Clustering and Dissimilarity Measures

Clustering

- What salient structures exist in the data?
- How many clusters?

Cluster Analysis

- Grouping observations based on [dis]similarity
- E.g. data mining [exploration, searching for concepts in data]
  - Clustering species based on genetic similarity
  - Finding typical customer behaviour
  - Reducing amount of data to be analysed, helps defining concept [class]
- Selecting typical class examples
  - Multi-modal classes may be represented using typical examples
  - Interpretation is not a goal here!

Dissimilarity Measures

- Let \( d(r, s) \) be the dissimilarity between objects \( r \) and \( s \)
- Formally, dissimilarity measures should satisfy
  - \( d(r, s) \geq 0 \)
  - \( d(r, r) = 0 \)
  - \( d(r, s) = d(s, r) \)
- If triangle inequality holds measure is a metric
  - \( d(r, t) + d(t, s) \geq d(r, s) \)

E.g. Measures Between Distributions

- Histogram intersection
- Kullback-Leibler divergence
  - Efficiency coding distribution using other as code-book
- Kolmogorov-Smirnov
  - Maximum difference between cumulative distributions
- Chi squared statistic
  - Likelihood of one distribution drawn from the other

Perceptually-Inspired Measures

- Earth-mover’s distance (EMD)
  - Transforms one object into another by shifting “evidence” in a feature space
  - Compare to L1 metric
- Tversky counting similarity
  - Large set of “predicates” [detectors] is defined [e.g. is the object round?]?
  - Similarity increases with increasing number of matching predicates
- Dynamic partial function [DPF]
  - Large number of features is computed for both images
  - Compare m smallest feature differences with Minkowski metric

May 19, 2008
Data-Specific Measures

- Measures defined for binary data:
  - Spectral angle mapper \( \theta_{\text{SA}}(x,y) = \arccos \left( \frac{x \cdot y}{||x|| \cdot ||y||} \right) \)
  - Derivative-based distances
  - Using derivatives of spectra, emphasizing shape differences

- Dissimilarity measures for spectra:
  - Very large field, huge number of methods
  - See for example Theodoridis and Koutroumbas, Pattern Recognition, 2003
  - More than 240 page overview of cluster analysis

<table>
<thead>
<tr>
<th>object x</th>
<th>object y</th>
<th>Similarity measure</th>
<th>Metric</th>
<th>Euclidean</th>
<th>Mahalanobis</th>
<th>Distance data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>X_1 = 1 - \theta_{\text{SA}}(x,y)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>X_1 = 1 - \theta_{\text{SA}}(x,y)</td>
<td></td>
</tr>
</tbody>
</table>

Hard vs. Soft

- Hard assignments
  - k-Means
  - Hierarchical clustering

- Soft assignments
  - Fuzzy c-means
  - Probabilistic mixture models

k-Means Algorithm

- Input: dataset, desired number of clusters m
- Output: sample labels
- Choose arbitrary initial prototypes \( \mu_i, i = 1, \ldots, m \)
- Loop:
  - Determine the closest prototype for each observation (label the observations)
  - Compute new prototypes as cluster means
- Repeat the loop until there is no change in prototypes

Clustering Algorithms

- Very large field, huge number of methods
- See for example Theodoridis and Koutroumbas, Pattern Recognition, 2003
- More than 240 page overview of cluster analysis

- Dissimilarity: squared Euclidean distance
- Minimize the criterion:
  \[
  J(\mathbf{x}, D) = \sum_{i=1}^{N} \sum_{j=1}^{m} w_{ij} || x_i - \mu_j ||^2
  \]

- Iterative procedure started from random prototypes
- Produces crisp assignment [binary sample weights]
### k-Centers / k-Medoids
- Minimizes maximum distance within objects in the cluster
- Selects existing objects as prototypes

![k-means clustering](image1) ![k-centres clustering](image2)

### Fuzzy c-Means Clustering
- Set of N observations is soft-assigned to m parameterized clusters
- Cost function:
  \[
  J_c(\mathbf{x}, \mathbf{u}) = \sum_{i=1}^{N} \sum_{j=1}^{m} u_{ij}^c d(x_i, \mathbf{v}_j)
  \]
- Minimized subject to sample weights and cluster model parameters
- Again, sample weights and model parameters are obtained by an iterative algorithm

### Probabilistic Mixture Model
- Probabilistic mixture model:
  \[
  p(x|\theta) = \sum_{k=1}^{K} u_k \pi(x_k|\theta_k)
  \]
- Mixing proportions: \( u_k \geq 0, \sum_{k=1}^{K} u_k = 1 \)
- Often Gaussian mixture is used:
  \[
  \pi(x_k|\theta_k) = \frac{1}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)
  \]
- Probabilistic clustering allows for overlapping clusters
- Model parameters are usually estimated by maximum-likelihood approach using Expectation-Maximization (EM) algorithm

### Maximum-Likelihood Estimation
- Model and parameters:
  \[
  p(x|\theta) = \sum_{k=1}^{K} u_k \pi(x_k|\theta_k)
  \]
- Likelihood:
  \[
  L(\theta|X) = \prod_{i=1}^{N} p(x_i|\theta)
  \]
- Likelihood is a function of parameters, data samples remain fixed

### EM Algorithm
- EM = expectation maximization
- Expectation step computes an expectation of the likelihood by including the unknown labels as if they were observed
- Maximization step computes maximum likelihood estimates of parameters by maximizing expected likelihood found in the E step
- This process is iterated

### EM Algorithm
- E step
  \[
  \gamma(x_k|\theta) = \frac{u_k \pi(x_k|\theta_k)}{\sum_{j=1}^{K} u_j \pi(x_k|\theta_j)}
  \]
- M step
  \[
  \mu_k = \frac{1}{\sum_{i=1}^{N} \gamma(x_i|\theta)} \sum_{i=1}^{N} \gamma(x_i|\theta) x_i
  \]
  \[
  \Sigma_k^{\text{new}} = \frac{1}{\sum_{i=1}^{N} \gamma(x_i|\theta)} \sum_{i=1}^{N} \gamma(x_i|\theta) (x_i - \mu_k^{\text{new}}) (x_i - \mu_k^{\text{new}})^T
  \]
- This process is iterated
**EM for Mixture Models**

- EM clustering
  - Assumes a priori known number of clusters K
  - Guarantees finding of [only] local optimum
  - May converge slowly
  - Is dependent on initialization
- An example:

**“Generalized” EM Clustering**

- Replacing probability model by an arbitrary classifier
- E step: assign each observation $x$ by classifier $S$ to one of the classes
- M step: use the labels to train new classifier $S$
- Stopping criterion: Labels do not change
- Note that:
  - `emclust` provides a final trained classifier which may be applied to new data
  - Move from soft to hard sample assignments
  - Some classifiers allow for soft labels [see dataset `labtype`]

**Agglomerative Hierarchical Clustering**

- Agglomerative algorithms: starting from individual observations, produce a sequence of clusterings of increasing cluster size
- At each level, two clusters chosen by a criterion are merged

**Different Combining Rules**

- Two nearest objects in the clusters: single linkage
  \[ d(X, S) = \min_{x, y \in X, S} d(x, y) \]
- Two most remote objects in the clusters: complete linkage
  \[ d(X, S) = \max_{x, y \in X, S} d(x, y) \]
- Cluster centers: average linkage
  \[ d(X, S) = \frac{1}{|X||S|} \sum_{x \in X, y \in S} d(x, y) \]
Minimum Spanning Tree

- Spanning tree is a weighted graph connecting all vertices without loops.
- Minimum spanning tree (MST) is a spanning tree with minimum total weight.
  - Weights on edges unique then unique MST.

Evaluation of Clustering Validity

- Every clustering algorithm will produce some result, but which one is better?
- Clustering is an ill-posed problem and results should be evaluated.

Evaluation Strategies

- Expert judgment: can the identified clusters be interpreted?
- External criterion: if clustering is used to define a set of prototypes for building a classifier, what is the eventual classification performance?
- Stability: which solution remains unchanged under data perturbation, parameter change or over scales?
- Based on the user-defined “ground-truth” data partitioning [Problematic: if user knows the grouping of data, why not use supervised techniques?]

Number of Clusters?

- Hierarchical clustering: maximum lifetime criterion
  - Problems: noise sensitivity in single linkage
- Based on clustering stability
  - Choose clustering which is the most stable to data perturbation, parameter choice or initialization

Number of Clusters?

- Probabilistic methods: penalized likelihood
  - $-\log$ likelihood + degrees of freedom
- AIC, BIC, MDL, etc.
- Validity indices: many methods based on different definitions of cluster compactness and intra-cluster diversity
  - Dunn index, Davies-Bouldin (DB) index, SD index, Xie-Beni index, and so on and so forth
- Problem: often derived on simple artificial problems strongly imposing data structure

Conclusions

- Many decisions to be made:
  - Measure (dis)similarity between observations
  - What type of structures we search for [blobs, elongated, whatever but stable, …]
  - Choice algorithm parameters [number of clusters, thresholds, scale, …]
  - How to evaluate clustering result? [panel of experts, final classification error, …]
- Clustering is an ill-posed problem
References

- Rubner, Perceptual Metrics for Image Database Navigation, 1999
- Theodoridis and Koutroumbas, Pattern Recognition, 2003