Clustering

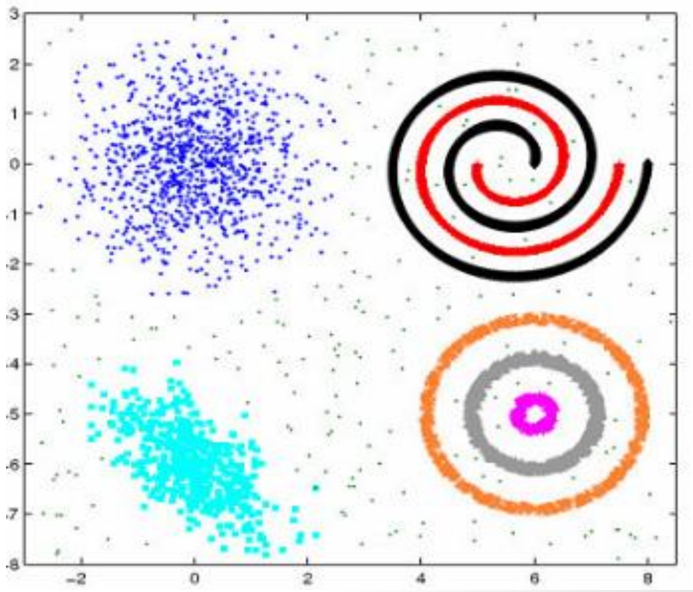
A Deeper Look at Clustering as a Machine Learning Technique

Overview

- Form of unsupervised learning
  - i.e. goal is not to learn the underlying mechanism, instead explore data and try to best represent it
  - Results require subjective interpretation
  - Interpretation is good match if results fit the intended outcomes
- Does not require assumptions such as in *confirmatory data analysis*
- Explosion of visual data captured by over 1 billion camera phones has produced over 2810 exabytes (or billions of gigabytes) of data.
- To archive and effectively use this data, we need tools for data visualization and categorization.

Overview

- Goal of clustering is to discover the natural grouping(s) of a set of patterns, points, or objects.
- More formally, given *N* data objects, we seek to kind *K* groups where the objects in the *same* grouping are similar in a way they are not in another.
- Groupings are determined based on a measure of *similarity* between data objects
- What constitutes similarity can vary



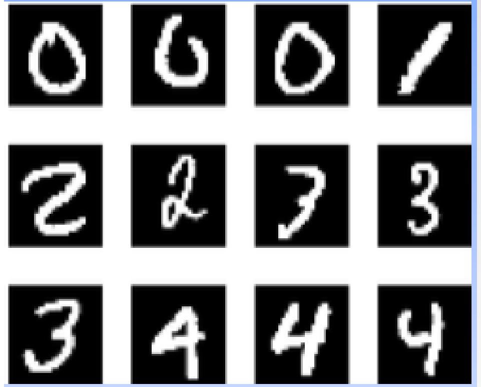
Over Time

- Cluster analysis first appeared in the title of a 1954 article analyzing anthropological data (JSTOR)
- Hierarchical Clustering: Sneath (1957), Sorensen (1957)
- **K-Means**: Steinhaus (1956), Lloyd (1957), Cox (1957), Ball & Hall (1967), MacQueen (1967)
- Mixture models (Wolfe, 1970)
- Graph-theoretic methods (Zahn, 1971)
- K Nearest neighbors (Jarvis & Patrick, 1973)
- Fuzzy clustering (Bezdek, 1973)
- Self Organizing Map (Kohonen, 1982)
- Vector Quantization (Gersho and Gray, 1992)

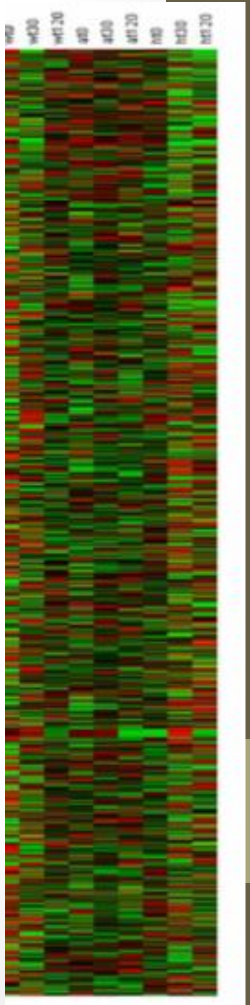
Issues

- What features and normalization scheme to use?
- How do we define pair-wise similarity?
- How many clusters?
- Which clustering method?
- How to choose algorithmic parameters?
- Does the data actually have any clustering tendency?
- Are the discovered clusters & partition valid/real?
- How to visualize , interpret & evaluate clusters?

What’s a Cluster?



- Clusters are a set of objects which are *similar*
- Objects from different clusters will be *dissimilar*
- An ideal cluster is compact and connected.
    - Compactness: inter-cluster data point distance is smaller than data point distance between clusters
    - Connectivity: inter-cluster data point connectivity is greater than data point connectivity between clusters
- Objects: pixels, images, time series, documents
- Represented via features, similarity



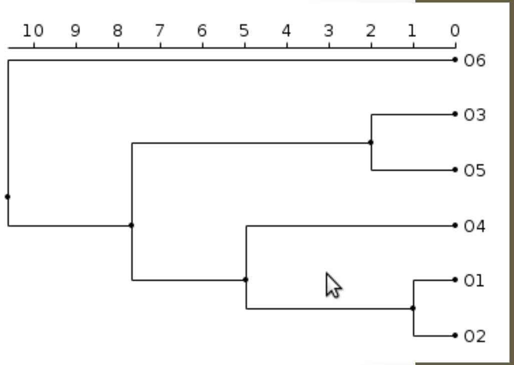
Types of Clustering Algo’s

- 2 broad types: Hierarchical and partitional
- Hierarchical
    - Top Down (Divisive)
    - Bottom Up (Agglomerative)
    - Single link, complete link

- Partitional
    - Tries to find all clusters simultaneously
    - No imposed hierarchy
    - K - means

Hierarchical

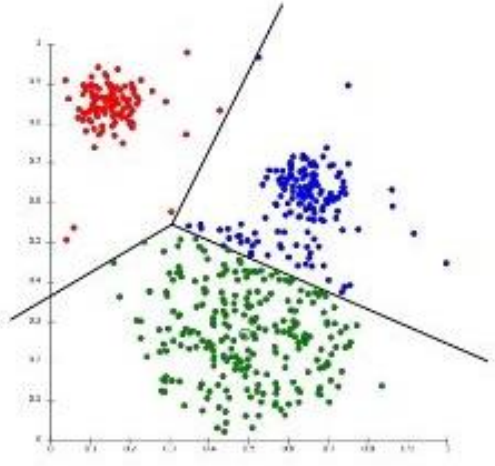
- All clustering functions first check the arguments for correctness.
- They then set up the model
  - Clusters
  - Linkages
  - Hierarchy(clusters, linkages)
- The normal and weighted clustering methods loop continuously to build out a tree according to the linkage strategy
- The flat clustering method calls a separate agglomeration function that takes a threshold as an additional parameter.
- Clusters are stored in an ArrayList & store a left and right pair of clusters associated with each link.



K Means

1. Select initial partition with K clusters;
   repeat 2 & 3 until clusters stabilize
2. Create new partition by assigning all patterns to closest cluster center
3. Generate new cluster centers

- Requires the number of clusters K (How do we pick?), the initialization of the clusters, and a distance metric
- Stabilizes during execution as means fit the data in their clusters more accurately.



Questions?