TIM 245 Lecture 14 (5/22/17)

Agenda

1) Finish Lecture 13

2) General Process for Unsupervised Learning

3) Types of Clustering Algorithms

4) Partition Based Clustering and the K-Means Algorithm

5) Return HW1 Corrections

② General Process for Unsupervised Learning Problems

Objective: discover useful patterns (structure) in the attributes and instances of the dataset.

Attributes: Which attribute values occur together (Frequent itemset rules)

pattern: $\{V_{1,1}, V_{3,1}\} \rightarrow \{V_{4,1}\}$

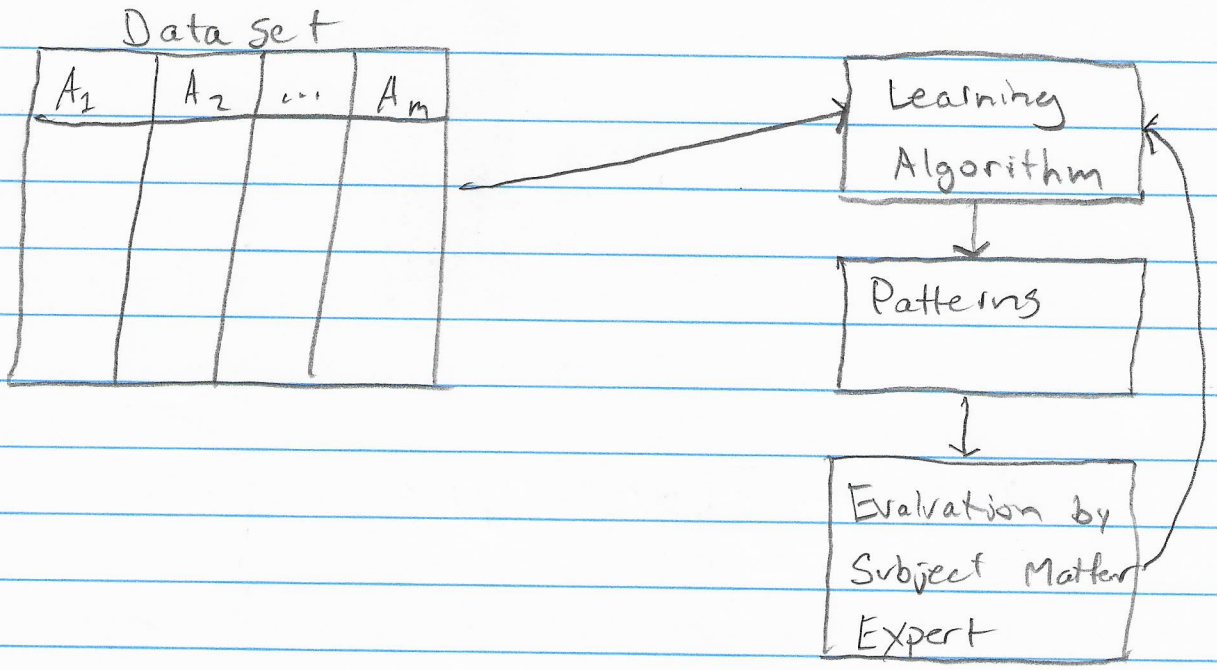
Where V_{jk} is the K-th value of A_j

Instances: Which instances are similar to eachother (clusters)

pattern: $C_L = \{d_4, d_5 \ldots d_{17}\}$

Where d_i is the ith instance

These patterns are discovered through an unsupervised Learning process



Two general applications for unsupervised Learning (clustering and association analysis)

Domain Understanding

Subject Matter expert uses patterns to better understand the underlying process creating the data

Pre Processing

Preparing data for other mining tasks

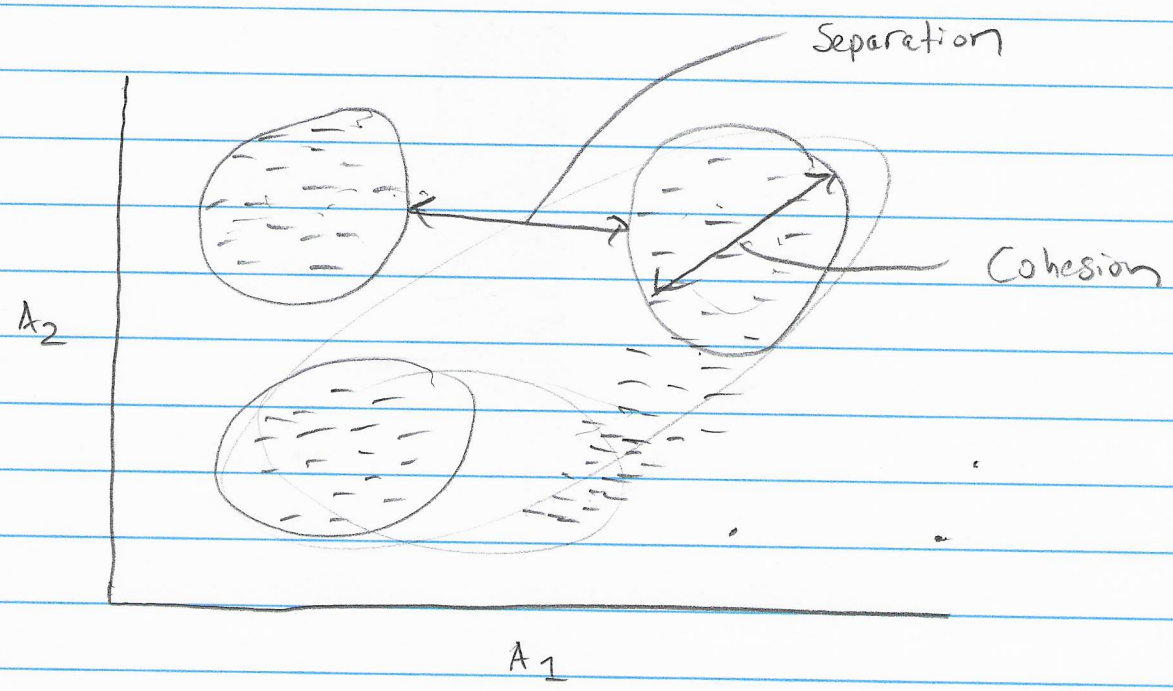
- Anomaly Detection
- Dimensionality Reduction
- Sampling

Key issue: Interpretability and usability of the extracted patterns

There is no objectively "best" pattern, it depends on the use case and the judgement of the subject matter expert.

(If we knew the best pattern, it would be a supervised learning problem)

(3) Types of Clustering Algorithms



We want to find groups or clusters where

1) Instances in the same cluster are similar (cluster Cohesion)

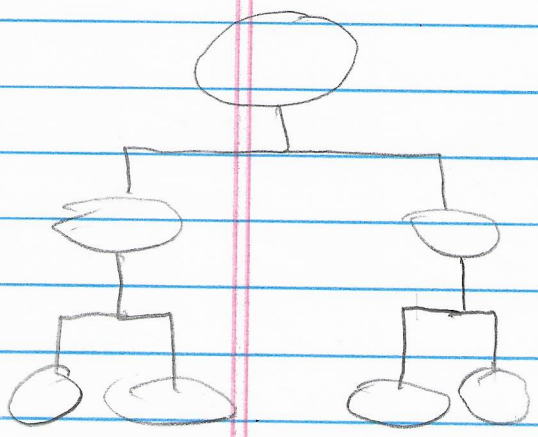
2) Instances in different clusters are dissimilar (cluster Separation)

There are four general approaches

1) Partitioning Methods: organize (classify) the instances into k disjoint sets or partitions so that intra cluster distance is minimized and inter-cluster distance is maximized

Popular Algorithms: K-Means

2) Hierarchical Methods: Create a hierarchical (nested) decomposition of the instances based on either merging similar instances (Bottom-up) or splitting groups (top down)



Agglomerative Clustering: Start with each instance as its own cluster. At each step merge the closest pair of clusters until there is only one cluster

Divise Clustering: Start with one cluster that includes all instances. Split until each instance is its own cluster.

3) Density-Based Methods: grow clusters as long as the density (number of instances) in the surrounding area exceeds a user specified threshold.

Popular Algorithms: DBScan, Optics

4) Model Based Methods: hypothesize a statistical distribution for K clusters, e.g. multivariate normal. Find the parameters that maximize the likelihood of the observed data.

Popular algorithm: Expectation Maximization (EM)

The choice of the clustering algorithm will depend on the data, subject matter expert, and use case.

(4) Partition Based Clustering and the K-Means Algorithm

Basic Process for Partition Based Clustering

1) Select K, the number of clusters

2) Assign each instance to a single cluster based on minimizing a cost or objective function

Total numbers of possible clusters assignments is K^n, intractible for most datasets

K-Means Algorithm:

- Cost function is sum of the squared error, where error is the distance from the cluster mean or centroid.

- Cluster assignment process is iterative