Clustering Validation

Davide Mottin

Data Mining

Clustering evaluation and validation

- **External measures:**
  - Take criteria into account that are not part of the clustering data
    - E.g. class labels
- **Internal measures:**
  - Based only on clustering data
    - E.g. distances as in TD or silhouette coefficient
- **Measures can be relative**
  - Compare two clusterings instead of obtaining ”objective” goodness value
  - Some measures, e.g. silhouette can be used in both ways

External measures (using ground-truth clustering)

Correct or ground-truth clustering is known *a priori*

- $y_i \in \{1,2,\dots,k\}$ ground-truth cluster membership
  - $\mathcal{T} = \{T_1, T_2, \dots, T_k\}$ ground-truth clustering
  - $T_j = \{\mathbf{x_i} \in \boldsymbol{D} | y_i = j\}$
- $C = \{C_1, \dots, C_r\}$ a clustering in r clusters
  - $\hat{y}_i \in \{1,2,\dots,r\}$ cluster label for $\mathbf{x_i}$

Contingency table

- $r \times k$ table $\mathbf{N}$ induced by clustering $\mathcal{C}$ and ground truth $\mathcal{T}$

$$\mathbf{N}(i,j) = \mathrm{n_{ij}} = |\mathrm{C_i} \cap T_j|$$

- n_{ij} : number of points **common** to cluster C_i and ground-truth T_j
- $n_i = |C_i|$ number of points in cluster C_i
- $m_j = |T_j|$ number of points in ground-truth T_j
- Computed in $O(n)$ time
  - Examine the ground-truth and cluster labels $y_i, \hat{y}_i$ for each point $x_i \in D$
  - Increment the count $n_{y_i,\hat{y}_i}$

Purity

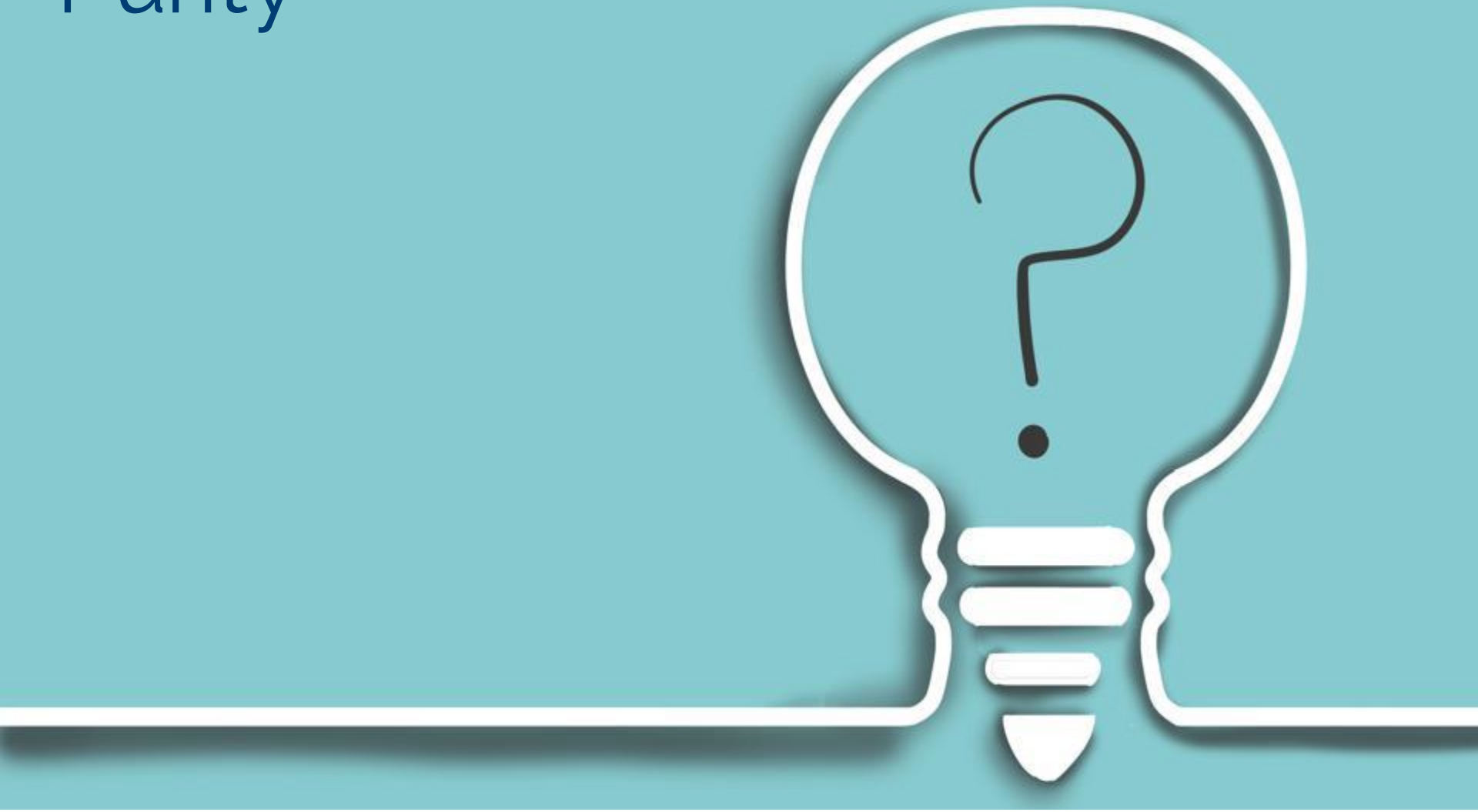
- Quantifies the extent to which a cluster C_i contains entities from **only** one partition

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

- **Purity of clustering $\mathcal{C}$**: weighted sum of each cluster's purity

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

Purity



Go to **www.menti.com** and use the code **84 92 61 9**

Maximum matching

Select mapping between clusters and ground-truth, such that the sum of the number of common points is maximized

Computation

- Create a **weighted bipartite graph**
  - $V = \mathcal{C} \cup \mathcal{T}$ vertex set
  - $E = \{(C_i, T_j\}\}$ edge set
  - $w(C_i, T_j) = n_{ij}$
- Matching M: subset of pairwise nonadjacent edges from E (i.e., no common vertex)
- Maximum weight matching
  - Computed , e.g. .with Hungarian algorithm for assignment problem

$$match = \arg\max_{M} \left\{ \frac{\sum_{e \in M} w(e)}{n} \right\}$$

F-measure

- j_i ground-truth cluster with max number of points from C_i, $j_i = \max_{j=1}^{k}\{n_{ij}\}$
- **Precision**
  - Same as purity

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

- **Recall**

$$recall_i = \frac{n_{ij_i}}{|T_{j_i}|} = \frac{n_{ij_i}}{m_{j_i}}$$

  - $m_{j_i} = |\mathrm{T_{j_i}}|$
- **F-measure**
  - Harmonic mean of precision and recall for each cluster

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

- F-measure of clustering $\mathcal{C}$ is the mean clusterwise F-measure:

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

Conditional entropy

- Entropy of a clustering

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

Where $p_{C_i} = \frac{n_i}{n}$ probability of cluster C_i

- Entropy of ground-truth $H(\mathcal{T})$ is similarly defined
- Cluster specific entropy of $\mathcal{T}$

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

- Conditional entropy of clustering

$$H(\mathcal{T}|\mathcal{C}) = -\sum_{i=1}^{r}\sum_{j=1}^{k} \frac{n_{ij}}{n} \log \frac{n_{ij}}{n_i} = -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij} \log \frac{p_{ij}}{p_{C_i}}$$

  - Perfect clustering $H(\mathcal{T}|\mathcal{C}) = 0$
  - Worst entropy $\log k$

Mutual-Information

Quantify the amount of shared information between the clustering $\mathcal{C}$ and the ground-truth $\mathcal{T}$

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

- Measures the dependence between the observed joint probability p_{ij} of the clustering and the gropund-truth, and the expected joing probability $p_{C_i} p_{T_j}$ **under independence assumption**
  - If C and T are independent → $p_{C_i} p_{T_j} = \mathrm{p_{ij}}$ → $I(\mathcal{C},\mathcal{T}) = 0$
- No upper bound on $I(\mathcal{C},\mathcal{T})$!!!

Normalized version

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

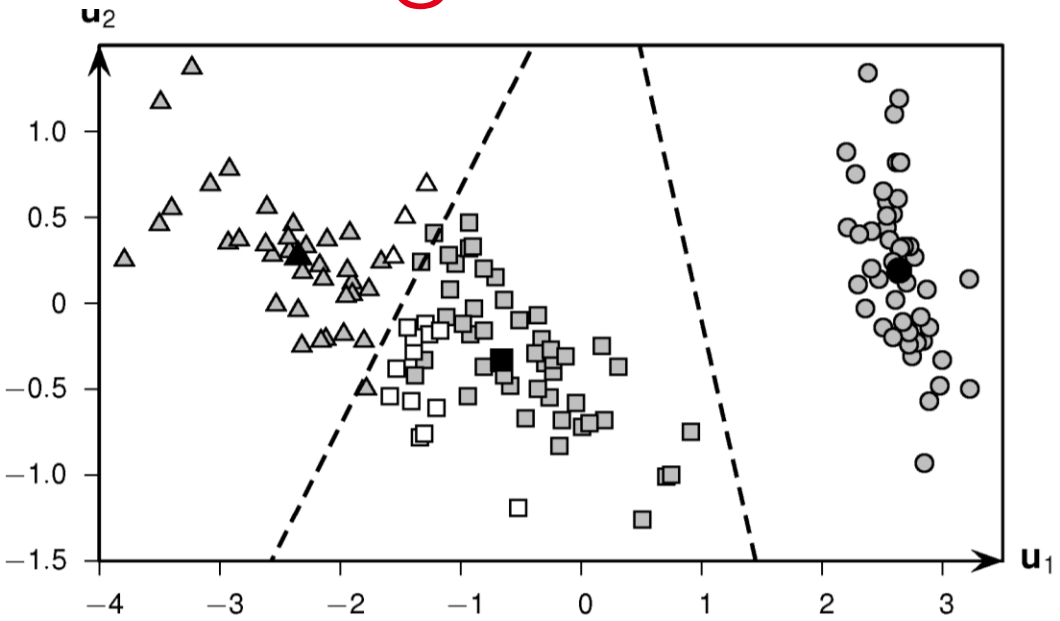
- NMI is in the range [0,1] with 1 indicate a good clustering!

Iris prinpiple component
Good clustering

Purity 0.887

Match 0.887

F1 0.85



Contingency table:

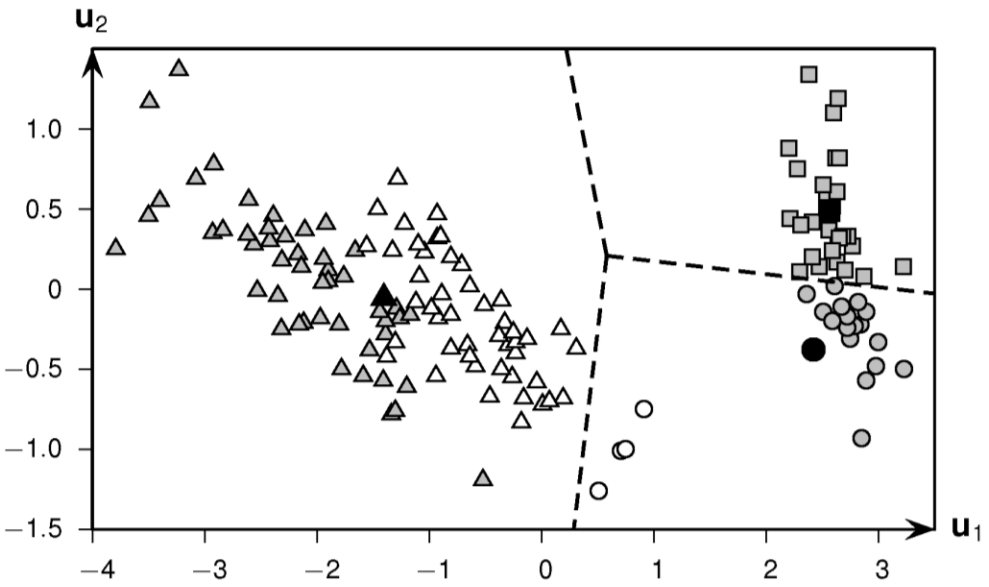
	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1 (squares)	0	47	14	61
C_2 (circles)	50	0	0	50
C_3 (triangles)	0	3	36	39
m_j	50	50	50	$n = 100$

Iris principle component data
Poor clustering

Purity 0.667

Match 0.560

F 0.658



Contingency table:

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1 (*squares*)	30	0	0	30
C_2 (*circles*)	20	4	0	24
C_3 (*triangles*)	0	46	50	96
m_j	50	50	50	$n = 150$

Evaluation measures - Discussion

- **Internal measures:**
  - Make no assumptions about ground truth
  - In unsupervised learning this is very often the case in practice
- **External measures:**
  - Make use of external information such as domain expertise
  - Sometimes to be generated for testing sample

Different measures capture different aspects of a good clustering

- This is actively discussed in the research community
- For example, can a cluster be “matched” to two ground truth clusters (or the other way around? Should this be punished somehow?