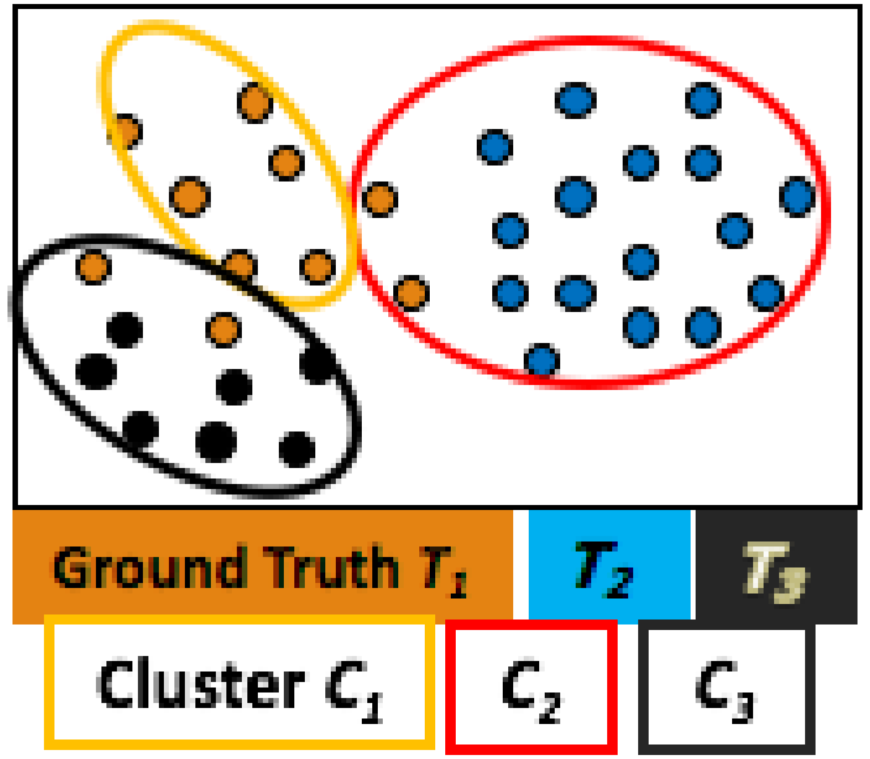
External measures

- External measures **assume that the correct or ground-truth clustering is known a priori**, which is used to evaluate a given clustering
- Let $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^{N}$ be a dataset consisting of N points in a D-dimensional space, partitioned into K clusters. Let $y_n \in \{1,2,\dots,K\}$ 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_k consists of all the points with label k, i.e. $T_k = \{\mathbf{x}_n \in \mathbf{X} | y_n = k\}$. We refer to $\mathcal{T}$ as the ground-truth *partitioning*, and to each T_k as a *partition*.
- Let $\mathcal{C} = \{C_1, C_2, \dots, C_R\}$ denote a clustering of the same dataset into R clusters, obtained via some clustering algorithm, and let $\hat{y}_n \in \{1,2,\dots,R\}$ denote the cluster label for $\mathbf{x}_n$.

- So K is the number of ground-truth partitions ($\mathcal{T}$) and R is the number of clusters ($\mathcal{C}$) obtained by algorithm
- $n_{rk} =$ Number of data points in cluster $\boldsymbol{r}$ which are also in ground-truth partition $\boldsymbol{k}$

Matching-based measures: Purity

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



$$purity_r = \frac{1}{n_r} \max_{k=1}^{K}\{n_{rk}\}$$

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

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

- The total purity of clustering $\mathcal{C}$ is the weighted sum of the cluster-wise purity:

$$purity = \sum_{r=1}^{R} \frac{n_r}{N} purity_r = \frac{1}{N} \sum_{r=1}^{R} \max_{k=1}^{K}\{n_{rk}\}$$

- What is purity value for a perfect clustering? $purity = 1$

Purity: example

$$purity_r = \frac{1}{n_r} \max_{k=1}^{K}\{n_{rk}\}$$

$$purity = \sum_{r=1}^{R} \frac{n_r}{N} purity_r = \frac{1}{N} \sum_{r=1}^{R} \max_{k=1}^{K}\{n_{rk}\}$$

$C \backslash T$	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

$$purity_1 = 30/50$$
$$purity_2 = 20/25$$
$$purity_3 = 25/25$$

$$purity = \frac{30 + 20 + 25}{100} = 0.75$$

Purity: example

- Two clusters may be matched to the same partition

C_1 is more paired with T_3
C_2 is more paired with T_2

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

$$purity = \frac{30 + 20 + 25}{100} = 0.75$$

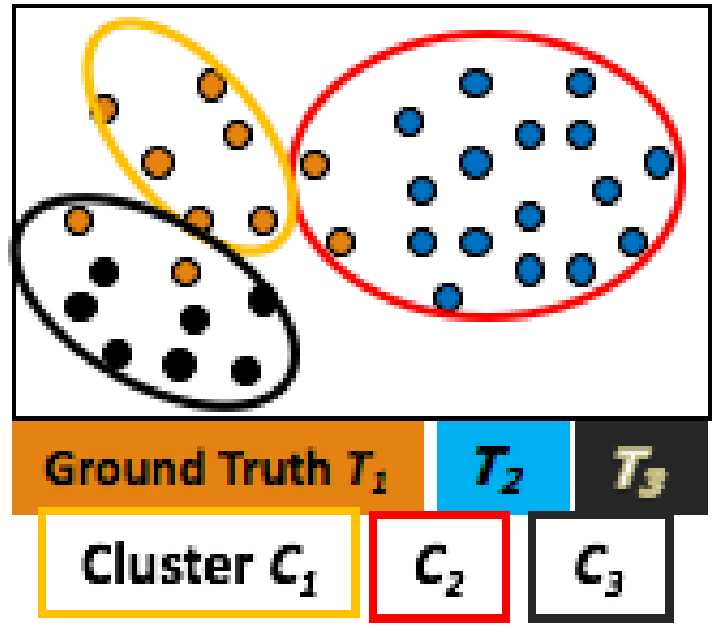
C_1 is more paired with T_2
C_2 is more paired with T_2

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

$$purity = \frac{30 + 20 + 25}{100} = 0.75$$

Matching-based measures: 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)



$C \backslash T$	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

Example:
Maximum matching = 0.65 > 0.6

Purity: example

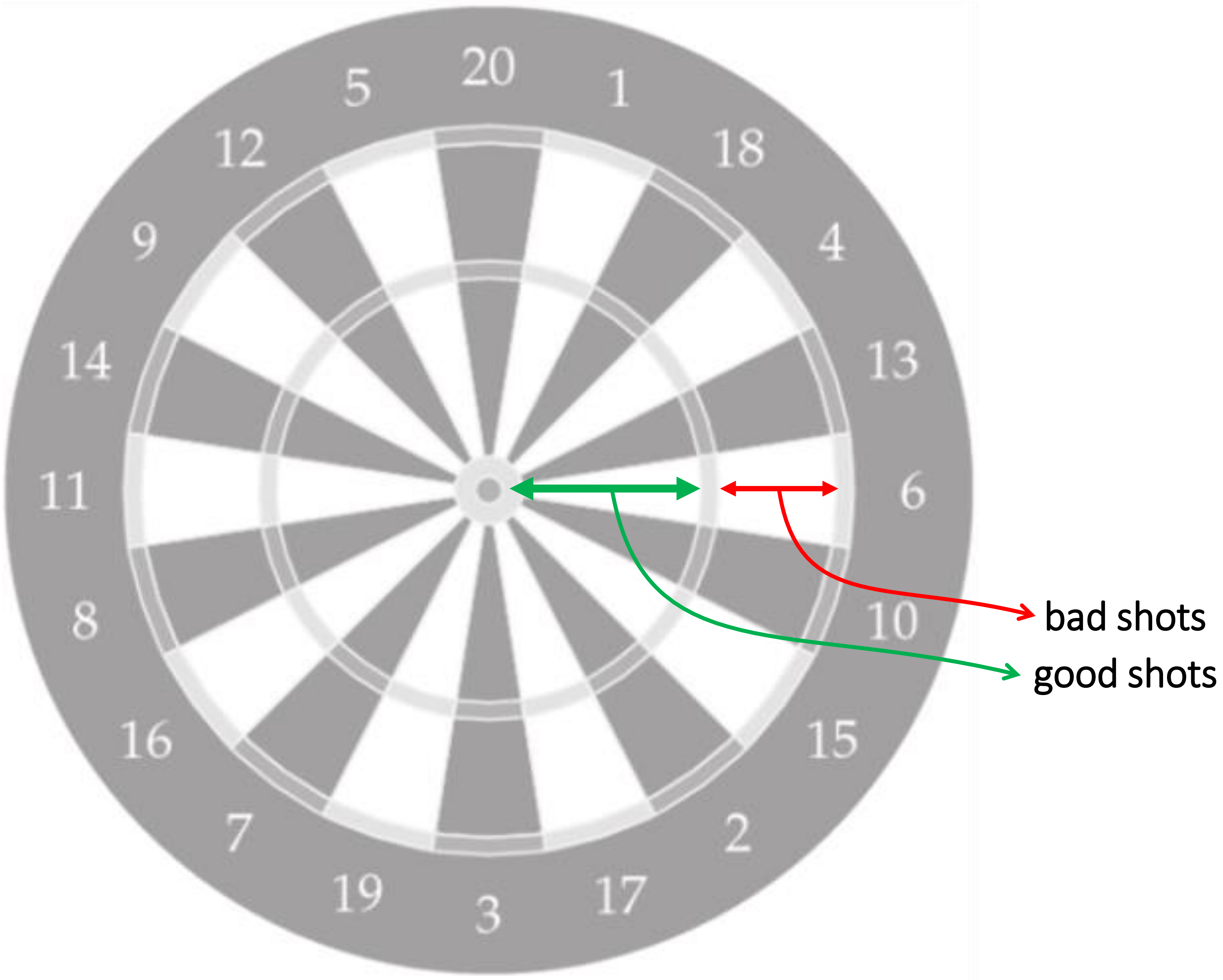
- **Maximum weight matching:** Only one cluster can match one partition
  - **Example:** If C_1 is more paired with T_2 THEN C_2 and C_3 cannot paired with T_2

$C\backslash T$	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

MAX

$$\left.\begin{array}{l} C_1 \text{ is more paired with } T_2, purity = \frac{30+5+25}{100} = 0.6 \\ C_1 \text{ is more paired with } T_3, purity = \frac{20+20+25}{100} = 0.65 \end{array}\right\} purity = 0.65$$

Precision, accuracy and recall



Precision, accuracy and recall

- Number of predicted "positive" labeled data = True Positive + False Positive
- Number of predicted "negative" labeled data = True Negative + False Negative

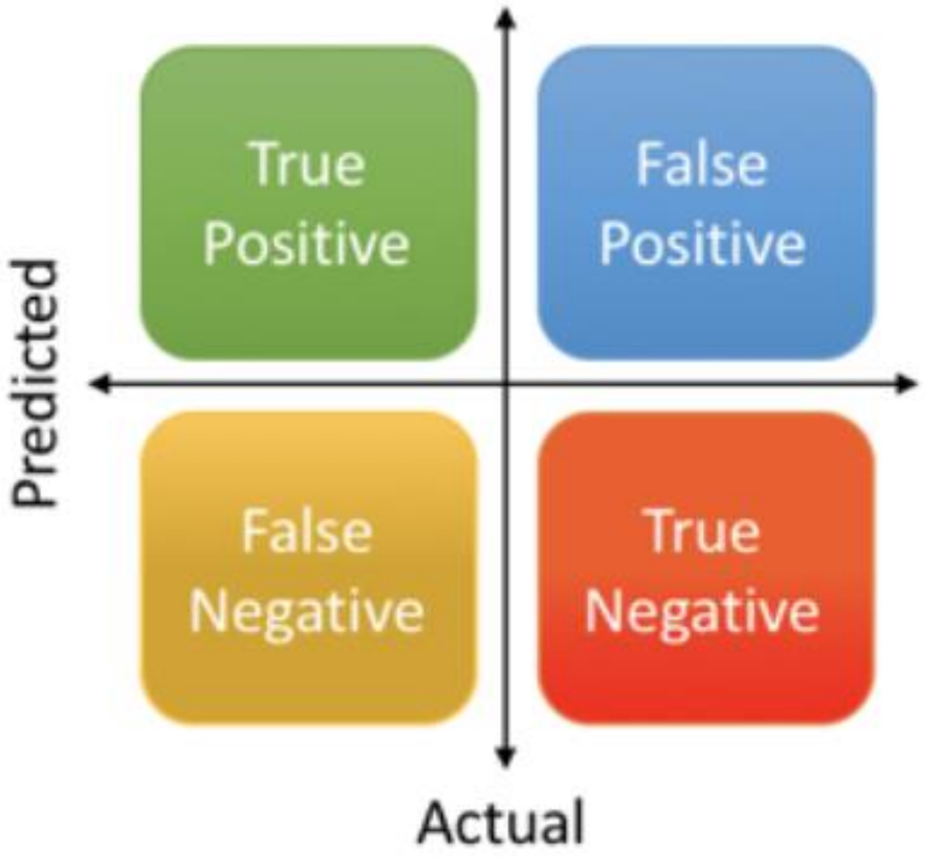
Correct prediction

Wrong prediction

$$\textbf{Precision} = \frac{True\ Positive}{Predicted\ Results} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\textbf{Recall} = \frac{True\ Positive}{Actual\ Results} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

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



False positive is also called false alarm

Matching-based measures: F-measure

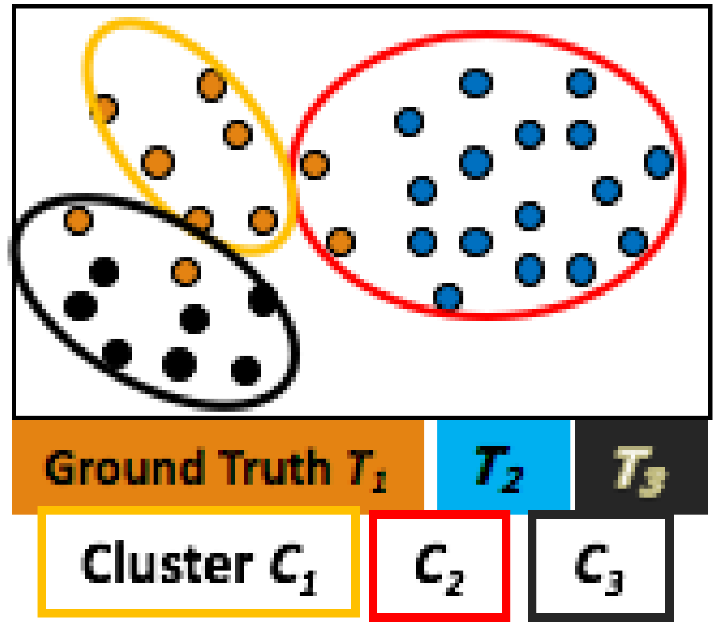
- **Precision:** which measures **quality**, is the same as purity:
  - How precisely does each cluster represent the ground truth?

$$precision_r = \frac{1}{n_r} \max_{k=1}^{K} \{n_{rk}\} = \frac{n_{rk_r}}{n_r}$$

- **Recall: measures completeness**
  - How completely does each cluster recover the ground truth?

$$recall_r = \frac{n_{rk_r}}{|T_{k_r}|} = \frac{n_{rk_r}}{m_{k_r}}$$

The fraction of point in partition T_k shared with cluster C_r



Example: $prec_1 = \frac{6}{6}$ $recall_1 = \frac{6}{10}$

Precision and recall

(Precision here is same as the purity)

Precision:

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

Recall:

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

$C\backslash T$	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

Matching-based measures: F-measure

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

$$F_r = \frac{2}{\frac{1}{prec_r} + \frac{1}{recall_r}} = \frac{2 \times prec_r \times recall_r}{prec_r + recall_r} = \frac{2 \times n_{rk_r}}{n_r + m_{k_r}}$$

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

$$F = \frac{1}{R}\sum_{r=1}^{R} F_r$$

$C \backslash T$	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

Example:

$$F_1 = \frac{2\times 30}{35+50} = \frac{60}{85} \qquad F_2 = \frac{2\times 20}{40+25} = \frac{40}{65} \qquad F_3 = \frac{2\times 25}{25+25} = 1$$

$$F = 0.774$$

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

Entropy-based measures: Conditional entropy

- Amount of information orderness in different partitions
- The entropy for clustering $\mathcal{C}$ and partition $\mathcal{T}$ is:

$$H(\mathcal{C}) = -\sum_{r=1}^{R} p_{C_r} \log_2 p_{C_r} \qquad H(\mathcal{T}) = -\sum_{k=1}^{K} p_{T_k} \log_2 p_{T_k}$$

where $p_{C_r} = \frac{n_r}{N}$ (n_r: row-wise summation, i.e. the probability of cluster C_r, $n_r = n_{r1} + \cdots + n_{rK}$) and $p_{T_k} = \frac{m_k}{N}$ (m_k: column-wise summation, i.e. the probability of cluster T_k)

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

$$H(\mathcal{T}|C_r) = -\sum_{k=1}^{K} \left(\frac{n_{rk}}{n_r}\right) \log \left(\frac{n_{rk}}{n_r}\right)$$

How ground truth is distributed within each cluster

n_{rk}

Cluster ($\boldsymbol{C}$)

Ground-truth ($\boldsymbol{T}$)

Entropy-based measures: Conditional entropy

- The conditional entropy of $\mathcal{T}$ given clustering $\mathcal{C}$ is defined as the weighted sum:

$$H(\mathcal{T}|\mathcal{C}) = \sum_{r=1}^{R} \frac{n_r}{N} H(\mathcal{T}|C_r) = -\sum_{r=1}^{R}\sum_{k=1}^{K} p_{rk} \log \frac{p_{rk}}{p_{C_r}}$$

$$p_{rk} \rightarrow \frac{n_{rk}}{n_r} \qquad p_{C_r} \rightarrow \frac{n_r}{N}$$

$$= H(\mathcal{C},\mathcal{T}) - H(\mathcal{C})$$

- The more clusters members are split into different partitions, the higher the conditional entropy (**not a desirable condition** and the max value is $\log_2 K$)
- $H(\mathcal{T}|\mathcal{C}) = 0$ 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 $H(\mathcal{T}|\mathcal{C}) = H(\mathcal{T})$.

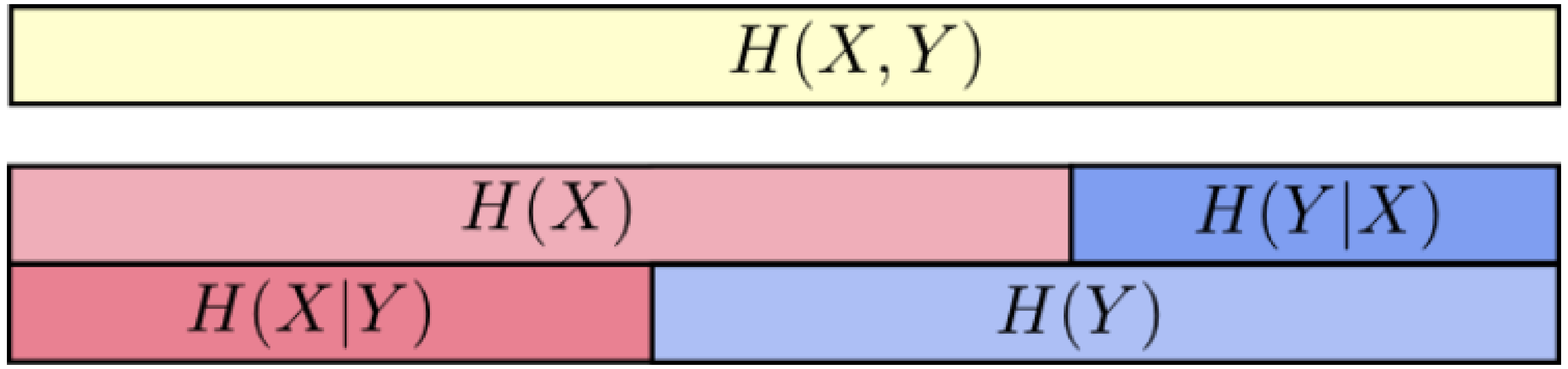
- Refresher: $H(Y|X) = \sum_{x\in X} p(x) H(Y|X=x)$, $H(Y|X) = H(X,Y) - H(X)$

Entropy-based measures: Conditional entropy

$$H(\mathcal{T}|\mathcal{C}) = -\sum_{r=1}^{R}\sum_{k=1}^{K} p_{rk}\log\frac{p_{rk}}{p_{C_r}} = -\sum_{r=1}^{R}\sum_{k=1}^{K} p_{rk}\left(\log p_{rk} - \log p_{C_r}\right)$$

$$= -\sum_{r=1}^{R}\sum_{k=1}^{K} p_{rk}(\log p_{rk}) + \sum_{r=1}^{R}(\log p_{C_r}\sum_{k=1}^{K} p_{rk}) =$$

$$-\sum_{r=1}^{R}\sum_{k=1}^{K} p_{rk}\log p_{rk} + \sum_{r=1}^{R}(p_{C_r}\log p_{C_r}) = H(\mathcal{T},\mathcal{C}) - H(\mathcal{C})$$



Entropy-based measures: example

- For each cluster:

$$H(\mathcal{T}|C_1) = -\left(\frac{0}{50}\right)\log_2\frac{0}{50} - \left(\frac{20}{50}\right)\log_2\frac{20}{50} - \left(\frac{30}{50}\right)\log_2\frac{30}{50} = 0.97$$

$$H(\mathcal{T}|C_2) = -\left(\frac{0}{25}\right)\log_2\frac{0}{25} - \left(\frac{20}{25}\right)\log_2\frac{20}{25} - \left(\frac{5}{25}\right)\log_2\frac{5}{25} = 0.72$$

$$H(\mathcal{T}|C_3) = -\left(\frac{25}{25}\right)\log_2\frac{25}{25} - \left(\frac{0}{25}\right)\log_2\frac{0}{25} - \left(\frac{0}{25}\right)\log_2\frac{0}{25} = 0.0$$

- Conditional entropy

$$H(\mathcal{T}|\mathcal{C}) = \frac{50}{100}\times 0.97 + \frac{25}{100}\times 0.72 + \frac{25}{100}\times 0.0 = 0.67$$

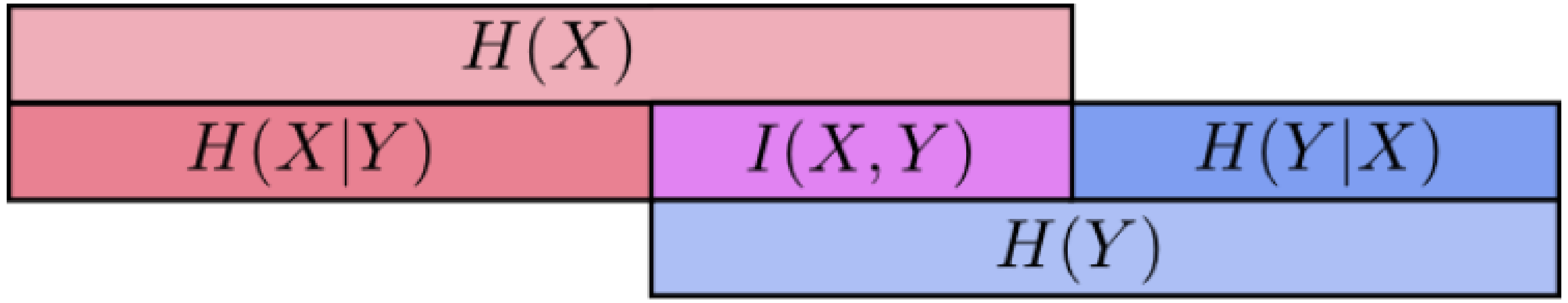
$C \backslash T$	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

Entropy-based measures: 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

$$I(\mathcal{C},\mathcal{T}) = \sum_{r=1}^{R}\sum_{k=1}^{K} p_{rk} \log\left(\frac{p_{rk}}{p_{C_r} \times p_{T_k}}\right) = H(\mathcal{T}) - H(\mathcal{T}|\mathcal{C})$$

- When $\mathcal{C}$ and $\mathcal{T}$ are independent then $p_{rk} = p_{C_k} \times p_{T_k}$, and thus $I(\mathcal{C},\mathcal{T}) = 0$. There is no upper bound on the mutual information.



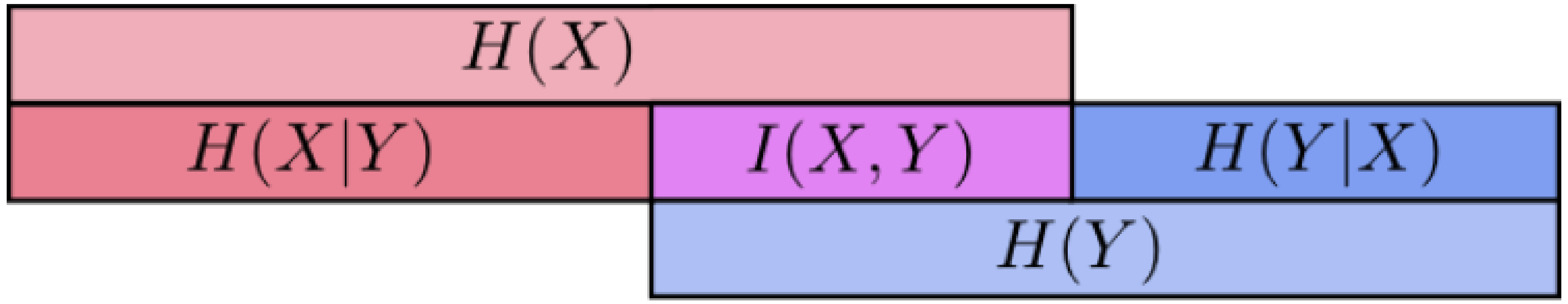
- We measure the dependency between the observed joint probability p_{rk} of C and T, and the expected joint probability $p_{C_r} \times p_{T_k}$ under the independence assumption

Entropy-based measures: Mutual information

- The **normalized mutual information** is defined as the geometric mean:

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

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



Entropy-based measures: example

- For clusters

$$H(\mathcal{C}) = -\left(\frac{50}{100}\right)\log_2\frac{50}{100} - \left(\frac{25}{100}\right)\log_2\frac{25}{100} - \left(\frac{25}{100}\right)\log_2\frac{25}{100} = 1.50$$

- For partitions

$$H(\mathcal{T}) = -\left(\frac{25}{100}\right)\log_2\frac{25}{100} - \left(\frac{40}{100}\right)\log_2\frac{40}{100} - \left(\frac{35}{100}\right)\log_2\frac{35}{100} = 1.56$$

- Mutual information

$$I(\mathcal{C},\mathcal{T}) = H(\mathcal{T}) - H(\mathcal{T}|\mathcal{C}) = 1.56 - 0.67 = 0.88$$

- Normalized mutual information

$$NMI(\mathcal{C},\mathcal{T}) = \frac{I(\mathcal{C},\mathcal{T})}{\sqrt{H(\mathcal{C}) \times H(\mathcal{T})}} = \frac{0.88}{\sqrt{1.5 \times 1.56}} = 0.57$$

$C\backslash T$	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

Outline

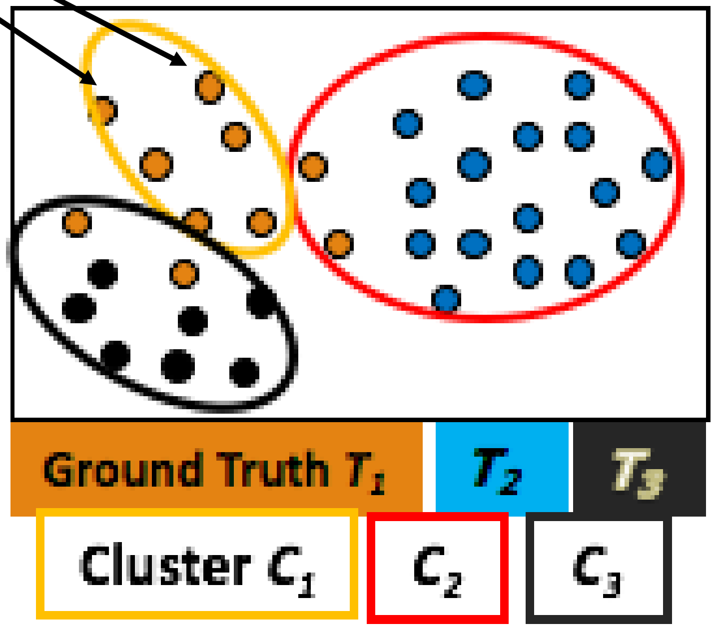
- External measures for clustering evaluation
  - Matching-based measures
  - Entropy-based measures
  - Pairwise measures

- Internal measures for clustering evaluation
  - Graph-based measures
  - Davies-Bouldin Index
  - Silhouette Coefficient

Pairwise measures

- Given clustering $\mathcal{C}$ and ground-truth partitioning $\mathcal{T}$, let $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$ 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 $\mathbf{x}_i$.
- **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

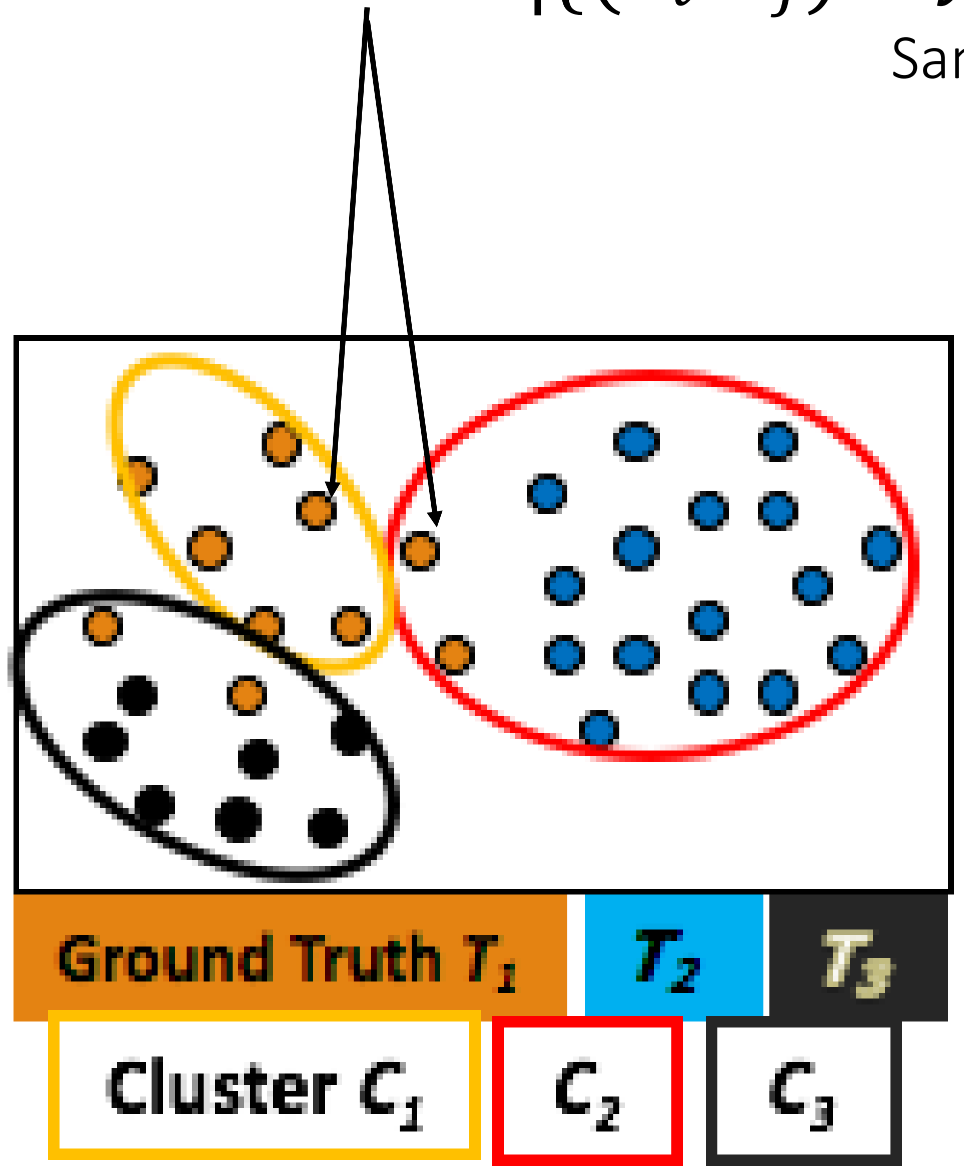
$$TP = \left|\left\{(\mathbf{x}_i, \mathbf{x}_j):\ \underset{\text{Same partition}}{y_i = y_j}\ \ \text{and}\ \ \underset{\text{Same cluster}}{\hat{y}_i = \hat{y}_j}\right\}\right|$$



Pairwise measures

- **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

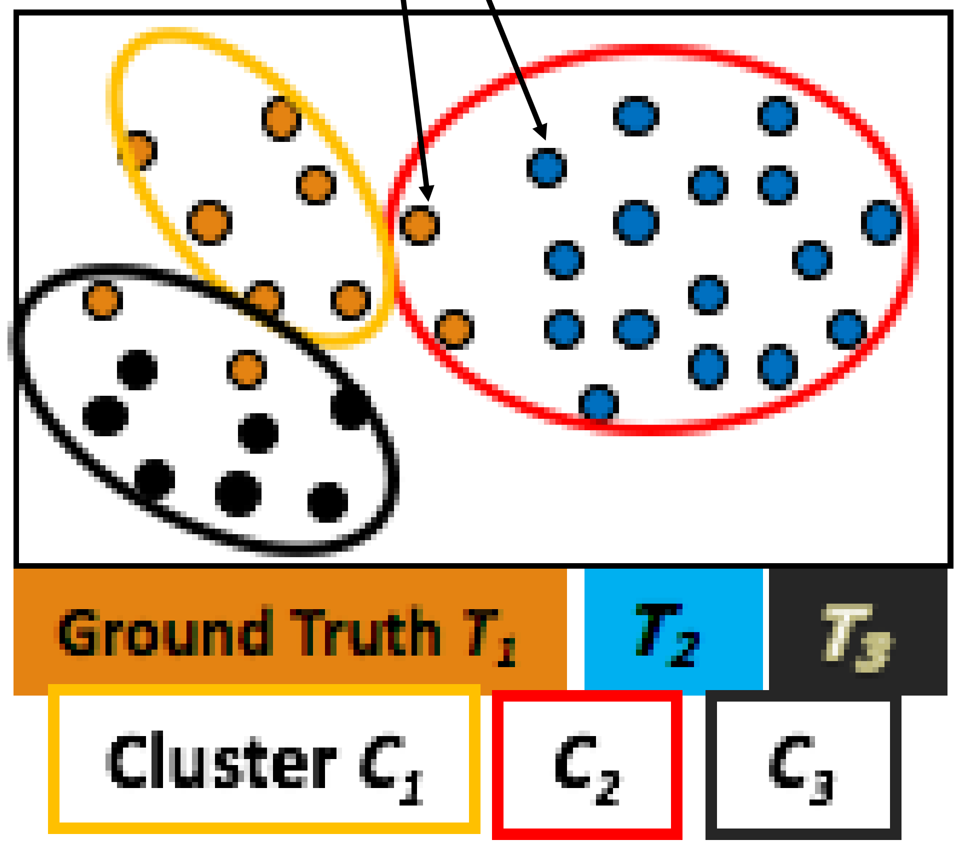
$$FN = \left|\left\{(\mathbf{x}_i, \mathbf{x}_j):\ \underbrace{y_i = y_j}_{\text{Same partition}}\ \text{ and }\ \underbrace{\hat{y}_i \neq \hat{y}_j}_{\text{Different cluster}}\right\}\right|$$



Pairwise measures

- **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 all false positive pairs is given as

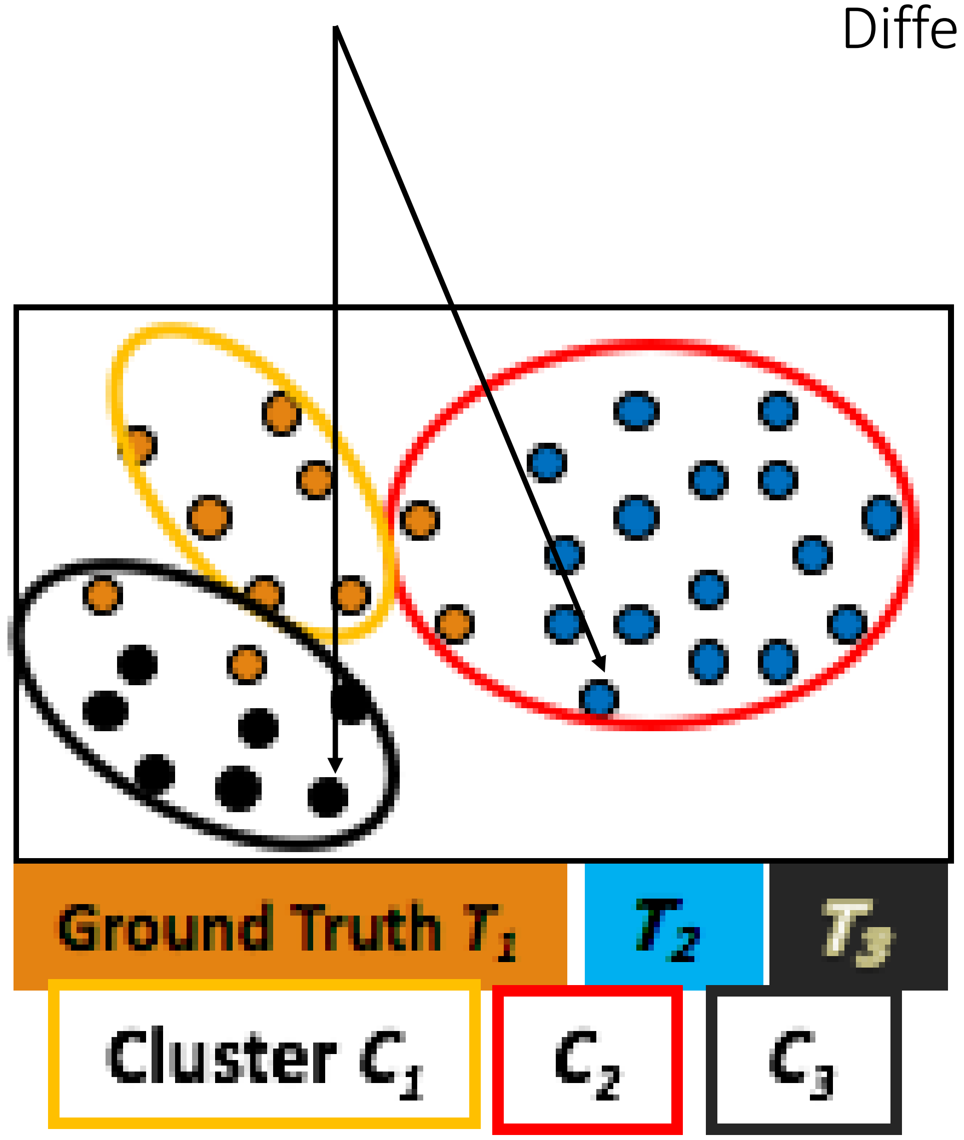
$$FP = \left|\left\{(\mathbf{x}_i, \mathbf{x}_j):\ \underbrace{y_i \neq y_j}_{\text{Different partition}}\ \text{and}\ \underbrace{\hat{y}_i = \hat{y}_j}_{\text{Same cluster}}\right\}\right|$$



Pairwise measures

- **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

$$FP = \left|\left\{\left(\mathbf{x}_i, \mathbf{x}_j\right): \quad \underbrace{y_i \neq y_j}_{\text{Different partition}} \quad \text{and} \quad \underbrace{\hat{y}_i \neq \hat{y}_j}_{\text{Different cluster}}\right\}\right|$$



Pairwise measures

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

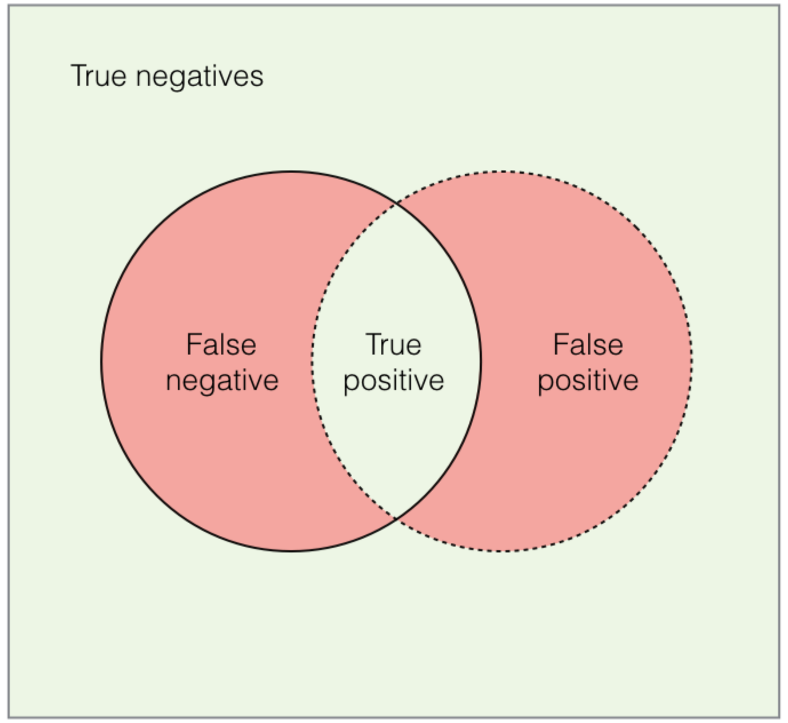
$$N = TP + FN + FP + TN$$

$$\text{TP} = \sum_{r=1}^{R}\sum_{k=1}^{K}\binom{n_{rk}}{2}$$

$$\text{FN} = \sum_{k=1}^{K}\binom{m_k}{2} - TP$$

$$\text{FP} = \sum_{r=1}^{R}\binom{n_r}{2} - TP$$

$$\text{TN} = N - (TP + FN + FP)$$



$C\backslash T$	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

$n_{12} = 20$ Points which have same C_1 and same T_2

Pairwise measures

- **Jaccard coefficient:** measures the fraction of true positive point pairs, but after ignoring the true negative:

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

Perfect clustering = 1

- **Rand statistic:** measures the fraction of true positives and true negatives over all point pairs:

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

Perfect clustering = 1 (like accuracy)

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

$$prec = \frac{TP}{TP+FP} \qquad recall = \frac{TP}{TP+FN}$$

The Fowlkes-Mallows (FM) measure is defined as the geometric mean of the pairwise precision and recall (higher value means a better clustering)

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

Pairwise measures

$$N = TP + FN + FP + TN = \frac{100(100-1)}{2} = 4{,}950$$

$$\mathrm{TP} = \sum_{r=1}^{R}\sum_{k=1}^{K}\binom{n_{rk}}{2} = \frac{20(20-1)}{2} + \frac{30(30-1)}{2} + \frac{20(20-1)}{2} + \frac{5(5-1)}{2} + \frac{25(25-1)}{2} = 1{,}125$$

$$\mathrm{FN} = \sum_{k=1}^{K}\binom{m_k}{2} - TP = \frac{25(25-1)}{2} + \frac{40(40-1)}{2} + \frac{35(35-1)}{2} - 1{,}125 = 550$$

$$\mathrm{FP} = \sum_{r=1}^{R}\binom{n_r}{2} - TP = \frac{50(50-1)}{2} + \frac{25(25-1)}{2} + \frac{25(25-1)}{2} - 1{,}125 = 700$$

$$\mathrm{TN} = N - (TP + FN + FP) = \frac{100(100-1)}{2} - (1{,}125 + 550 + 700) = 2{,}575$$

$C \backslash T$	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_k	25	40	35	100

Pairwise measures

- Jaccard coefficient:

$$Jaccard = \frac{TP}{TP + FN + FP} = \frac{1{,}125}{1{,}125 + 550 + 700} = 0.47$$

- Rand statistic:

$$Rand = \frac{TP + TN}{N} = \frac{550 + 2{,}575}{4{,}950} = 0.63$$

- Fowlkes-Mallows measure:

$$prec = \frac{TP}{TP+FP} = \frac{1{,}125}{1{,}125+700} = 0.616 \qquad recall = \frac{TP}{TP+FN} = \frac{1{,}125}{1{,}125+550} = 0.672$$

$$FM = \sqrt{prec \times recall} = \sqrt{0.616 \times 0.672} = 0.643$$

Outline

- External measures for clustering evaluation
  - Matching-based measures
  - Entropy-based measures
  - Pairwise measures

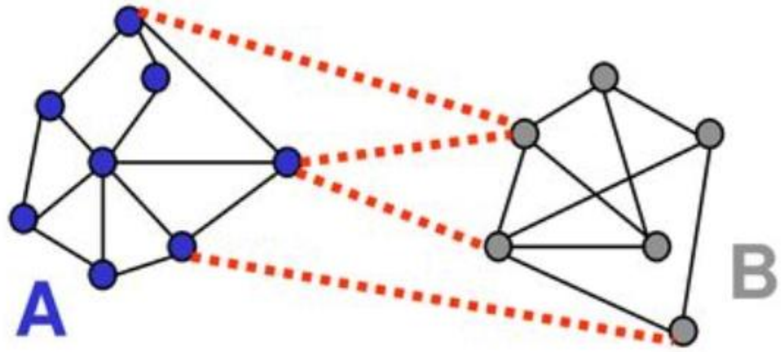
- Internal measures for clustering evaluation
  - Graph-based measures
  - Davies-Bouldin Index
  - Silhouette Coefficient

We want intra-cluster datapoints to be as close as possible to each other and inter-clusters to be as far as possible from each other

Beta-CV measure

- Let W be the pairwise distance matrix for all the given points. For any two point sets S and R, we define:

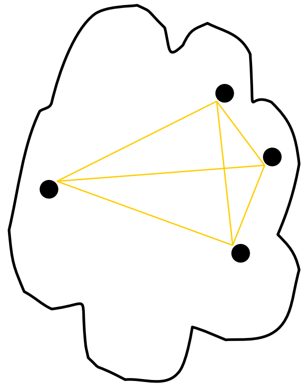
$$W(S,R) = \sum_{\mathbf{x}_i \in S} \sum_{\mathbf{x}_j \in R} w_{ij}$$



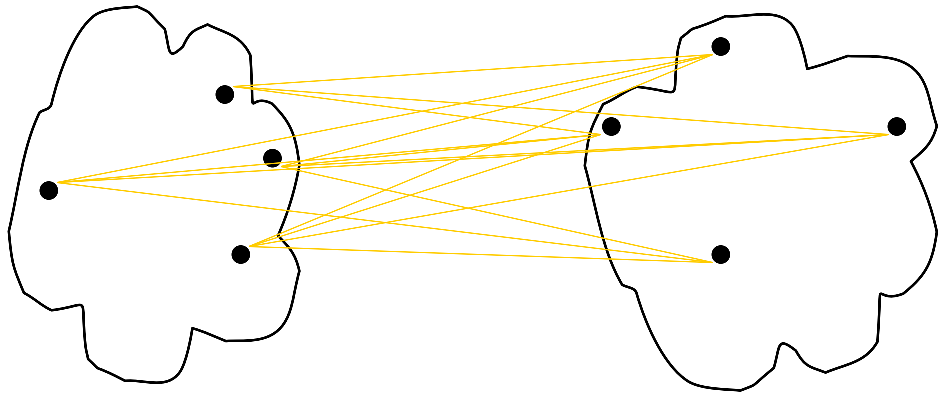
- **The sum of all the intracluster and intercluster weights are given as**

$$W_{in} = \frac{1}{2} \sum_{i=1}^{K} W(C_i, C_i) \qquad W_{out} = \frac{1}{2} \sum_{i=1}^{K} W(C_i, \overline{C_i}) = \sum_{i=1}^{K-1} \sum_{j>i} W\left(C_i, C_j\right)$$

The distance of each point is measured two times



cohesion



separation

Beta-CV measure

- **The number of distinct intracluster and intercluster edges is given as**

$$N_{in} = \sum_{i=1}^{K} \binom{n_i}{2} \qquad N_{out} = \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} n_i \times n_j$$

- **Beta-CV measure:** the Beta-CV measure is the ratio of the mean intracluster distance to the mean intercluster distance:

$$BetaCV = \frac{\frac{W_{in}}{N_{in}}}{\frac{W_{out}}{N_{out}}} = \frac{N_{out}}{N_{in}} \times \frac{W_{in}}{W_{out}} = \frac{N_{out}}{N_{in}} \frac{\sum_{i=1}^{K} W(C_i, C_i)}{\sum_{i=1}^{K} W(C_i, \overline{C}_i)}$$

The smaller the Beta-CV ratio, the better the clustering.

Normalized cut

- Normalized cut:

$$NC = \sum_{i=1}^{K} \frac{W(C_i, \overline{C_i})}{vol(C_i)} = \sum_{i=1}^{K} \frac{W(C_i, \overline{C_i})}{W(C_i, V)} = \sum_{i=1}^{K} \frac{W(C_i, \overline{C_i})}{W(C_i, \overline{C_i}) + W(C_i, C_i)} = \sum_{i=1}^{K} \frac{1}{\frac{W(C_i, C_i)}{W(C_i, \overline{C_i})} + 1}$$

where $vol(C_i) = W(C_i, V)$ is the volume of cluster C_i. The higher normalized cut value, the better the clustering

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

The Davies-Bouldin Index

- Let μ_i denote the cluster mean

$$\mu_i = \frac{1}{n_i} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$$

- Let σ_{μ_i} denote the dispersion or spread of the points around the cluster mean

$$\sigma_{\mu_i} = \sqrt{\frac{\sum_{\mathbf{x}_j \in C_i} \delta(\mathbf{x}_j, \mu_i)^2}{n_i}} = \sqrt{var(C_i)}$$

- The Davies-Bouldin measure for a pair of clusters C_i and C_j is defined as the ratio

$$DB_{ij} = \frac{\sigma_{\mu_i} + \sigma_{\mu_j}}{d(\mu_i, \mu_j)}$$

- DB_{ij} measures how compact the clusters are compared to the distance between the cluster means. The Davies-Bouldin index is then defined as:

$$DB = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \{DB_{ij}\}$$

A lower value means that the clustering is better.

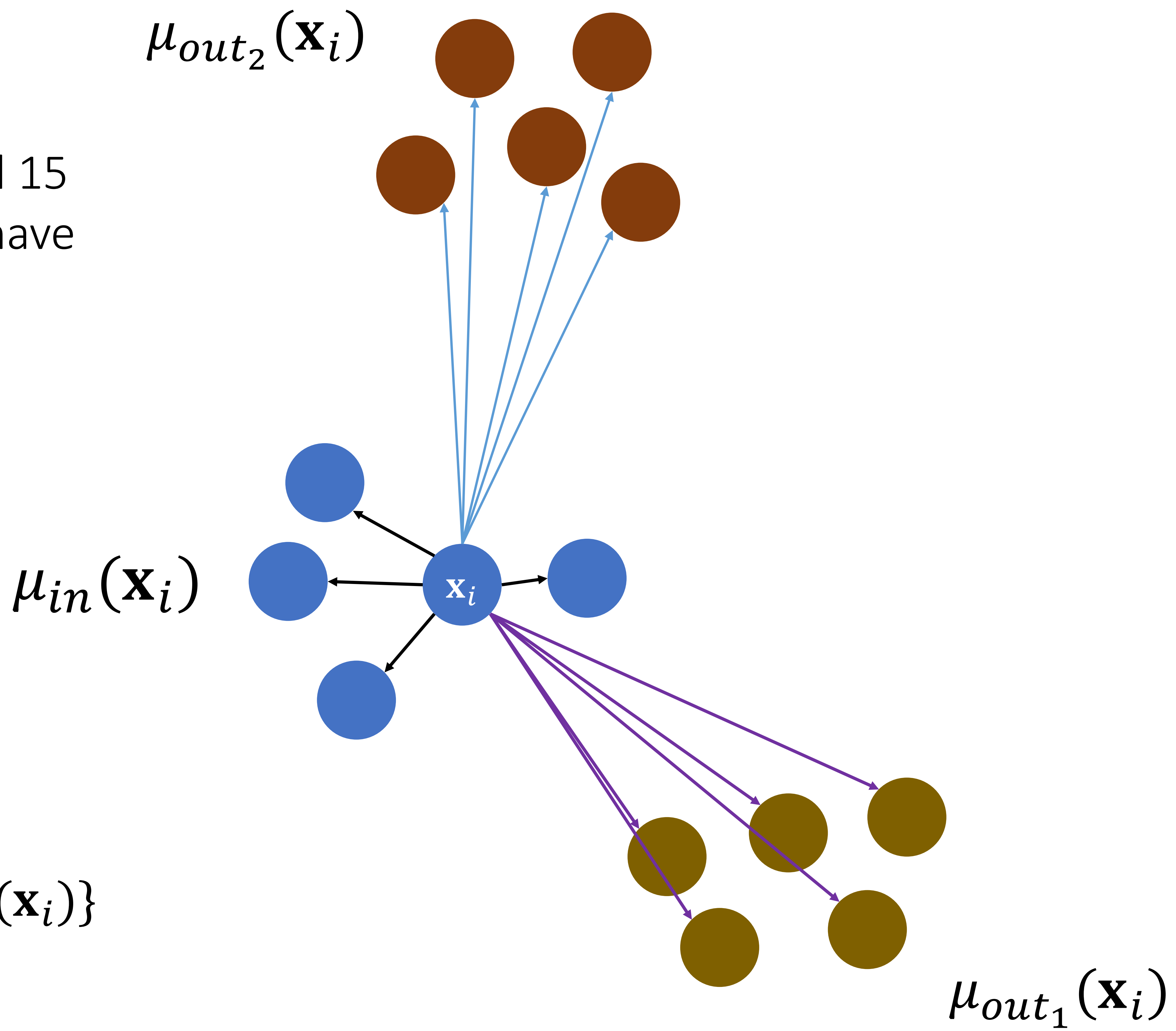
Outline

- 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

Silhouette coefficient

- Total of 15 mean distances μ_{in} and 15 mean distances μ_{out} because we have 15 datapoints



$$\mu_{out}^{min}(\mathbf{x}_n) = \min\{\mu_{out_2}(\mathbf{x}_i)\,, \mu_{out_1}(\mathbf{x}_i)\}$$

Silhouette coefficient

- **Define the silhouette coefficient of a point $\mathbf{x}_n$** as

$$S_i = \frac{\mu_{out}^{min}(\mathbf{x}_i) - \mu_{in}(\mathbf{x}_i)}{\max\{\mu_{out}^{min}(\mathbf{x}_i), \mu_{in}(\mathbf{x}_i)\}}$$

where $\mu_{in}(\mathbf{x}_i)$ is the mean distance from $\mathbf{x}_i$ to points in its own cluster $\hat{y}_i$:

$$\mu_{in}(\mathbf{x}_i) = \frac{\sum_{\mathbf{x}_i \in C_{\hat{y}_i}, j \neq i} d(\mathbf{x}_i, \mathbf{x}_j)}{n_{\hat{y}_i} - 1}$$

and $\mu_{out}^{min}(\mathbf{x}_i)$ is the mean of the distances from $\mathbf{x}_i$ to points in the closest cluster:

$$\mu_{out}^{min}(\mathbf{x}_i) = \min_{j \neq \hat{y}_i} \left\{ \frac{\sum_{\mathbf{x} \in C_j} d(\mathbf{x}_i, \mathbf{x}_j)}{n_j} \right\}$$

- The Silhouette Coefficient for clustering C:

$$SC = \frac{1}{N} \sum_{i=1}^{N} Si$$

- **SC close to 1 implies a good clustering** (points are close to their own clusters but far from other clusters)