Outline

- Partitional Methods
  - K-Means
  - Spectral Clustering
  - EM-based Gaussian Mixture Decomposition

Part 3.: Validation of clustering solutions
- Cluster Validity Measures

Part 4.: Ensemble Methods
- Evidence Accumulation Clustering

Different Aspects of Cluster Validation

- Determining the clustering tendency of a set of data, i.e., distinguishing whether non-random structure actually exists in the data.
- Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels – Performance evaluation
- Evaluating how well the results of a cluster analysis fit the data without reference to external information - Use only the data
- Comparing the results of two different sets of cluster analyses to determine which is better - Select a good algorithm for a data set.
- Select values of parameters of a clustering algorithm
  - Determining the ‘correct’ number of clusters.
- Select a good distance measure on the data
  - Feature selection and feature extraction are special cases of this
Cluster Validity Measures

Different Algorithms are Suitable for Different Data Sets

Spectral Clustering, $k=5$, $\sigma=.1$

Spectral Clustering, $k=5$, $\sigma=.05$

Hierarchical agglomerative, based on dissimilarity increments: $\alpha=3$

Single-link, $k=24$

Different Algorithms are Suitable for Different Data Sets
Cluster Validity Measures

Types:
- **External index**: compare with the “true” partition (known labels)
  - Entropy
  - Drawback: ground truth usually not available
- **Internal index**: compare the partition with intrinsic properties of the data set (without respect to external information).
  - Sum of Squared Error (SSE)
- **Relative index**: evaluate quality of a partition by comparing it to other clustering schemes, resulting by the same algorithm but with different parameter values.
  - Often an external or internal index is used for this function, e.g., SSE or entropy
  - Drawback: subjective, sometimes algorithm-dependent
- **Stability-based index**

Sometimes these are referred to as criteria instead of indices

Cluster Validity Measures: External Indices

Type of criterion by “principle”:
- Criteria based on counting pairs
  - Rand, Fowlkes-Mallows
- Criteria based on cluster matching
  - “classification error” (Ci), Larsen
- Other criteria
  - Based on mutual information

Type of criterion by mathematical properties
- Metric (distance) \(=0\) for identical clusterings
- “index” \(=1\) for identical clusterings
Cluster Validity Measures: External Indices

Criteria based on counting pairs

- For two partitions P1 and P2
  - N11: number of pairs of data points that are in the same cluster in both P1 and P2
  - N01: number of pairs of data points that are in the same cluster in P2 but not in P1
  - N10: number of pairs of data points that are in the same cluster in P1 but not in P2
  - N00: number of pairs of data points that are in different clusters in both P1 and P2

- m1: number of pairs of data points in the same cluster in P1
- m2: number of pairs of data points in the same cluster in P2

\[ N00 + N01 + N10 + N11 = n(n-1)/2 = M \]

### Criteria Based on Counting Pairs

- Rand = \( \frac{N_{00} + N_{11}}{N_{00} + N_{01} + N_{10} + N_{11}} = \frac{N_{00} + N_{11}}{M} \)
- Jaccard = \( \frac{N_{01} + N_{10} + N_{11}}{N_{11}} \)
- F & M = \( \frac{N_{11}}{\sqrt{m_1 m_2}} \) (Fowlkes-Mallows)
- Hubert = \( \frac{MN_{11} - m_1 m_2}{\sqrt{m_1 m_2 (M - m_1)(M - m_2)}} \)

\[ M = n(n-1)/2 \]
Cluster Validity Measures: External Indices

Criteria based on cluster matching

- Larsen
  - Asymmetric

$$\mathcal{L}(P, P') = \frac{1}{K} \sum_{k=1}^{K} \max_{k'} \frac{2n_{kk'}}{n_k + n_{k'}}$$

- "Classification error", Consistency Index
  - (Meila – Heckerman, Fred)

$$\mathcal{H}(P, P') = C_i(P, P') = \frac{1}{n} \sum_{k = \text{match}(k)} n_{kk'}$$

- Van Dongen
  - Is a metric

$$\mathcal{D}(P, P') = 2n - \sum_{k} \max_{k'} n_{kk'} \max_{k'} n_{kk'}$$

Criteria based on mutual information

- Mutual Information
  - For any random variable

$$H(X) = - \sum_{x=1}^{K} P_X(x) \log P_X(x)$$

$$I(X, Y) = \sum_{x=1}^{K} \sum_{y=1}^{K'} P_{XY}(x, y) \log \frac{P_{XY}(x, y)}{P_X(x) P_Y(y)}$$

- Entropy = 0 iff RV deterministic
- Mutual information = 0 iff RV’s independent

- For two clusterings
  - $H(P^o), H(P^b) = $ the entropy of the associated RV’s
  - $I(P^o, P^b) = $ the mutual information of the associated RV’s
Cluster Validity Measures: External Indices

- A partition $P^a$ describes a labeling of the patterns into $k_a$ clusters.

- Taking frequency counts as approximations for probabilities, the entropy of the data partition $P^a$ is expressed by
  \[
  H(P^a) = -\sum_{i=1}^{k_a} \frac{n^a_i}{n} \log \left( \frac{n^a_i}{n} \right) \quad n^a_i = \text{#patterns in cluster } C^a_i \in P^a
  \]

- The agreement between two data partitions $P^a$ and $P^b$ is measured by the Mutual Information
  \[
  I(P^a, P^b) = \sum_{i=1}^{k_a} \sum_{j=1}^{k_b} \frac{n_{ij}^{ab}}{n} \log \left( \frac{n_{ij}^{ab}}{n_i^a n_j^b} \right) \quad n_{ij}^{ab} = \text{#shared patterns between clusters } C^a_i \in P^a \text{ and } C^b_j \in P^b
  \]

- Normalized Mutual Information
  \[
  NMI(P^a, P^b) = \frac{-2 \sum_{i=1}^{k_a} \sum_{j=1}^{k_b} \frac{n_{ij}^{ab}}{n} \log \left( \frac{n_{ij}^{ab}}{n_i^a n_j^b} \right)}{ \sum_{i=1}^{k_a} n_i^a \log \left( \frac{n_i^a}{n} \right) + \sum_{j=1}^{k_b} n_j^b \log \left( \frac{n_j^b}{n} \right) } = \frac{2I(P^a, P^b)}{H(P^a) + H(P^b)}
  \]

Since

\[
I(P^a, P^b) \leq \frac{H(P^a) + H(P^b)}{2}
\]

\[
0 \leq NMI(P^a, P^b) \leq 1
\]
Cluster Validity Measures: External Indices

- **Variation of Information (Meila)**
  - Measures the sum of information lost + gained between the two clusterings
  
  \[
  VI(P^a, P^b) = H(P^a) + H(P^b) - 2I(P^a, P^b)
  \]
  
  \[
  VI(P^a, P^b) = H(P^a|P^b) + H(P^b|P^a)
  \]

Cluster Validity Measures: Internal Indices

- **Measuring Cluster Validity Via Correlation**
  - Two matrices
    - Proximity Matrix
    - “Incidence” Matrix
      - One row and one column for each data point
      - An entry is 1 if the associated pair of points belong to the same cluster
      - An entry is 0 if the associated pair of points belongs to different clusters
  - Compute the correlation between the two matrices
    - Since the matrices are symmetric, only the correlation between n(n-1)/2 entries needs to be calculated
  - High correlation indicates that points that belong to the same cluster are close to each other.
  - Not a good measure for some density or contiguity based clusters.
Cluster Validity Measures: Internal Indices

Measuring Cluster Validity Via Correlation

- Correlation of incidence and proximity matrices for the K-means clusterings of the following two data sets.

![Correlation plots with Corr = -0.9235 and Corr = -0.5810.]

Using Similarity Matrix for Cluster Validation

- Order the similarity matrix with respect to cluster labels and inspect visually.
Cluster Validity Measures: Internal Indices - Cohesion and Separation

- Cluster Cohesion: Measures how closely related are objects in a cluster
  - Example: SSE
- Cluster Separation: Measure how distinct or well separated a cluster is from other clusters

Example: Squared Error
- Cohesion is measured by the within cluster sum of squares (SSE)
  \[ WSS = SSE = \sum_{i=1}^{k} \sum_{x \in C_i} (x - m_i)^2 \]
- Separation is measured by the between cluster sum of squares
  \[ BSS = \sum_{i=1}^{k} |C_i| (m_i - m)^2 \]
  where \(|C_i|\) is the size of cluster \(i\)

Cluster Validity Measures: Internal Indices - Silhouette Coefficient

- Silhouette Coefficient combine ideas of both cohesion and separation, but for individual points, as well as clusters and clusterings
- For an individual point, \(i\)
  - Calculate \(a\) = average distance of \(i\) to the points in its cluster
  - Calculate \(b\) = min (average distance of \(i\) to points in another cluster)
  - The silhouette coefficient for a point is then given by
    \[ s = 1 - \frac{a}{b} \text{ if } a < b, \quad \text{or } s = \frac{b}{a} - 1 \text{ if } a \geq b, \text{ not the usual case} \]
    - Typically between 0 and 1.
    - The closer to 1 the better.
- Can calculate the Average Silhouette width for a cluster or a clustering
### Spectral Clustering: Criteria for Parameter Selection

#### MSE:

\[ MSE_f = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \sqrt{(y_i - m_i)^2} \]

#### Eigengap

\[ \delta(A) = 1 - \frac{\lambda_2}{\lambda_1} \quad \delta_K = \min_{i=1\ldots K} \delta(L(A_k^{(w)})) \]

#### Rcut

\[ RCut_K = \frac{\sum_{k=1}^{K} \sum_{d=1}^{K} \sum_{j \in S_k} \sum_{j' \in S_{k'}} A_{jk}}{\sum_j \sum_j A_{ij}} \]

### Alignment

\[ A = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}} = \frac{\langle K, yy^* \rangle_F}{n \sqrt{\langle K, K \rangle_F}} \]

#### Modified Alignment: consider only Nearest-Neighbors

<table>
<thead>
<tr>
<th>( K_i )</th>
<th>( A_1 )</th>
<th>( A_2 )</th>
<th>( A_3 )</th>
<th>( A_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( e \left( \frac{D^2}{2 \sigma^2} \right) )</td>
<td>( e \left( \frac{D^2}{2 \cdot 0.08^2} \right) )</td>
<td>( e \left( \frac{D^2}{2 \cdot 5^2} \right) )</td>
<td>( d_{max} - d )</td>
<td></td>
</tr>
</tbody>
</table>
Selection of parameter values: $\sigma$ and $K$

None of the methods systematically selected good parameter values.

Using a majority voting mechanism does not significantly improve results.

Percentage of correct identification:
- $\sigma$
- $K$

From Single Clustering to Ensemble Methods - April 2009
Cluster Validity Measures: Relative Criteria

The fundamental idea is to choose the best clustering scheme of a set of defined schemes according to a pre-specified criterion.

The problem can be stated as follows:

Let $P$ the set of parameters associated with a specific clustering algorithm (e.g. the number of clusters). Among the clustering schemes, $C_i=1,\ldots,i=1,\ldots,nc$, defined by a specific algorithm, for different values of the parameters in $P$, choose the one that best fits the data set.

Complexity-Based Validation:

- Used in model-based clustering: add a complexity term to the negative Log-likelihood
- Basic Idea: Occam’s razor
  - Choose the model that provides the shortest description of the data
- Rissanen’s Minimum Description Length (Rissanen, 1978)
- Schwartz’s Bayesian Information Criterion (Schwartz, 1978)
Cluster Validity Measures: Relative Criteria

MDL minimizes the overall description length of the data, where description consists of the data and the model parameters

\[ \hat{k} = \arg \min_{1 \leq k \leq K_{\max}} \left\{ -\log \hat{f}(x | \hat{\Theta}_k) + \frac{k'}{2} \log(n) \right\} \]

where \( k' \) is the number of independent parameters in the model \( \Theta_k \)

Schwartz’s Bayesian Information Criterion: takes a Bayesian perspective

- Select the model with the largest posterior probability

\[ \log(\Pr(\Theta_k | x)) \propto \log(\Pr(\Theta_k)) + \log(\Pr(x | \Theta_k)) \]

\[ \approx \log \hat{f}(x | \hat{\Theta}_k) + \frac{k'}{2 \log(n)} \]

MDL and BIC are formally equivalent

Cluster Validity Measures: Relative Criteria

Stability-Based Validation

Basic Idea:

- Solutions on two data sets from the same source should be similar
Cluster Validity Measures: Relative Criteria

Variant of stability-based cluster validity:

1. Regard a clustering algorithm as a statistical estimator for the partition of data space.
2. Use bootstrapping or sub-sampling to estimate the variability of this estimator.
3. The variability is interpreted as the validity of the clusters.

Stability-Based Validity Index

Summary of the technique:

1. B bootstrap samples, \( Y(1), \ldots, Y(B) \), each of size \( n \), are generated by sampling \( Y \) with replacement.
2. The clustering algorithm \( A \) is run on each \( Y(b) \) and we obtain the corresponding partition.
3. Variability is calculated using one of the dissimilarity measures of the partitions.
4. Note: normalization is needed to compensate for the difference in the number of possible partitions with different number of clusters.
Stability-Based Validity Index

Three data sets from UCI repository are used for different clustering algorithms:

- **Wine recognition**
- **Wisconsin diagnostic breast cancer**
- **Image segmentation**

The graphs show the partition variability for different clustering algorithms:
- **k-means**
- **EM**
- **complete**

The graphs illustrate the stability of the clustering results across different numbers of clusters for each data set.