Clustering

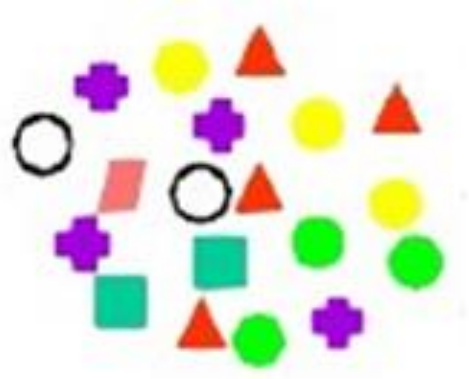
Data Science and Machine Learning Workshop' 2017

Habib University

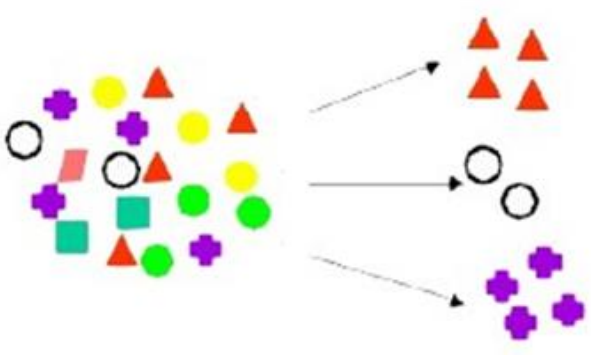
Acknowledgement

- Some slides of this lecture have been taken from:
  - lecture notes of Tan, Steinbach, Kumar Introduction to Data Mining

Know your data!



Know your data!



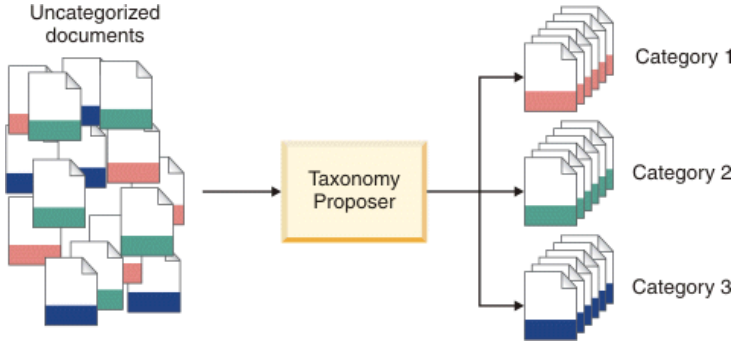
Related News

Why today's **earthquake** – 1200 km away – was felt in Delhi
The Indian Express - 26-Oct-2015
Almost exactly six months after the **Nepal earthquake** that killed nearly 10,000 people, an earthquake of similar magnitude hit north-west ...
Over 260 dead as 7.5 **earthquake** rocks Afghanistan, Pak and India
In-Depth - Hindustan Times - 26-Oct-2015
Explore in depth (4,890 more articles)

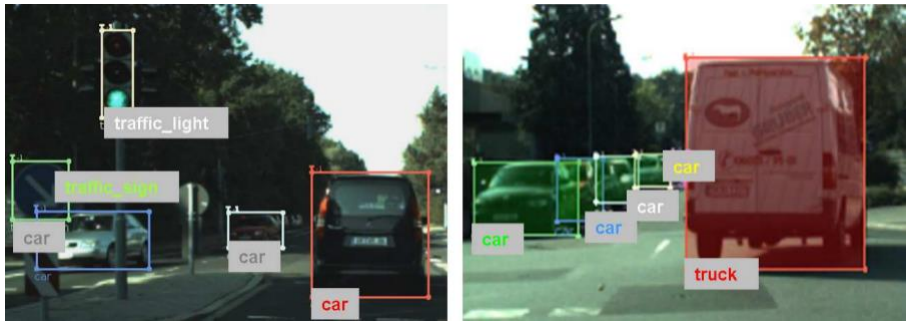
400-Plus Quakes Strike San Ramon in 2 Weeks: USGS
NBC Bay Area - 27-Oct-2015
San Ramon, California, appears to have broken a new **earthquake** record over the last two weeks: A total of 408 small ... (Published Tuesday, Oct. 27, **2015**).
Record-Breaking 408 **Earthquakes** Hit Bay Area City Over Past 2 ...
International - Live Science - 27-Oct-2015
Explore in depth (107 more articles)

Afghanistan **earthquake 2015**: Which country has the mo…
City A.M. - 26-Oct-2015
Today, the world was struck by yet another major **earthquake**. This time it was in the mountainous Kush region in northern Afghanistan, close to ...
Man clears rubble after **earthquake**
In-Depth - Economic Times - 27-Oct-2015
Explore in depth (166 more articles)

Documents Sorting

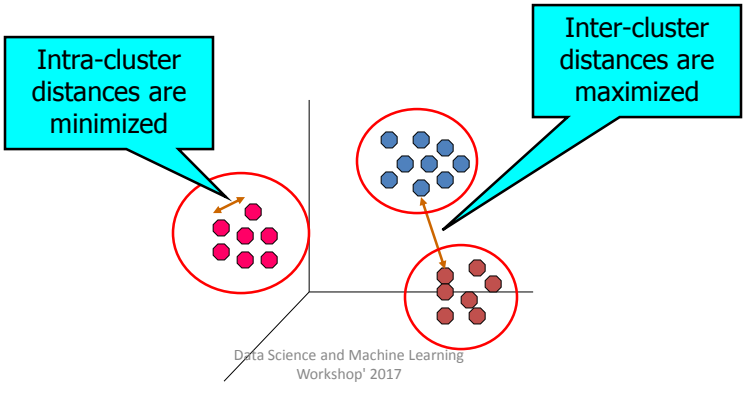


Object Detection



Cluster Analysis

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Cluster Analysis (Cont'd)

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters

What is not Cluster Analysis?

- Supervised classification
  - Have class label information
- Simple segmentation
  - Dividing students into different registration groups alphabetically, by last name
- Results of a query
  - Groupings are a result of an external specification
- Graph partitioning
  - Some mutual relevance and synergy, but areas are not identical

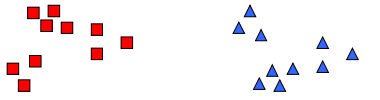
Applications

- Recommendation engines
- Market segmentation
- Social network analysis
- Search result grouping
- Medical imaging
- Image segmentation
- Anomaly detection
-

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters



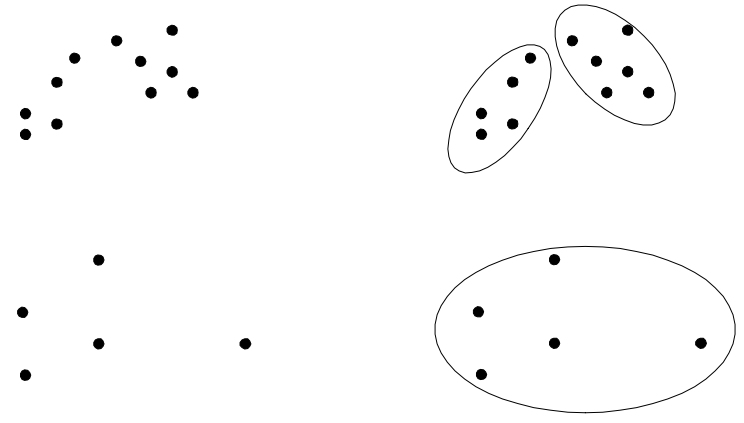
Two Clusters

Four Clusters

Types of Clustering

- Partitional Clustering
  - A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- Hierarchical clustering
  - A set of nested clusters organized as a hierarchical tree

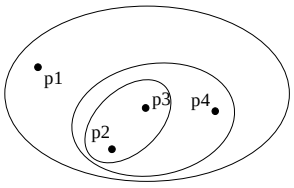
Partitional Clustering



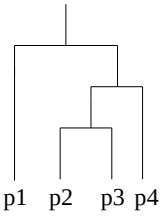
Original Points

A Partitional Clustering

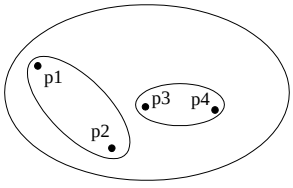
Hierarchical Clustering



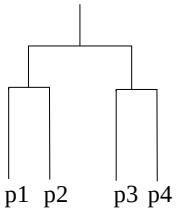
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

Notions of Distance and Center

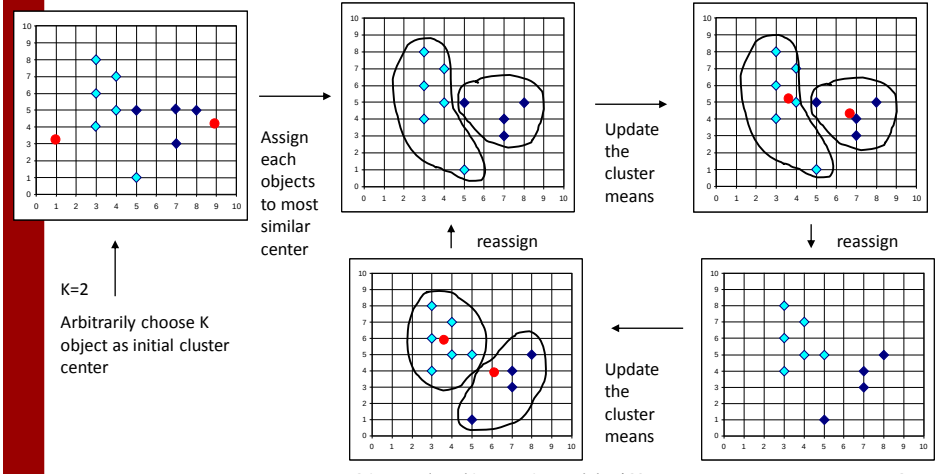
- Dissimilarity/Similarity metric: Similarity is expressed in terms of a distance function, typically metric: $d(i, j)$
- The definitions of distance functions are usually very different for interval-scaled, boolean, categorical, ordinal ratio, and vector variables.
- Weights should be associated with different variables based on applications and data semantics.
- Centroid of dataset

The *K-Means* Clustering Method

- Given *k*, the *k-means* algorithm is implemented in four steps:
  - Partition objects into *k* nonempty subsets
  - Compute seed points as the centroids of the clusters of the current partition (the centroid is the center, i.e., *mean point*, of the cluster)
  - Assign each object to the cluster with the nearest seed point
  - Go back to Step 2, stop when no more new assignment

The *K-Means* Clustering Method (Cont'd)

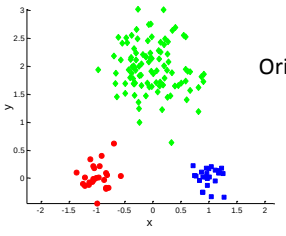
- Example

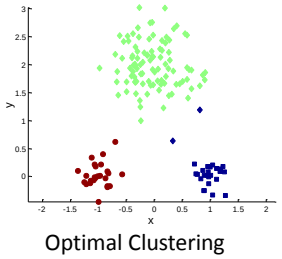


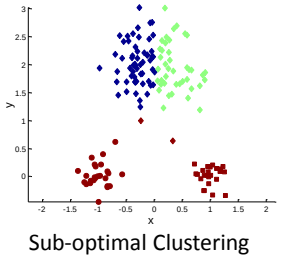
K-Means Clustering – Details

- Initial centroids are often chosen randomly.
  - Clusters produced vary from one run to another.
- The centroid is (typically) the mean of the points in the cluster.
- 'Closeness' is measured by Euclidean distance, cosine similarity, correlation, etc.
- K-means will converge for common similarity measures mentioned above.
- Most of the convergence happens in the first few iterations.
  - Often the stopping condition is changed to 'Until relatively few points change clusters'
- Outlier removal and feature normalization are important data pre-processing steps before applying K-Means.

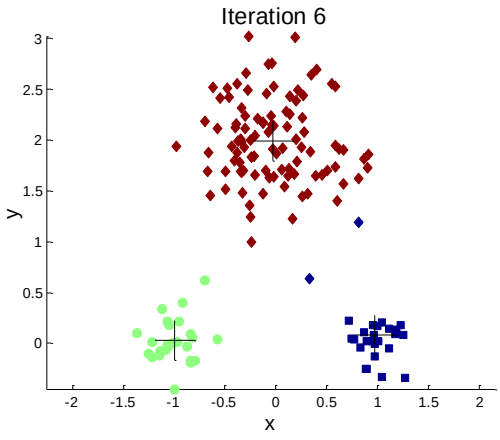
Two different K-means Clusterings



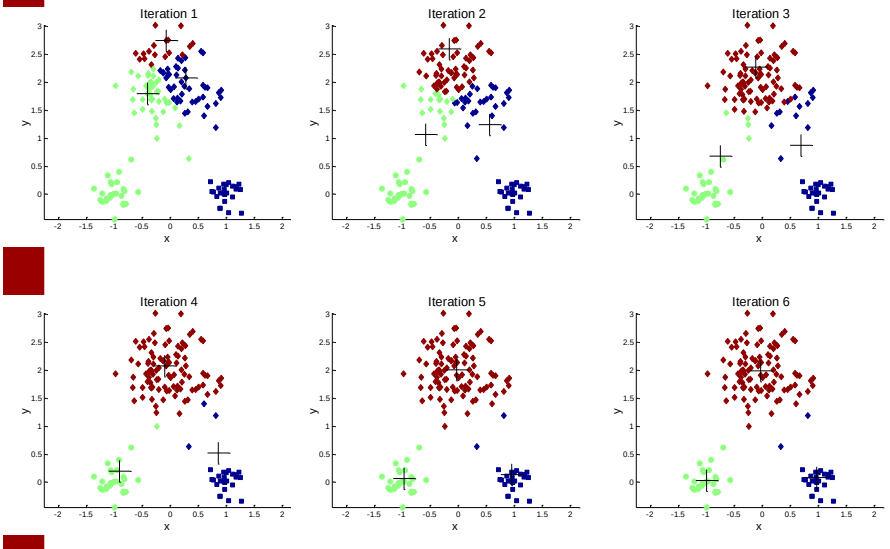

Optimal Clustering


Sub-optimal Clustering

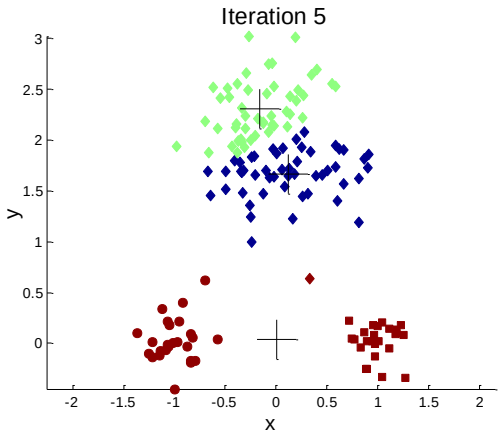
Importance of Choosing Initial Centroids



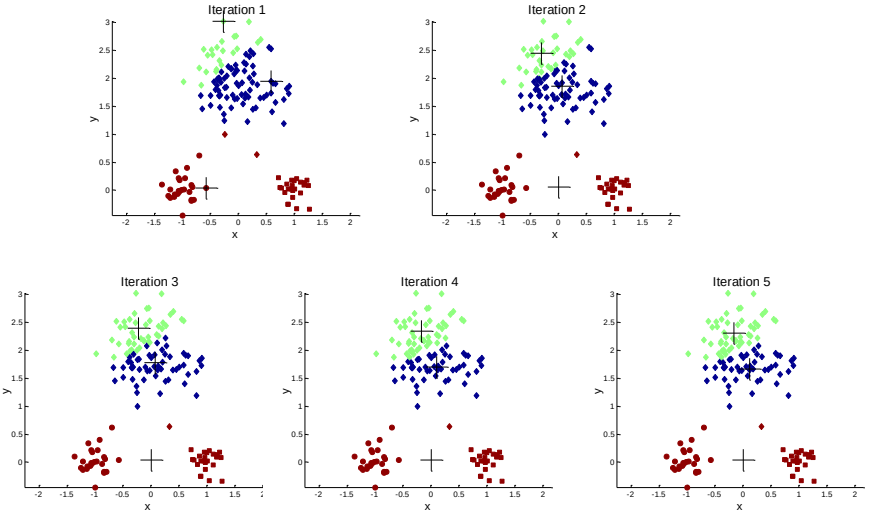
Importance of Choosing Initial Centroids



Importance of Choosing Initial Centroids …



Importance of Choosing Initial Centroids …



Evaluating K-means Clusters

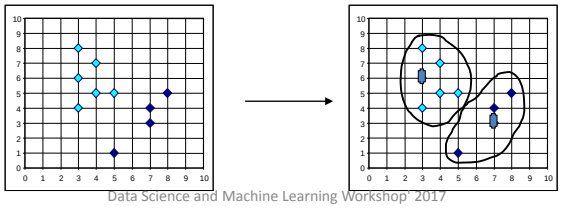
- Most common measure is Sum of Squared Error (SSE)
  - For each point, the error is the distance to the nearest cluster
  - To get SSE, we square these errors and sum them.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - x is a data point in cluster C_i and m_i is the representative point for cluster C_i
    - can show that m_i corresponds to the center (mean) of the cluster
  - Given two clusters, we can choose the one with the smallest error
  - One easy way to reduce SSE is to increase K, the number of clusters
    - A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

Limitations of K-means

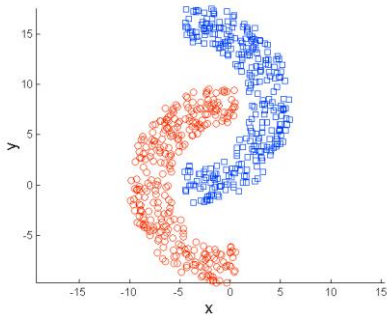
- The k-means algorithm is sensitive to outliers !
  - Since an object with an extremely large value may substantially distort the distribution of the data.
- K-Medoids: Instead of taking the **mean** value of the object in a cluster as a reference point, **medoids** can be used, which is the **most centrally located** object in a cluster.



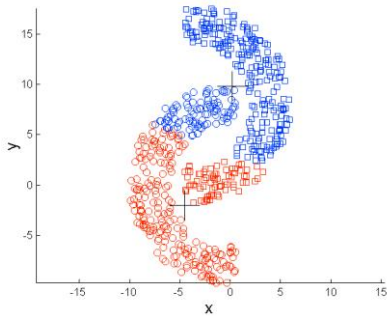
Limitations of K-means

- Applicable only when *mean* is defined, then what about categorical data?
- Need to specify *k,* the *number* of clusters, in advance
- Unable to handle noisy data and *outliers*
- Not suitable to discover clusters with *non-convex shapes*
- *K-means has problems when data contains outliers*
- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes

Limitations of K-means: Non-globular Shapes

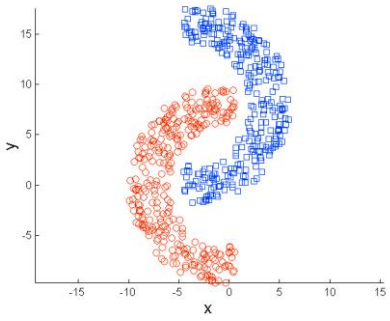


Original Points

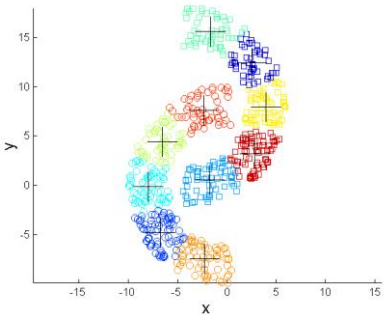


K-means (2 Clusters)

Overcoming K-means Limitations



Original Points



K-means Clusters

Pre-processing and Post-processing

- Pre-processing
  - Normalize the data
  - Eliminate outliers
- Post-processing
  - Eliminate small clusters that may represent outliers
  - Split 'loose' clusters, i.e., clusters with relatively high SSE
  - Merge clusters that are 'close' and that have relatively low SSE