Learning Analytics

Dr. Bowen Hui
Computer Science
University of British Columbia Okanagan

Last Class

- Overview of clustering
- Methods:
  - Hierarchical clustering (agglomerative)
    - MIN, MAX, Group Average
  - K-means
    - Given k, objective function, choice of initial centroids
  - Application with k-medoids
- Remaining issues:
  - How to choose k?
  - How to validate clusters?

How to Choose k

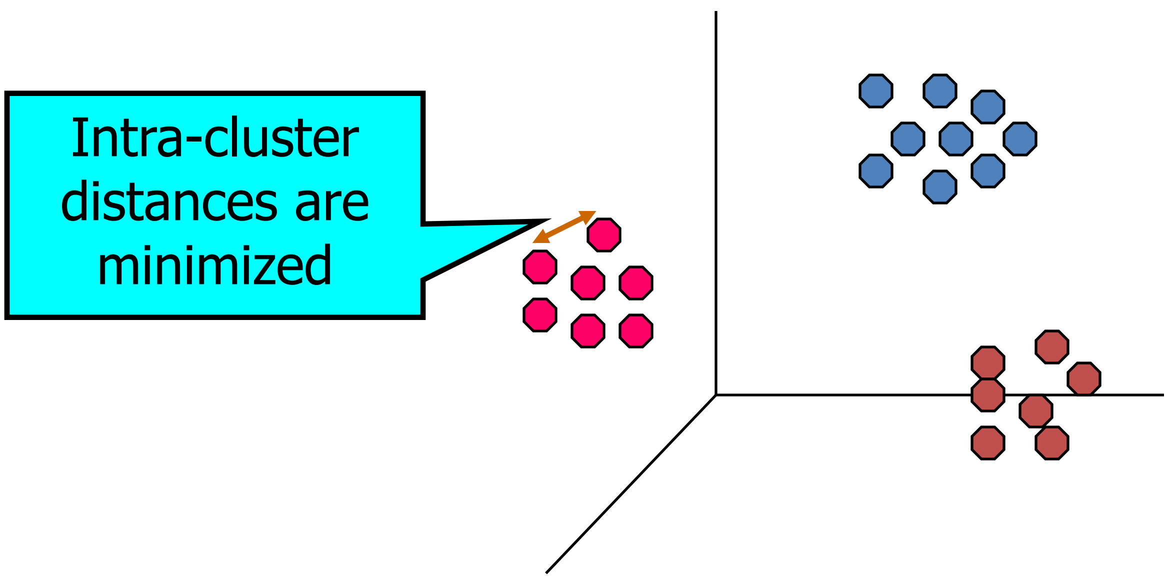
- Optimal number of clusters is somewhat subjective
  - Over 30+ approaches
  - Often determine k by “majority rule” approach
- Specific methods we will examine:
  - Elbow method
  - Silhouette method

Elbow method

- Recall SSE = $\sum_{i=1}^{k} \underbrace{\sum_{x \in C_i} dist(m_i, x)^2}_{\text{Within-cluster error}}$

where m_i is the mean of cluster C_i

Sum across all clusters



Algorithm for the Elbow Method

- Steps:
  - Compute clustering algorithm for different values of k
  - For each k, calculate SSE
  - Plot the curve of SSE as a function of k
  - The location of a bend (knee) in the plot is an indicator of an appropriate value for k

- Note: where the knee is can be ambiguous

Example

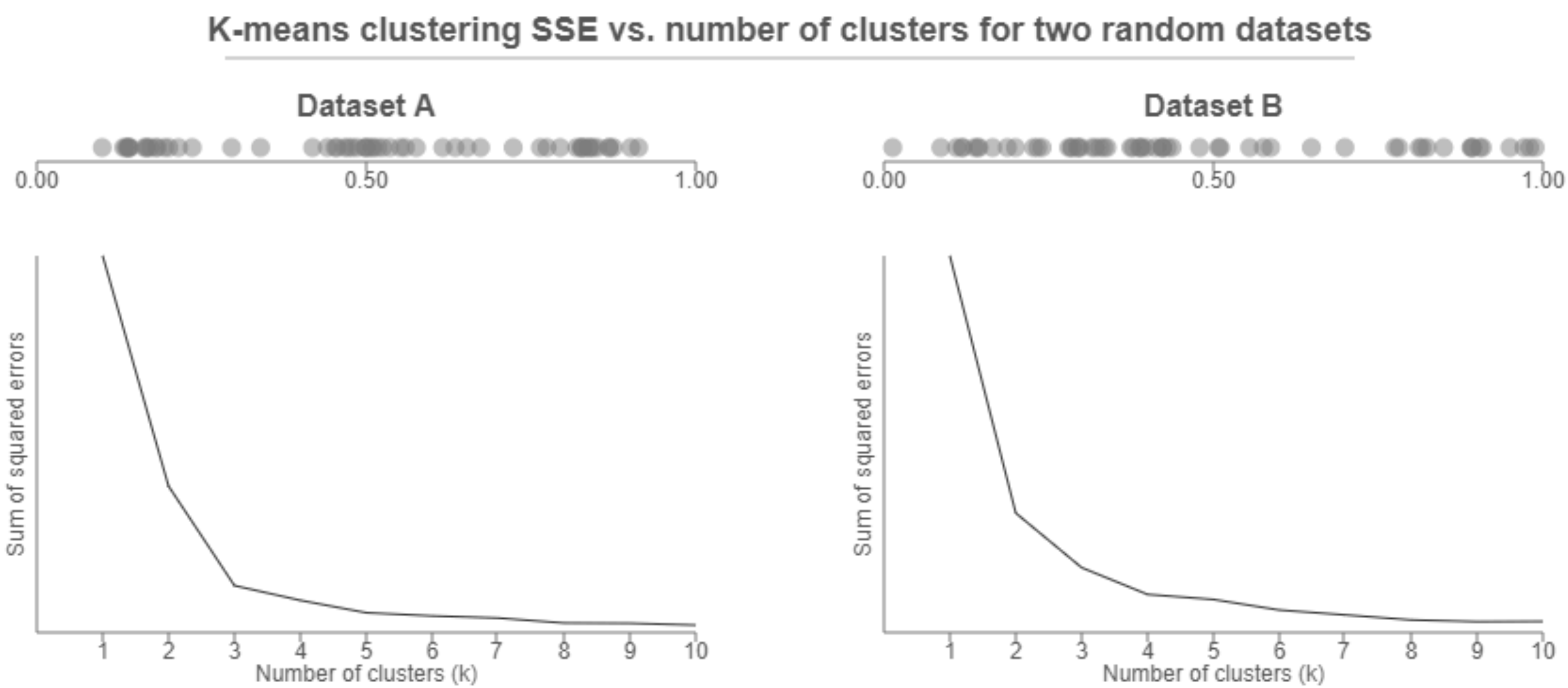


Image taken from medium.com

What should we use for k in either case?

Silhouette method

- Arguably more reliable than the elbow method
- Silhouette coefficient
  - Measures cohesion – how similar a point is to its own cluster
  - Measures separation – how far away a point is from other clusters
  - Ranges in [-1,+1], with higher value meaning a point is placed in the correct cluster
- Value reaches its global maximum at the optimal k
- If many points have negative value, it may suggest there are too many or too few clusters

Definition of the Silhouette Coefficient

- When $|C_i| = 1$: $s(i) = 0$
  defined this way to prevent an increase of singleton clusters
- When $|C_i| > 1$:

$$\boxed{s(i) = \frac{b(i) - a(i)}{\max(\,a(i), b(i))}}$$

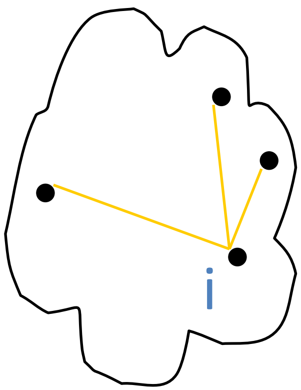
where:

$$\boxed{a(i) = \frac{1}{|C_i| - 1} \sum_{j \in Ci, i \neq j} d(i, j)}$$ is similarity of i to its own cluster

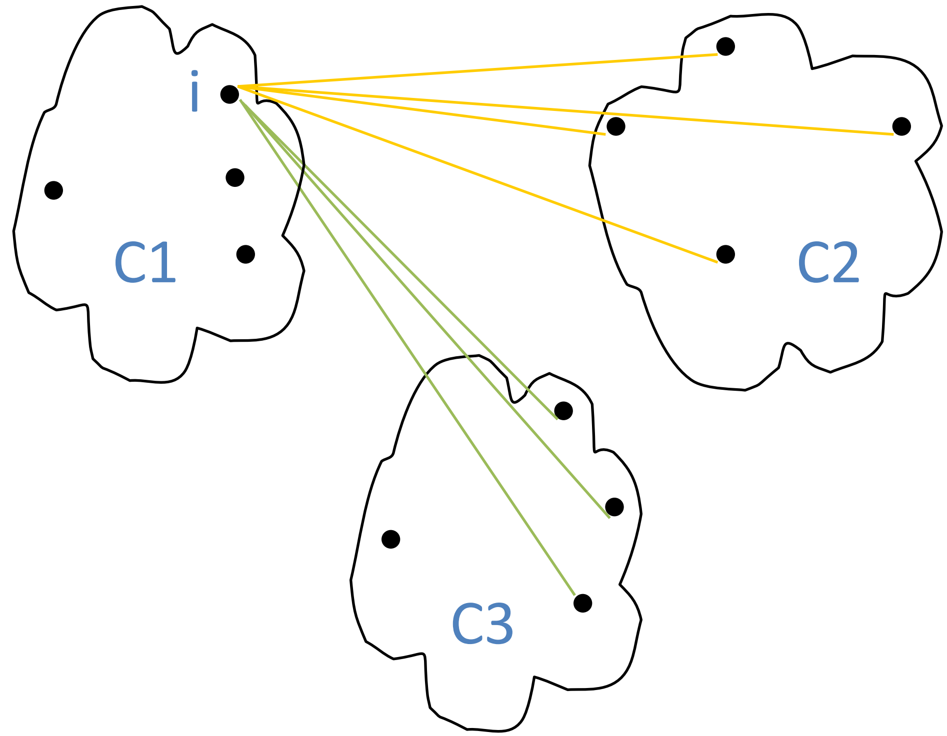
$$\boxed{b(i) = \min_{i \neq j} \frac{1}{|Cj|} \sum_{j \in Cj} d(i, j)}$$ is dissimilarity from i to other clusters

with d(i,j) defined as the distance between i and j (e.g. L2 norm)

Visualize a(i) and b(i)



Average distance from i to other points within cluster



Average distance from i to other points in one other cluster, then min of those averages

Algorithm for the Silhouette Method

- Steps:
  - Compute clustering algorithm for different values of k
  - For each k, calculate the average s(i) for all i
  - Plot the curve of average silhouette as a function of k
  - The location of a peak in the plot is an indicator of an appropriate value for k

Comparison Between Elbow and Silhouette Methods

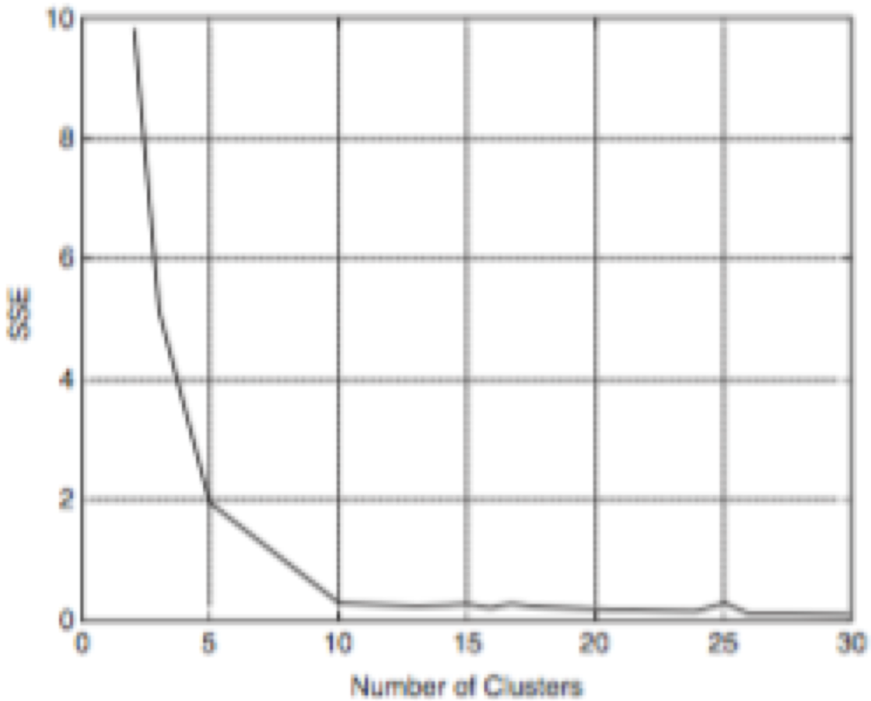


Figure 7.32. SSE versus number of clusters for the data of Figure 7.29 on page 582.

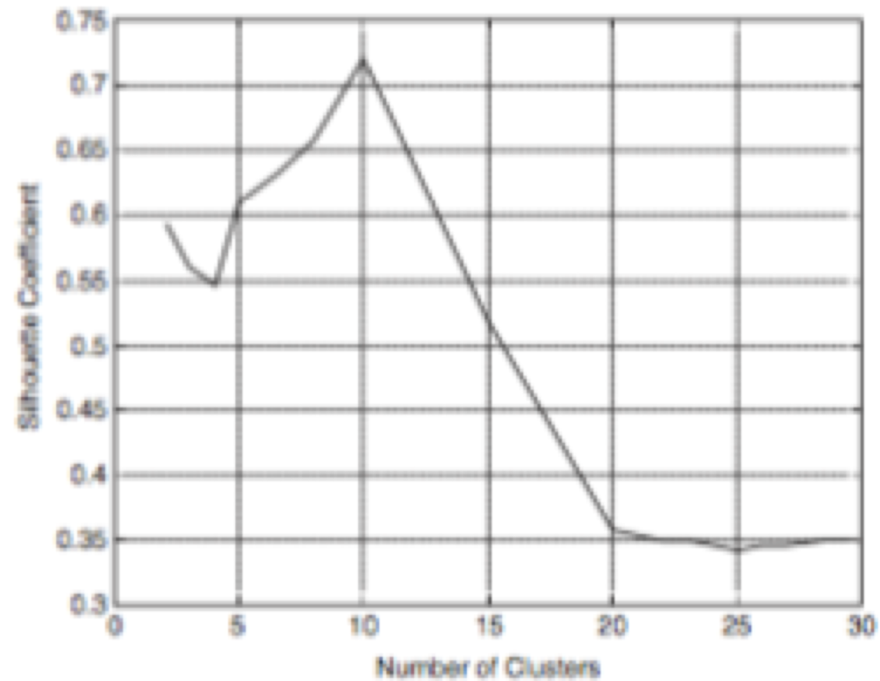


Figure 7.33. Average silhouette coefficient versus number of clusters for the data of Figure 7.29.

Clustering Tendency

- If you get poor cluster results, how to identify source of problem?
  - Is it the parameters chosen?
  - Is it the algorithm?
  - Is it the data set?
- If running multiple algorithms and parameter settings uniformly poor results, then this suggests there are no clusters in the data

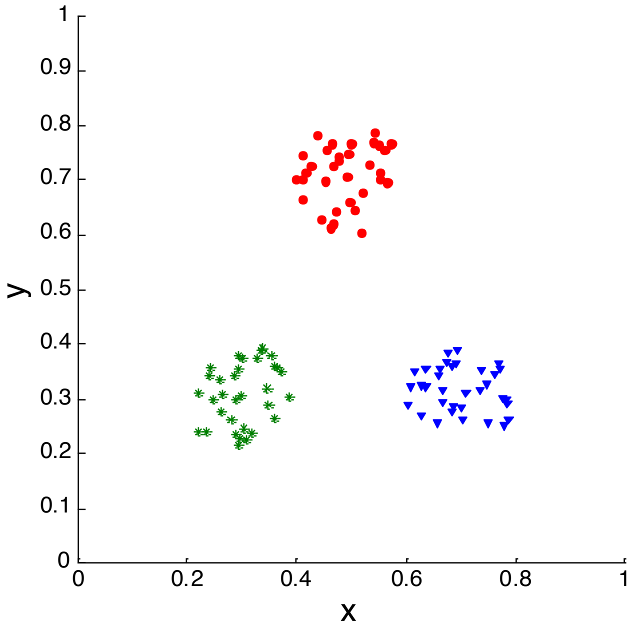
- Alternatively, use statistical measures to evaluate whether data has clusters without clustering
  - E.g. Hopkins statistic

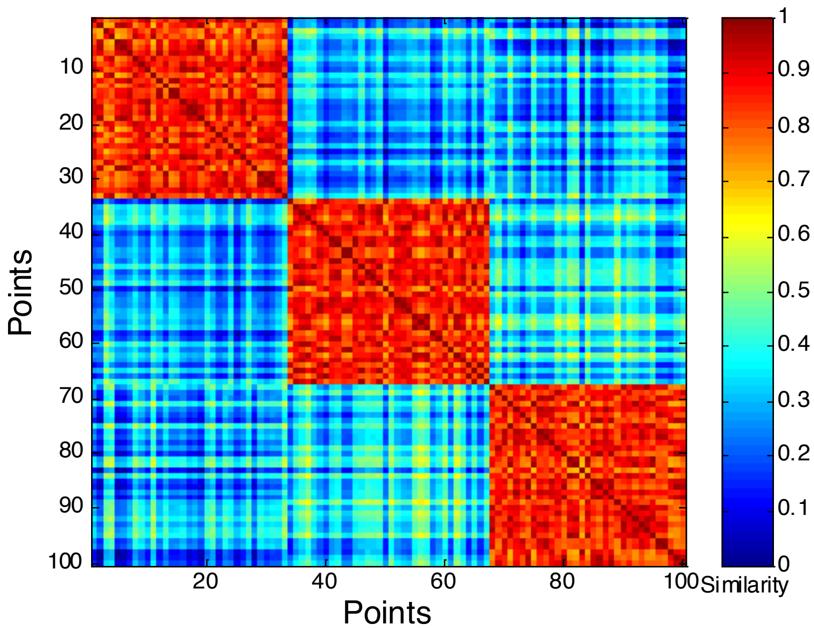
Measuring Cluster Validity via Correlation

- Idea: an ideal cluster is one whose points have similarity of 1 to all points in cluster, but 0 to all points in other clusters
- Two matrices
  - Proximity matrix
  - Ideal similarity matrix
    - One row and one column for each data point
    - Entry is 1 if the associated pair of points belong to same cluster
    - Entry is 0 if that pair of points belong to different clusters
- Compute the correlation between them
  - High correlation indicates points from the same cluster are close to each other
- Not a good measure for certain classes of algorithms

Using Similarity Matrix for Cluster Validation

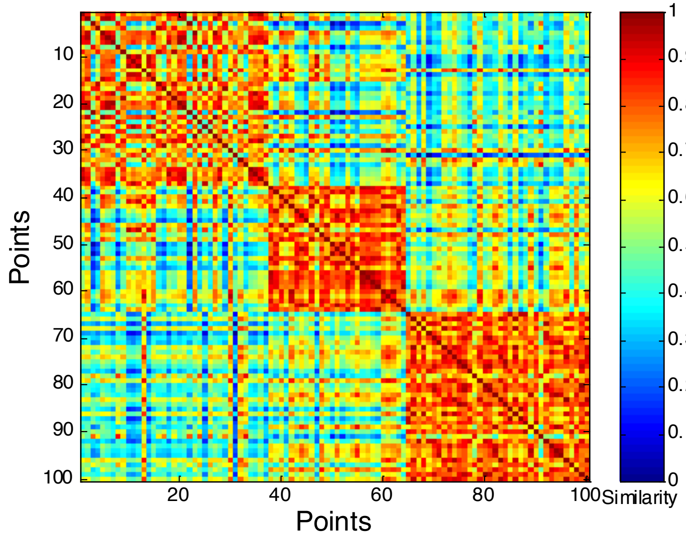
- Order the similarity matrix with respect to cluster labels and inspect visually

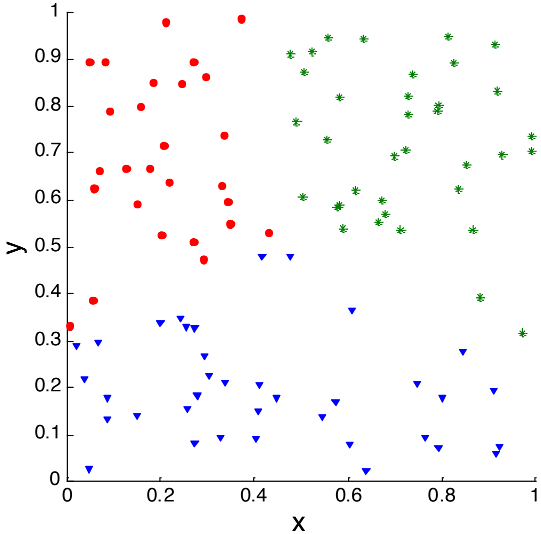




Using Similarity Matrix for Cluster Validation

- Clusters in random data are not so crisp

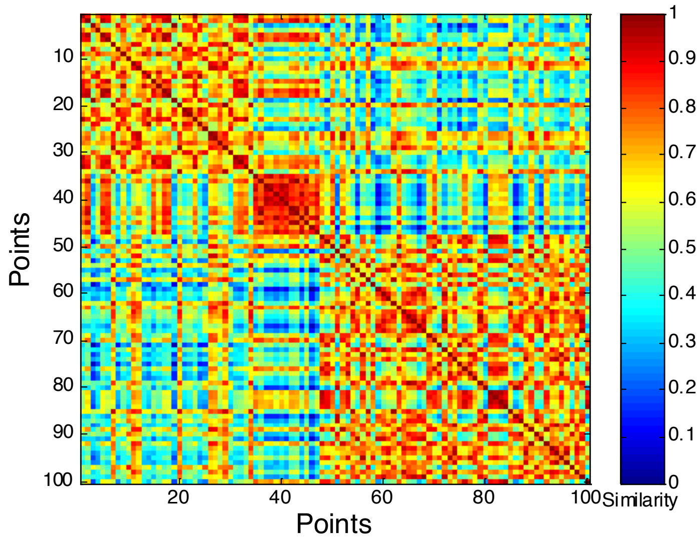


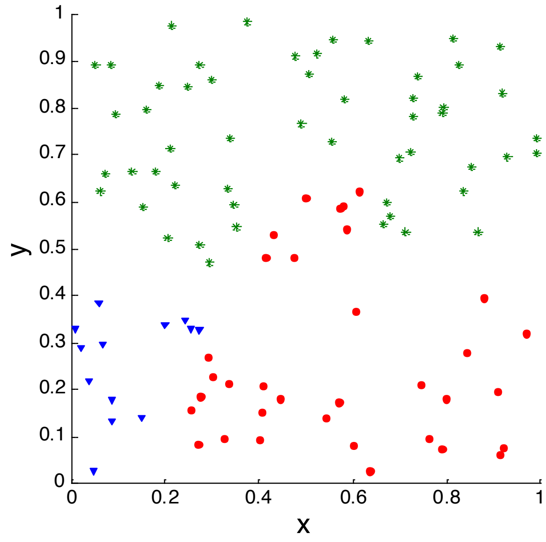


K-means

Using Similarity Matrix for Cluster Validation

- Clusters in random data are not so crisp





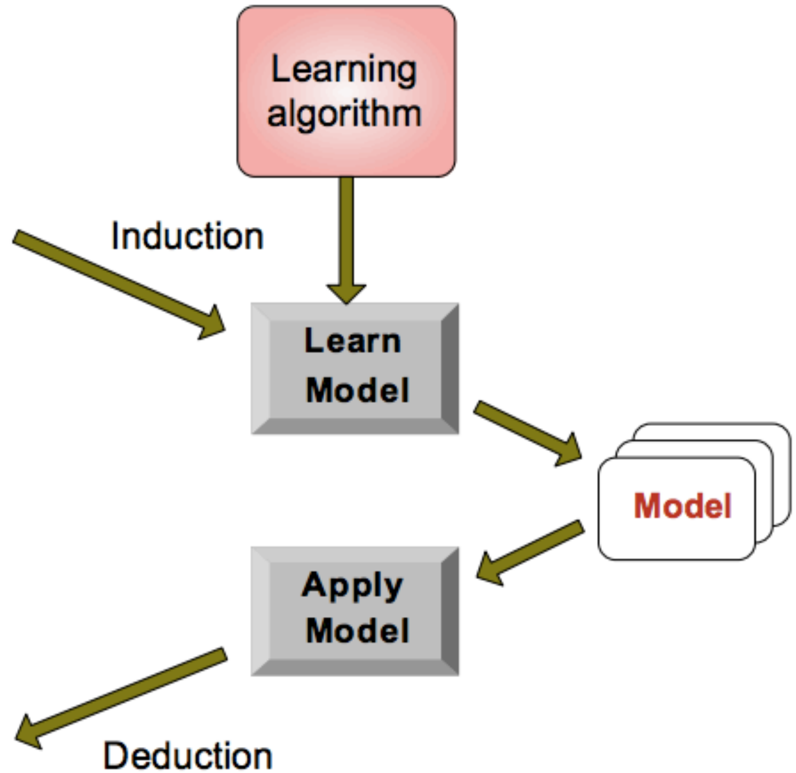
Agglomerative Hierarchical Clustering - MAX

Problem with Unlabeled Data

- Don't have labeled data like supervised learning

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

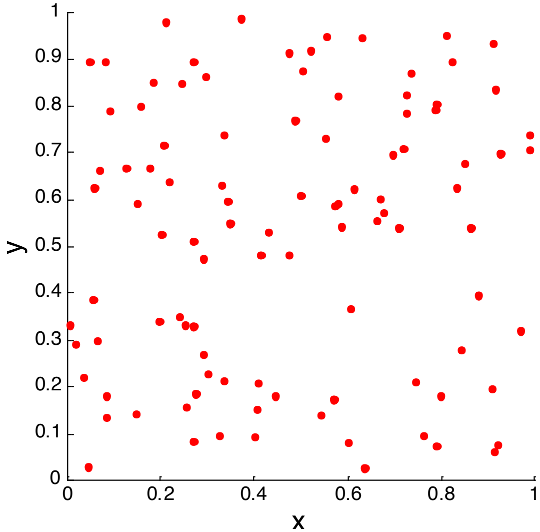
Image taken from https://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap4_basic_classification.pdf

Need for Validation

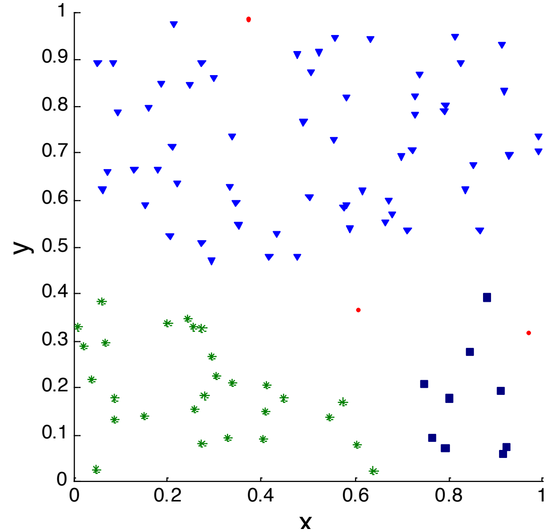
- Want to evaluate "goodness" of resulting clusters
  - When clustering is used for summarization
    - Max compression, use SSE or similar
  - When clustering is used for understanding
    - More complicated, more subjective
- Reasons:
  - Avoid finding patterns in noise
  - Compare clustering algorithms
  - Compare two sets of clusters
  - Compare two clusters

Clusters Found in Random Data

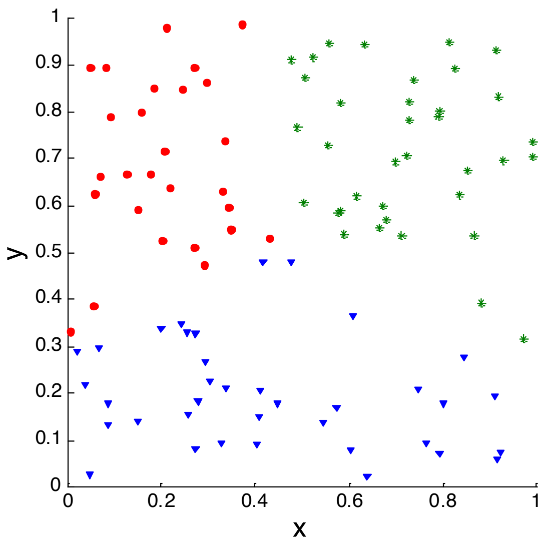
Random Points



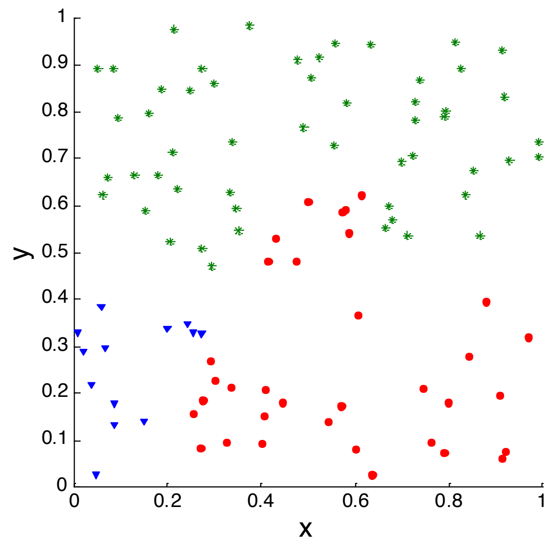
DBSCAN



K-means



Agglomerative Hierarchical Clustering - MAX



All clustering algorithms will find clusters (but are these meaningful?)

Issues for Cluster Validation

- Determine the clustering tendency of data
  - Whether non-random structure exists
- Determine the correct number of clusters
- Evaluate how well results of a cluster analysis fit the data without reference to external info (e.g., correlation)
- Compare the results of a cluster analysis to externally known results (i.e., known class labels)
- Compare two sets of clusters to determine which is better

Issues for Cluster Validation

- Determine the clustering tendency of data
- Determine the correct number of clusters
- Evaluate how well results of a cluster analysis fit the data without reference to external info
- Compare the results of a cluster analysis to externally known results
- Compare two sets of clusters to determine which is better

Unsupervised techniques that do not reference external info

Issues for Cluster Validation

- Determine the clustering tendency of data
- Determine the correct number of clusters
- Evaluate how well results of a cluster analysis fit the data without reference to external info
- Compare the results of a cluster analysis to externally known results
- Compare two sets of clusters to determine which is better

Supervised technique

Issues for Cluster Validation

- Determine the clustering tendency of data
- Determine the correct number of clusters
- Evaluate how well results of a cluster analysis fit the data without reference to external info
- Compare the results of a cluster analysis to externally known results
- Compare two sets of clusters to determine which is better

Can be either supervised or unsupervised

Issues for Cluster Validation

- Determine the clustering tendency of data
- Determine the correct number of clusters
- Evaluate how well results of a cluster analysis fit the data without reference to external info
- Compare the results of a cluster analysis to externally known results
- Compare two sets of clusters to determine which is better

Can be applied to individual clusters or the entire clustering

Types of Evaluation Measures

- Unsupervised
  - Measures goodness of clustering with no external info
  - Can measure cluster cohesion or cluster separation
  - E.g. SSE, silhouette coefficient
- Supervised
  - Measures extent of clustering results matching to some external structure
  - E.g. entropy
- Relative
  - Compares different clusterings
  - E.g. compares two k-means clusterings via SSE or entropy

Key Ideas

- No (easy) right answer to cluster validation unless external data is available
- Choosing k
  - Elbow method
  - Silhouette method
- Cluster validation
  - Need for a framework to interpret evaluation measure
  - Choice of measure depends on
    - Whether the goal is to understand vs summarize data
    - Whether external information is available
  - Still many open questions in this area