Clustering Validation

- No ground-truth data
- Internal measures
- Intra-cluster: pairwise internal distance
- Inter-cluster: across clusters

- External measures
- Hide the true class labels
- Use random class attribute for clustering

Algorithm $\rightarrow$ $S = \{ C_1, C_2, \ldots, C_Y \}$ clustering

Class labels $\rightarrow$ $T = \{ T_1, T_2, \ldots, T_k \}$ partitioning/ground-truth clustering

$V = k$

$N_j = \#j$ points in common between $C_i$ & $T_j$

$C_i \cap T_j$
1:1 matching

Permutation

Purity

\[
\text{Purity}_i = \frac{1}{m^2} \max_j \{ n_{ij} \}
\]

\[
\text{Purity} = \sum_{i=1}^{n} \frac{1}{n} \text{Purity}_i
\]

\[
= \frac{1}{n} \sum_{i} \max_j \{ n_{ij} \}
\]

\[
= \frac{1}{n} \sum_{i} \max_j \{ n_{ij} \}
\]

Maximum matching

\[
\gamma = k
\]

Hungarian also

\[O(k^3)\]
Precision vs Recall

\[ P_i = \frac{1}{n_i} \sum_j n_{ij} \]
\[ \text{Recall}_i = \frac{1}{m_i} \sum_j n_{ij} \]

For cluster \( C_i \), find the max intersection \( j^* \)

\[ f_i = \text{harmonic mean of} \ P_i \ \text{& recall}_i \]

**Info-theoretic**

Conditional entropy \( T \) given \( C_i \)

\[ H(T|C_i) = -\frac{1}{n} \sum_j \frac{n_{ij}}{n_i} \log \left( \frac{n_{ij}}{n_i} \right) \]

\[ H(T) = -\frac{1}{n} \sum_j n_j \log n_j \]

\[ H(C) = -\frac{1}{n} \sum_j n_j \log n_j \]

Over all clusters
Over all clusters

\[ H(T | X) = \sum_{i} \frac{n_i}{n} \cdot H(T | C_i) \]

**Normalized Mutual Information (NMI)**

\[
I(C, T) = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{P_{ij} \log \left( \frac{P_{ij}}{P_i \cdot P_j} \right)}
\]

\[
P_{ij} = \frac{n_{ij}}{n} \quad P_i = \frac{n_i}{n} \quad P_j = \frac{n_j}{n}
\]

**NMI** = geometric mean

\[
NMI = \sqrt{\frac{I(C, T)}{H(C) \cdot H(T)}} = \frac{I(C, T)}{\sqrt{H(C) \cdot H(T)}}
\]

Higher the better \( \in [0, 1] \)

**Pair-wise**

\[ \mathcal{C} = \{ C_1, C_2, \ldots, C_r \} \]

\[ \mathcal{T} = \{ T_1, T_2, \ldots, T_k \} \]

\[ O(n^2) \quad X_i \rightarrow X_j \]

\[ TP: \text{ if } C_i = C_j \text{ and } T_j \rightarrow T_i \]
Confusion matrix

\[ \begin{array}{c|c|c|c|c} & P & F & N & \text{Total} \\ \hline T & T_P & F_P & F_N & T_N \\ \hline \end{array} \]

\[ \text{Jaccard coefficient} = \frac{T_P}{T_P + F_N + F_P} \]

\[ \text{FM-measure: geometric mean} \]
\[ \text{Precision} = \frac{T_P}{T_P + F_P} \]
\[ \text{Recall} = \frac{T_P}{T_P + F_N} \]

\[ \sqrt{\frac{T_P T_P}{(T_P + F_P)(T_P + F_N)}} \]

Correlation (corr):

\[ X = \frac{1}{n} \sum_{i=1}^{n} x_{i} \]

\[ x_{i,j} = 1 \text{ if } c_i = c_j \]

\[ y_{i,j} = 1 \text{ if } t_i = t_j \]

\[ y_{i,j} = 0 \text{ otherwise} \]

\[ n \times n \text{ symmetric binary matrix} \]

Pearson correlation between 2 factors
**Internal Measure**

\[ G = \{ c_1, c_2, \ldots, c_k \} \]

\[ W = \begin{bmatrix} w_{ij} \end{bmatrix} \quad \text{distance/proximity matrix} \quad O(n^2) \]

\[ w_{ij} = \text{distance between } x_i \text{ and } x_j \]

\[ w_{ij} = \| x_i - x_j \|_2 \quad k = \text{mean } / \varepsilon_m \]

\[ w_{ij} = \text{length of the shortest path between } x_i \text{ and } x_j \]

\[ W(S,T) = \text{sum of the weights on edges between } S \text{ and } T \]

\[ W_{in} = \text{sum of the weights that are "intra"} \]

\[ W_{in} = \frac{1}{2} \sum_{i=1}^{k} W(c_i, c_i) \]

\[ W_{av} = \frac{1}{2} \sum_{i=1}^{k} W(c_i, \overline{c_i}) \]

\[ N_{in} = \# \text{ of internal edges} \]

\[ \sum_{i=1}^{k} \binom{n_i}{2} = \sum_{i=1}^{k} \frac{n_i(n_i-1)}{2} \]

\[ N_{av} = \# \text{ of external edges} \]

\[ N - N_{in} = W_{av} \]

\[ \# \text{ of total pairs of points} \]

\[ N = \frac{n(n-1)}{2} \]
Belz CV: fraction of intent to extend weight

\[ \text{Belz CV} = \frac{\frac{\text{Win}}{\text{Nin}}}{\frac{\text{Wout}}{\text{Nout}}} \]

Select the smaller Nin edges

\[ \zeta = \{ c_1, c_2, \ldots, c_k \} \]

Smaller is better

\[ \text{Win vs. } \min_{\text{Nin}} \]

Independent

Smaller is better \([0, 1]\)

\[ \text{Cond} = \frac{\text{Win} - \text{Wmin}}{\text{Wmax} - \text{Wmin}} \]

Smaller is better

\[ \sum_{i=1}^{k} \frac{\text{W}(c_i, \bar{c}_i)}{\text{W}(c_i, v)} \]

Larger is better

\[ \gamma_1 = \{ c_1, c_2, \ldots, c_k \} \]

\[ \gamma_2 = \{ c'_1, c'_2, \ldots, c'_k \} \]

Compare two clusterings
Modularity

$$\sum_{i=1}^{k} \left( \frac{w(c_i, c_j)}{w(v, v)} - \left( \frac{w(c_i, v)}{w(v, v)} \right)^2 \right)$$

expected prob internal edge

under independence assumption

larger is better

SC: Silhouette Coefficient

$$s_i = SC \text{ for } x_i^k$$

for point $$x_i$$.

$$\operatorname{Min}(x_i) = \text{ avg distance to points in its own cluster}$$

$$\operatorname{Min}(x_i) = \text{ avg internal distance from } x_i$$

$$\operatorname{Max}(x_i) = \wedge \text{ avg external distance}$$

$$s_i = \frac{\operatorname{Max}(x_i) - \operatorname{Min}(x_i)}{\max \{ \operatorname{Max}(x_i), \operatorname{Min}(x_i) \}}$$

$$s_i \in [-1, 1]$$

+1: $$x_i$$ is very close internally & very far externally

-1: $$x_i$$ is close to some other cluster

$$SC = \frac{1}{n} \sum_{i=1}^{n} s_i$$

How many clusters are there? $$k$$
Choose $k$ that has the largest gap! 

$\text{log}_k(\text{Win}_k(D))$

$\text{log}_j(\text{Win}_k(D))$

$\text{gap} = \mu(k) - \text{Win}_k(D)$

$\mu(k) = \underset{D_i}{\text{arg min}} \mu(k)$

$\sigma(k) = \text{std}$

$k$ should not be too large