Clustering

Patrick Cash
Outline

• Introduction
• Hierarchical Clustering
  – Top-down vs. Bottom-up clustering
    • Single link
    • Complete link
    • Group Average
• Non-Hierarchical (Flat) Clustering
  – K-means
  – EM algorithm
• Conclusion
Introduction

• Clustering is a widely used approach throughout AI (NLP, machine learning, etc.)
• Clustering is based on the idea that we can collect objects in the data into similar groups
  – Cluster so that: similar objects are within the same group and objects are dissimilar between groups
• Useful when there is no training data available and you are looking for natural patterns in the data
Introduction

- Objects are described and clustered using a set of features or attributes

- Clustering vs. Classification
  - Clustering is unsupervised and Classification is supervised
  - The result of clustering only depends on natural divisions in the data and not on any pre-existing categorization

- Hard vs. Soft clustering
  - Hard: each object belongs to one and only one cluster
  - Soft: each object can belong to more than one cluster
    - Object has a probability distribution over all clusters
Introduction

- Two main uses for clustering in NLP
  - Exploratory data analysis
    - Helps to understand the basic characteristics of a data set
    - Provides a visual representation of the data
  - Generalization
    - Forming bins or equivalence classes that are induced from the data
    - allows inference between same cluster members
Hierarchical Clustering

- Builds a tree-based hierarchical taxonomy from a set of unlabeled examples
- Often implies that a child node is a subclass of the parent node
- Two approaches:
  - Bottom-up: Agglomerative
  - Top-down: Divisive
Bottom-up Clustering

1. Start with a separate cluster for each object
2. Determine the two most similar clusters and merge into a new cluster. Repeat on the new clusters that have been formed
3. Terminate when one large cluster containing all objects has been formed
Cluster Distance Metrics

- Single link
  - Similarity of two most similar members
  - Good local cluster quality
- Complete link
  - Similarity of two least similar members
  - Good global cluster quality
- Group Average
  - Average similarity between members
  - A compromise between Single and Complete link
Single link
Complete link
Group Average

- Similarity metric is the average similarity between all the members for each cluster.
- This creates a compromise between single link and complete link clustering.
  - Can be faster than complete link and avoids chaining effect of single link.
Bottom-up Example

http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html
Top-down Clustering

• Starts with one large cluster containing all objects and iteratively splits the cluster based on coherence
• Can use single link, complete link or group average to determine cluster coherence
• Splitting the cluster is a clustering task in itself and any clustering algorithm can be used
  – The need for an additional clustering algorithm means top-down clustering is used less often, but it is a natural fit for some clustering tasks
Top-down Example
Non-Hierarchical Clustering

- Starts with a partition based on randomly selected seeds and then refine this initial partition
- Several passes of reallocating objects are needed (hierarchical algorithms need only one pass)
- Stop based on some measure of goodness or cluster quality
  - Heuristic: number of clusters, size of clusters, stopping criteria, etc.
- Non-Hierarchical clustering is usually faster than Hierarchical clustering
**k-means**

- Defines clusters by the center of the mass of cluster members
  1. Randomly pick a set of k cluster seed positions
  2. Assigning each object to a cluster based on some distance metric from the cluster seed positions
  3. Move seed to the new center of the cluster, determined by the cluster members position
  4. Repeat until centers do not change

- Can solve distance ties by randomly choosing a cluster or slightly moving the object
k-means

- Hard clustering – each object is assigned to only one cluster
- Determining k
  - Domain knowledge can be used to help determine k
  - Different values of k can be experimented with to determine the best value
  - Other learning methods can be used to learn k
- k-means needs a Euclidean based distance metric
Buckshot Algorithm

- Combines hierarchical bottom-up clustering and k-means clustering
  - First randomly take a sample of instances of size $\sqrt{n}$
  - Run group-average hierarchical bottom-up clustering on this sample, which takes $O(n)$ time
  - Use the results as the initial seed set for k-means
  - Avoids problems cause by bad seed selection and gives k-means efficiency
K-means Example

http://home.dei.polimi.it/matteucc/Clustering/tutorial/html/AppletKM.html
EM Algorithm

• The EM algorithm is a general template for a family of algorithms
  – Currently very popular and widely used in NLP and machine learning

• EM can be seen as a “soft” version of k-means clustering
  – Assigns objects to more than one cluster using a probability distribution
EM Algorithm

• Two steps
  – Expectation step: Use current parameters to reconstruct hidden structure
  – Maximization step: Use that hidden structure to re-estimate parameters

• Model
  – Parameters: k points representing cluster centers
  – Hidden structure: for each data point, which center generated it?
EM for Gaussian Mixtures

• EM is estimating a mixture of Gaussian probability distributions
  – Assumes the final distribution we see was generated by several independent underlying causes

• Represents the data as a pair: observable data and hidden data
  – Observable data is location of each object
  – Hidden data is probability that data point belongs to a cluster

• Once the estimation has been done we interpret each underlying cause as a cluster and determine a probability for each object
EM Example Applications

- Baum-Welsh re-estimation (forward-backward)
  - E step: computes expected number of transitions from each state in the observed data and for each pair of states the expected number of transitions between them
  - M step: computes new MLE for initial state, state transition and symbol emission probabilities

- Inside-outside algorithm
  - E step: expected number of times a rule is used
  - M step: computes MLE for rule probabilities
EM Example Applications

• Unsupervised word sense disambiguation
  – E step: expectations of cluster membership
  – M step: MLE probability of a cluster generating a specific word
• k-means
  – K-means can be seen as a special case of EM where the mean of the distribution is the only variable
  – E step: estimate cluster membership using distance metric
  – M step: move seeds to new cluster centers
Problems with EM

- EM can be very sensitive to initialization
  - Clustering can get stuck in local minima
  - Other clustering algorithms can be used for initialization
- EM convergence can be very slow
- EM is only really needed when there is not an easier way to solve the constraint problem
EM Example

http://www.cs.cmu.edu/~alad/em/
Properties of hierarchical and non-hierarchical clustering

Hierarchical Clustering:
- Preferable for detailed data analysis
- Provides more information than flat clustering
- No single best algorithm (dependent on application)
- Less efficient than flat (for n objects, n X n similarity matrix required)

Non-Hierarchical Clustering:
- Preferable if efficiency is a consideration or data sets are very large
- k-means is the conceptually simplest method
- k-means assumes a simple Euclidean representation space and so can’t be used for many data sets
- In such case, EM algorithm is chosen
References


• Image Clustering http://www.cs.bilkent.edu.tr/~canf/CS533/CS533Spr06stuPresent/imageClustering.ppt

• Natural Language Processing: Clustering http://www2.mta.ac.il/~gideon/courses/nlp/slides/chap15_clustering.ppt

Questions ?