PARSER EVALUATION

Emad Gohari & Ehsan Omidi

Intro to Computational Linguistics

Winter 2015

York University

Parsers (Intro)

- Parsing is the process of analyzing a string or text into logical syntactic components.
- Parsing is conducted in order to test conformability to a logical grammar
- Parsing often results in a parse tree showing the syntactic relation of the different constituents of a string or text to each other.
- The input to a parser is often text in some computer language, but may also be text in a natural language or less structured textual data

Types of Parsers

- **Top-Down Parsing:**
  - Top-down parsing is a parsing strategy where one first looks at the highest level of the parse tree and works down the parse tree
- **Bottom-Up Parsing:**
  - Bottom-up parsing identifies and processes the text's lowest-level small details first, before its mid-level structures, and leaving the highest-level overall structure to last

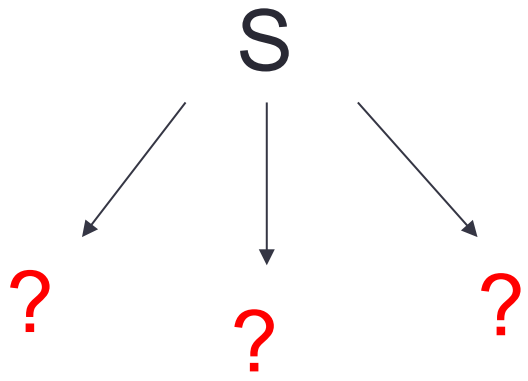
Top Down Parsing

- Example grammar:

  S → xyz | aBC

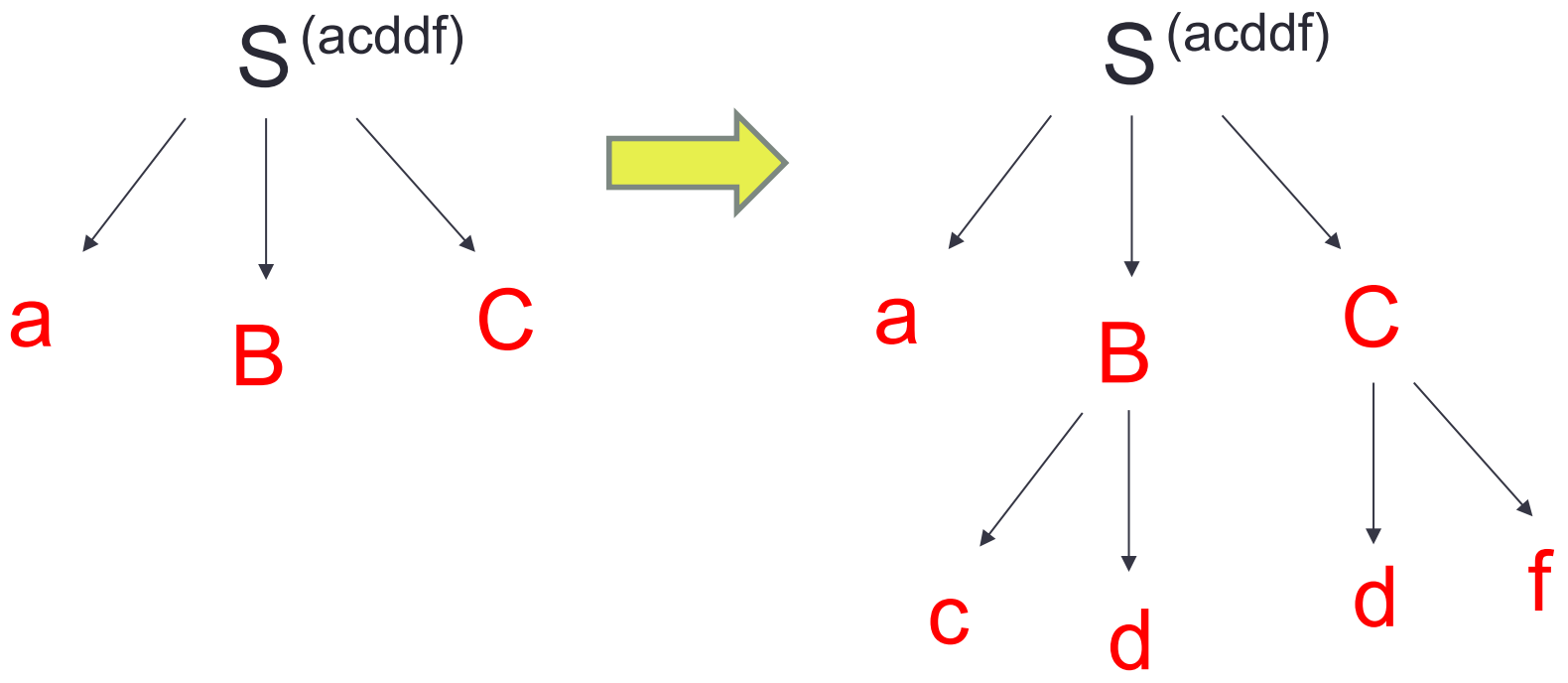
  B → c | cd

  C → eg | df



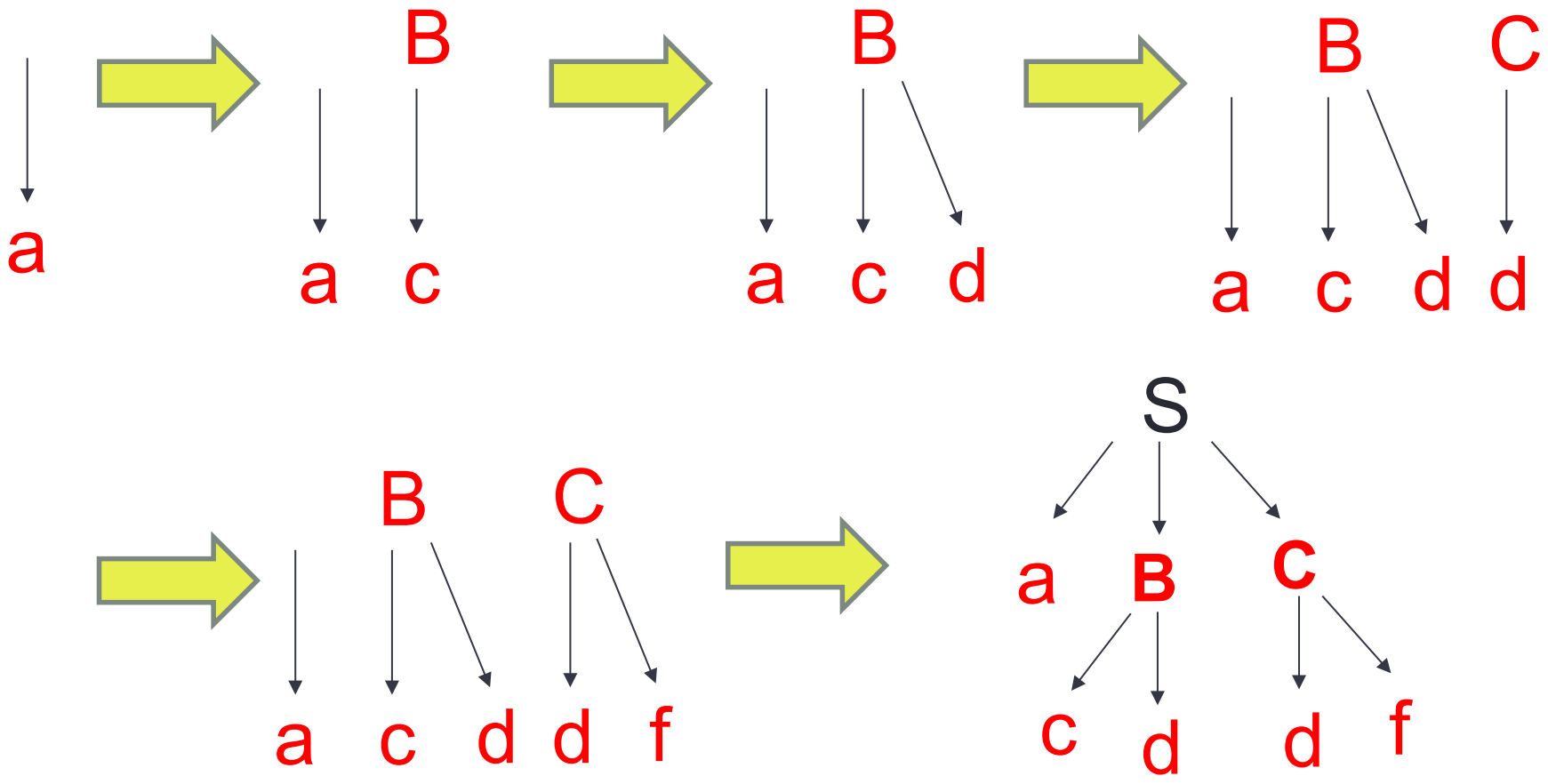
- We want to parse the string: acddf

Top Down Parsing



Bottom Down Parsing

- We want to parse the string: acddf



Parser Evaluation

- Because of the wide variety of contemporary practices used in the automatic syntactic parsing of natural languages, it has become necessary to analyze and evaluate the strengths and weaknesses of different approaches
- Parsing is not usually a goal in itself, but a parser is used as a component of NLP and artificial intelligence (AI) systems.
- An evaluation method defines the way in which the performance of a parser may be quantified

Parser Evaluation

- The “easy” measures for evaluation:
  - **Accuracy**: how many sentences are correctly parsed
  - **Coverage**: how many sentences can be parsed by the grammar
  - **n-best accuracy**: for how many sentences is the correct parse among the n best parses
- PARSEVAL: Workshop in 1991 to decide how to evaluate parsers
- PARSEVAL and used more “exact” measures by looking at single constituents

PARSEVAL Measures

- **Correct Constituent**: A constituent which has the correct yield
- **Unlabeled:**
  - **Precision**: number of correct constituents (yield) in parser output divided by number of constituents in the parser output
  - **Recall**: number of constituents from the gold standard (yield) that can be found in the parser output divided by the number of constituents in the gold standard

PARSEVAL Measures

- **Labeled:**
  - **Labeled Precision**: percentage of correct constituents (yield + label) in parser output
  - **Labeled Recall**: percentage of constituents from the gold standard (yield + label) that can be found in the parser output
- **F-Score**: $F\downarrow\beta = (\beta\uparrow 2 + 1) \times precision \times recall / \beta\uparrow 2 \times percision + recall$

TEXT CLUSTERING

Emad Gohari & Ehsan Omidi

Intro to Computational Linguistics

Winter 2015

York University

Topics

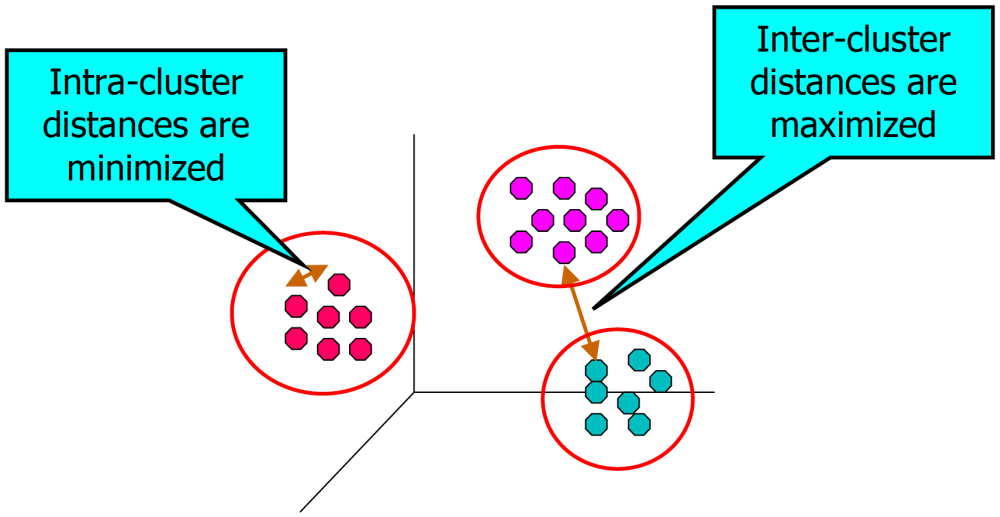
- Definition of clustering
- Types of clustering
- Types of clusters
- Measures of (Dis)Similarity
- K-means algorithm
- Hierarchical clustering
- Divisive and agglomerative clustering
- Inter-cluster similarity measures
- Evaluation of clustering
- Applications of text clustering
- Some examples of text clustering
- Conclusion

Topics

- Definition of clustering
- Types of clustering
- Types of clusters
- Measures of (Dis)Similarity
- K-means algorithm
- Hierarchical clustering
- Divisive and agglomerative clustering
- Inter-cluster similarity measures
- Evaluation of clustering
- Applications of text clustering
- Some examples of text clustering
- Conclusion

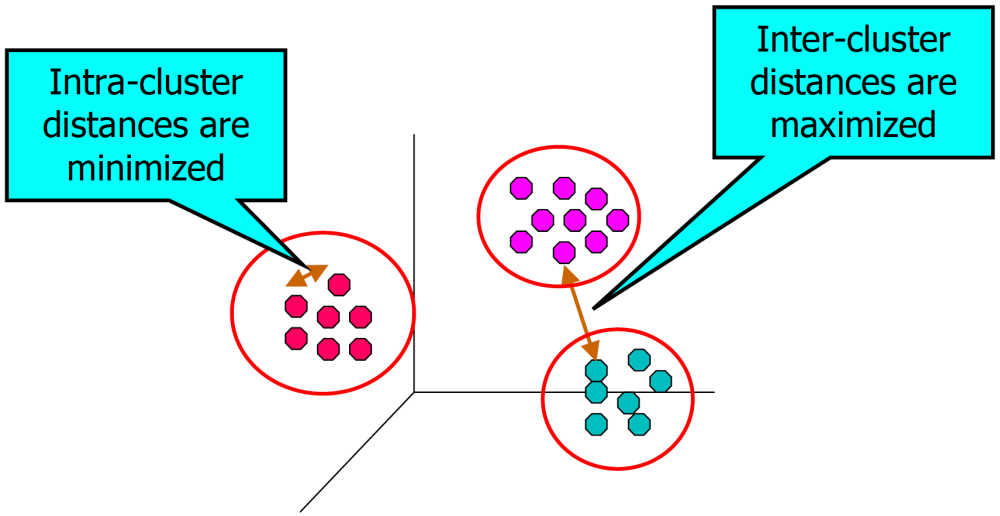
Clustering Problem

- The clustering problem is defined to be that of finding groups (clusters) of similar objects in the data
- Groups should be specified in a way to maximize similarity of objects in a group and minimize the similarity of objects in different groups



Clustering Problem

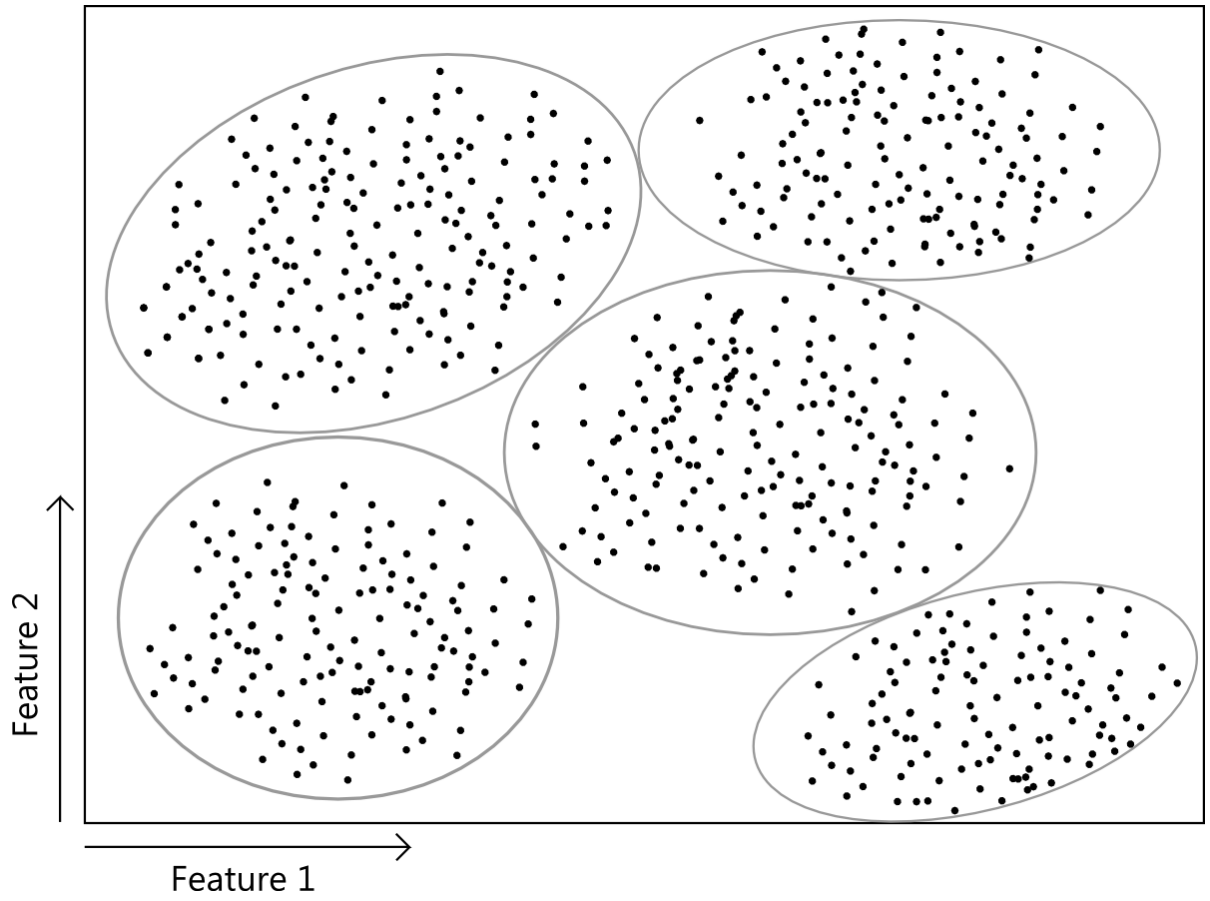
- The similarity between the objects is measured with the use of a **similarity function**
- Or the difference between the objects is measured with the use of a **distance function**
- Clustering is an **unsupervised** learning problem (there is no predefined cluster or group like in classification)



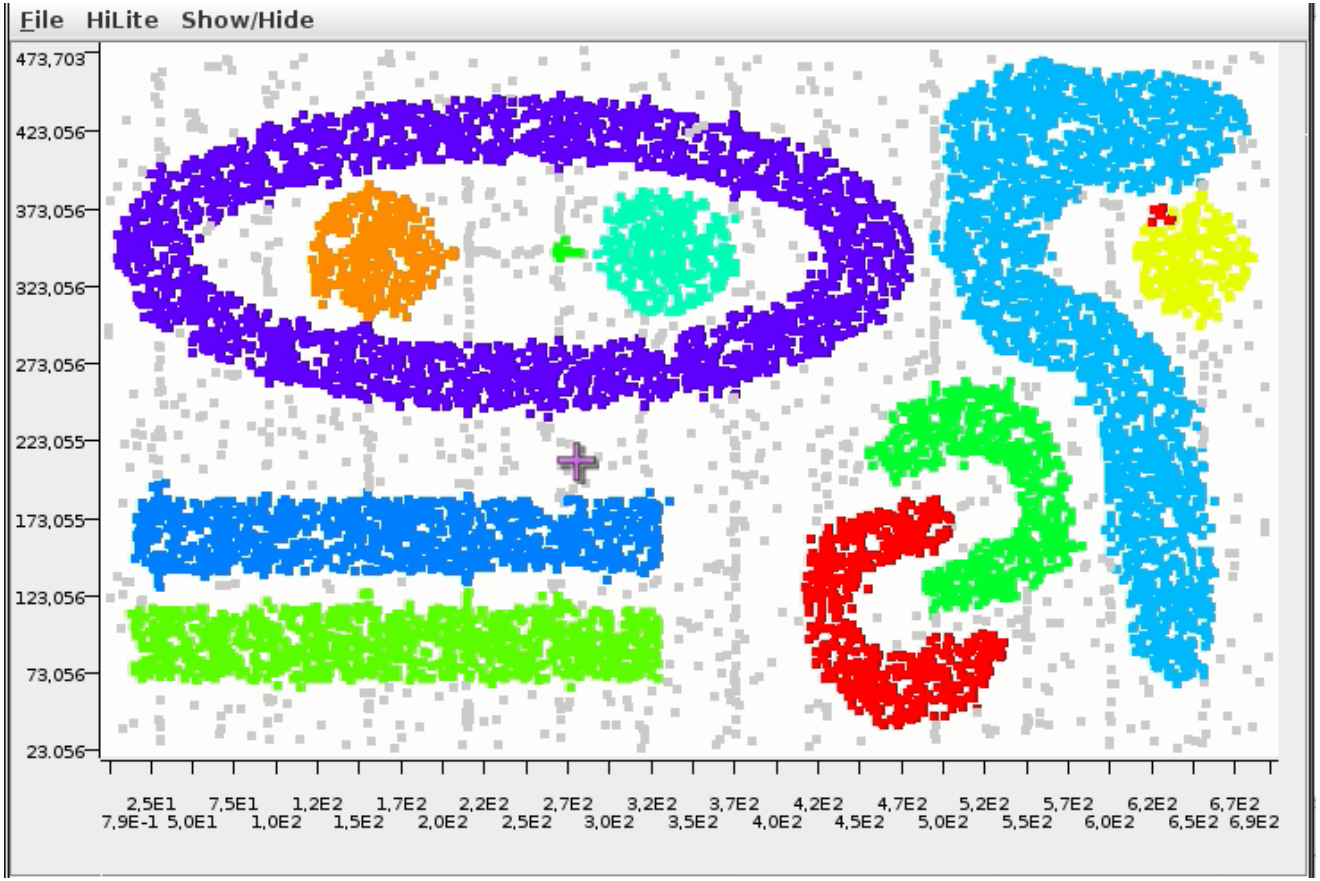
Clustering Problem

- Widely studied in the database and statistics literature in the context of a wide variety of data mining tasks
- Can be very useful in context of text domain
- When the objects can be of different granularities …
  - Documents, paragraphs, sentences, terms
- Specially useful for organizing documents to improve retrieval and support browsing

Examples of Clustering



Examples of Clustering



Topics

- Definition of clustering
- Types of clustering
- Types of clusters
- Measures of (Dis)Similarity
- K-means algorithm
- Hierarchical clustering
- Divisive and agglomerative clustering
- Inter-cluster similarity measures
- Evaluation of clustering
- Applications of text clustering
- Some examples of text clustering
- Conclusion

Different Types of Clustering

Hierarchical vs. partitional (nested and unnested)

- **Hierarchical** methods produce a **nested** sequence of partitions
- With a single, all inclusive cluster at the top and singleton clusters of individual points at the bottom.

Different Types of Clustering

Hierarchical vs. partitional (nested and unnested)

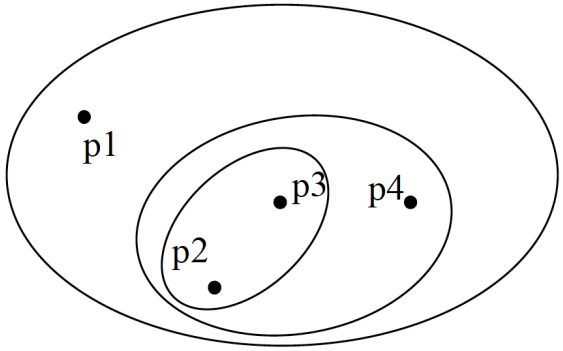


Figure 9a. Traditional nested set.

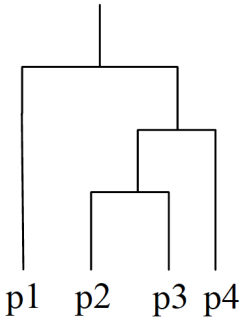


Figure 9b. Traditional dendogram

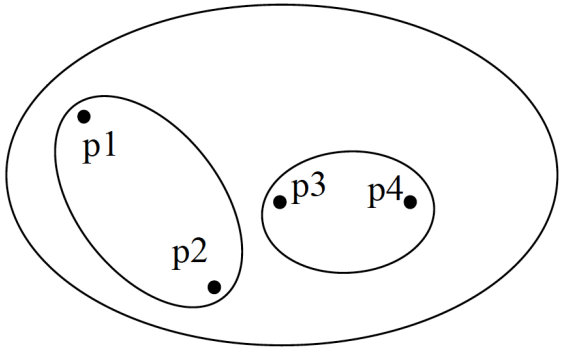


Figure 9c. Non-traditional nested set

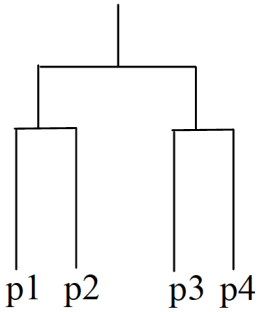


Figure 9d. Non-traditional dendogram.

Different Types of Clustering

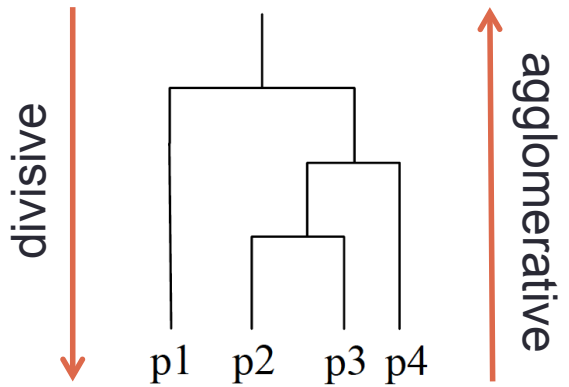
Hierarchical vs. partitional (nested and unnested)

- **Partitional** methods create a one-level unnested partitioning of the data points.
- If K is the desired number of clusters, then partitional approach typically find all K clusters at once.

Different Types of Clustering

Divisive vs. agglomerative

- Hierarchical clustering techniques proceed either from top to the bottom or from the bottom to the top
- **Divisive**: method starts with one large cluster and splits it
- **Agglomerative**: starts with clusters each containing a point, and then merges them



Different Types of Clustering

Incremental or non-incremental

- Some clustering techniques work with an item at a time and decide how to cluster it given the current set of points that have already been processed. (incremental)
- Other techniques use information about all the points at once. (non-incremental)
  - Non-incremental clustering algorithms are far more common.
- Exclusive vs. non-exclusive
- Fuzzy vs. non-fuzzy
- Partial vs. complete
- And …

Different Types of Clustering

Objective functions: clustering as an **optimization problem**

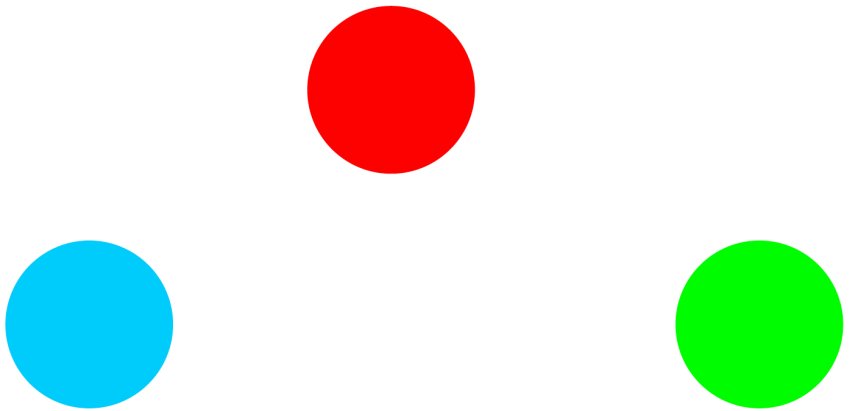
- Many clustering techniques are based on trying to minimize or maximize an objective function.
- The clustering problem then becomes an optimization problem
- In theory, can be solved by enumerating all possible ways of dividing the points into clusters and evaluating the "goodness" of each potential set of clusters using the objective function
- This "exhaustive" approach is computationally infeasible (NP complete)
- Therefore, **more practical** techniques for optimizing a global objective function have been developed.

Topics

- Definition of clustering
- Types of clustering
- Types of clusters
- Measures of (Dis)Similarity
- K-means algorithm
- Hierarchical clustering
- Divisive and agglomerative clustering
- Inter-cluster similarity measures
- Evaluation of clustering
- Applications of text clustering
- Some examples of text clustering
- Conclusion

Types of Clusters

- Well-separated clusters
  - Cluster: set of objects such that any point in a cluster is closer (more similar) to every other point in the cluster than to any point not in the cluster.



3 well-separated clusters

Types of Clusters

- Center-based
  - Cluster: set of objects such that an object in a cluster is closer (more similar) to the “center” of a cluster, than to the center of any other cluster.
  - The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most “representative” point of a cluster

4 center-based clusters

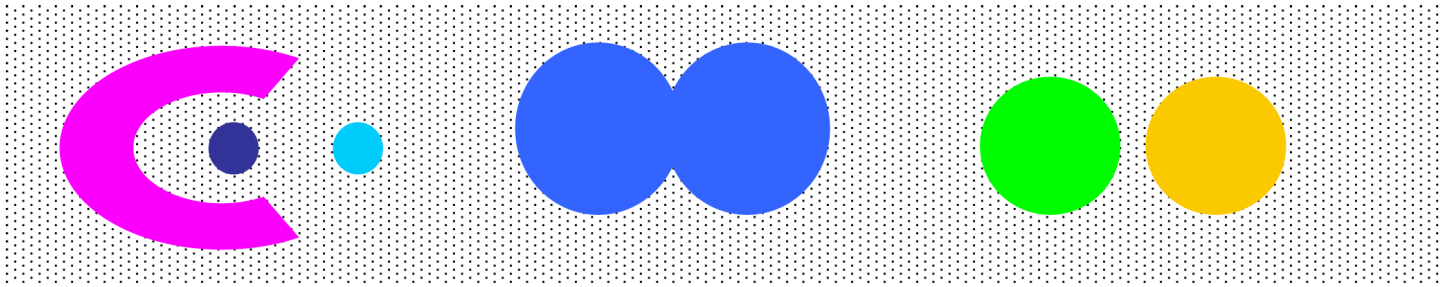
Types of Clusters

- Contiguous cluster (nearest neighbor or transitive)
  - Cluster: set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.

8 contiguous clusters

Types of Clusters

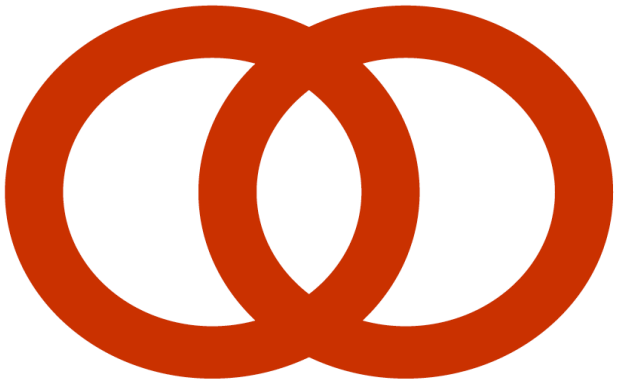
- Density-based
  - Cluster: a dense region of points, which is separated by low-density regions, from other regions of high density.
  - Used when the clusters are irregular or intertwined, and when noise and outliers are present.



6 density-based clusters

Types of Clusters

- Shared Property or Conceptual Clusters
  - Finds clusters that share some common property or represent a particular concept.



2 Overlapping Circles

Topics

- Definition of clustering
- Types of clustering
- Types of clusters
- Measures of (Dis)Similarity
- K-means algorithm
- Hierarchical clustering
- Divisive and agglomerative clustering
- Inter-cluster similarity measures
- Evaluation of clustering
- Applications of text clustering
- Some examples of text clustering
- Conclusion

Measures of (Dis)Similarity

- There are two main types of measures to determine whether two objects are similar or dissimilar
  - Distance measures
  - Similarity measures
- Many clustering methods use distance measures
- Distance between two objects x_i and x_j: $d(x_i, x_j)$
- A valid distance measure should be
  - Symmetric
  - Obtain minimum value (usually zero) in case of identical objects

Minkowski Distance Measure

- A distance measure for **numeric** attributes
- Given two p-dimensional instances x_i and x_j

$$d(x_i, x_j) = (|x_{i1} - x_{j1}|^g + |x_{i2} - x_{j2}|^g + \ldots + |x_{ip} - x_{jp}|^g)^{1/g}$$

- Euclidean distance achieved when g=2
- Given g=1, Manhattan distance
- Given g=∞, Chebychev distance
- If data is normalized, each variable can be assigned with a weight according to its importance (weighted distance)

$$d(x_i, x_j) = (w_1 |x_{i1} - x_{j1}|^g + w_2 |x_{i2} - x_{j2}|^g + \ldots + w_p |x_{ip} - x_{jp}|^g)^{1/g}$$

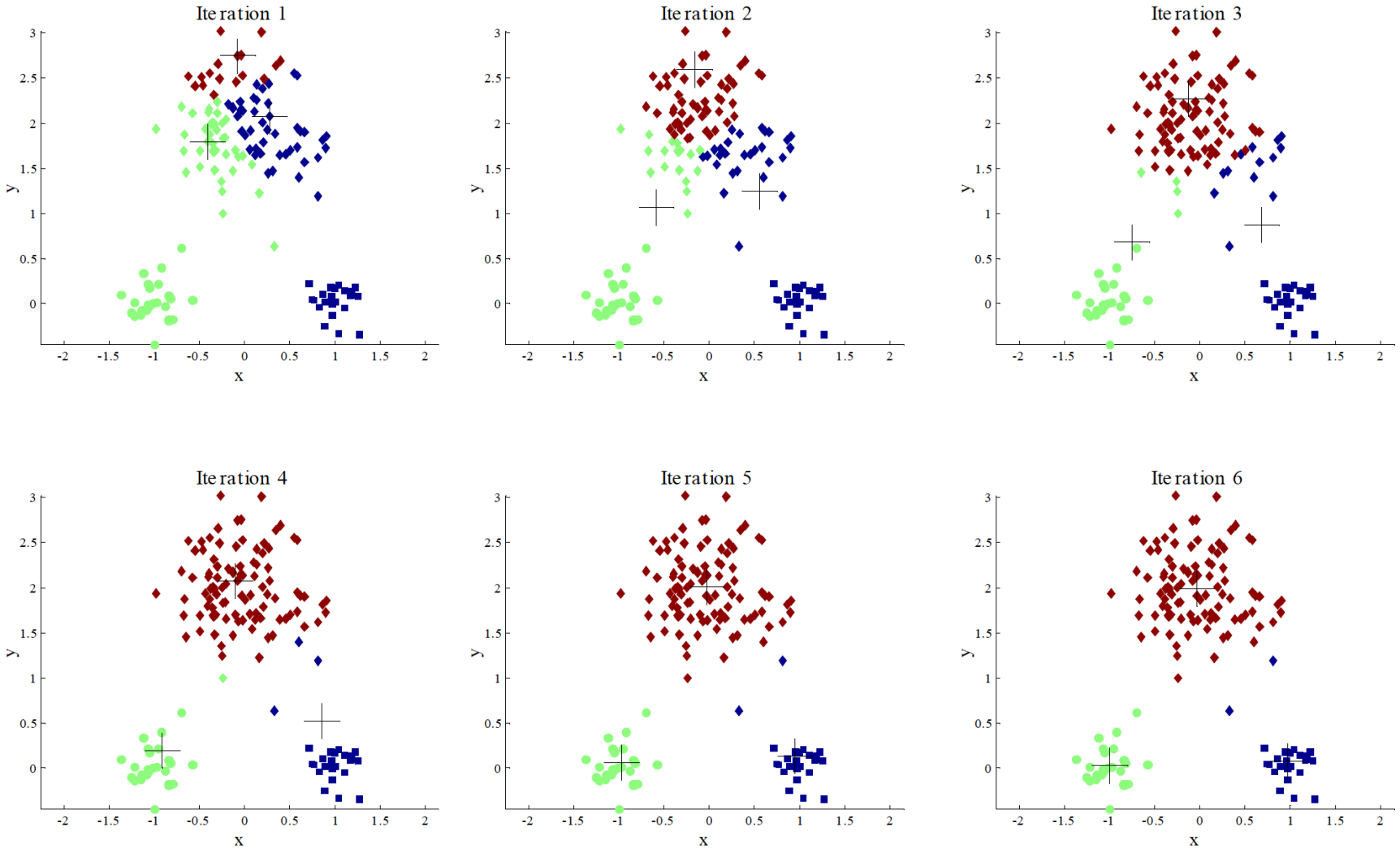
Topics

- Definition of clustering
- Types of clustering
- Types of clusters
- Measures of (Dis)Similarity
- K-means algorithm
- Hierarchical clustering
- Divisive and agglomerative clustering
- Inter-cluster similarity measures
- Evaluation of clustering
- Applications of text clustering
- Some examples of text clustering
- Conclusion

K-means Clustering

- It is a partitional clustering technique
- Based on the idea that a center point can represent a cluster (center-based)
- Basic K-means algorithm for finding K clusters
  1. Select K points as the initial centroids.
  2. Assign all points to the closest centroid.
  3. Recompute the centroid of each cluster.
  4. Repeat steps 2 and 3 until the centroids don't change.
- 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.

K-means Clustering



K-means Clustering

- "Closeness" is measured by Euclidean distance, cosine similarity, correlation, etc.
- 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.

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

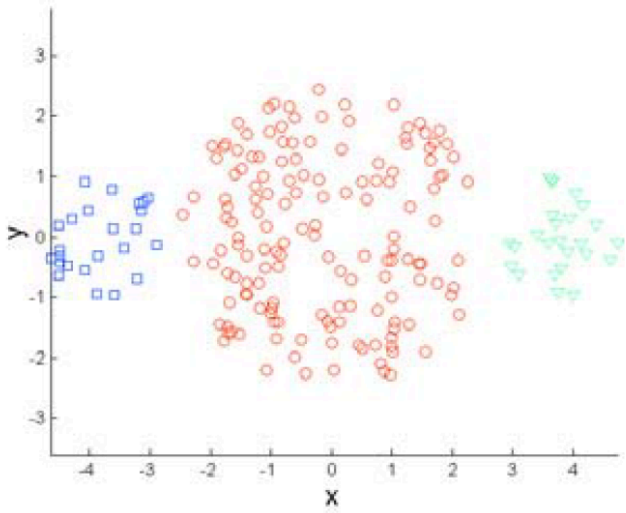
  - x is a data point in cluster C_i and m_i s the representative point for cluster C_i
  - One easy way to reduce SSE is to increase K
  - A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

K-means Clustering Problems

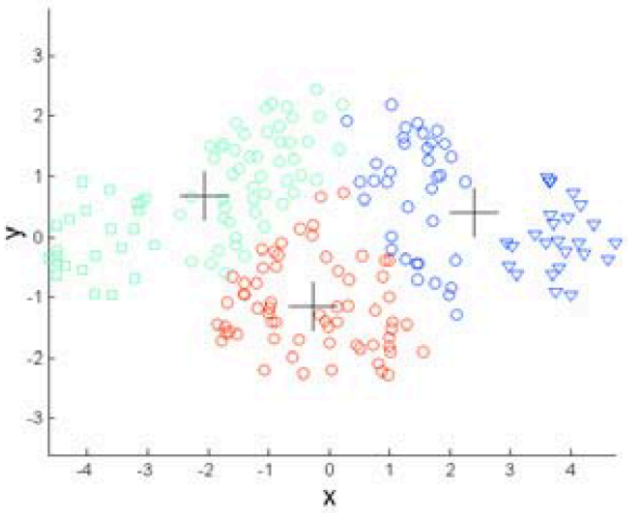
- Initial centroid problem
  - Multiple runs
  - Sample and use hierarchical clustering to determine initial centroids
  - Select more than K initial centroids (at the end select ones most widely separated)
- Basic k-means can yield empty clusters
  - Choose the point that contributes most to SSE
  - Choose a point from the cluster with the highest SSE
  - Update the centroids after each assignment (incremental approach)

Limitations of K-means

- Clusters with different sizes



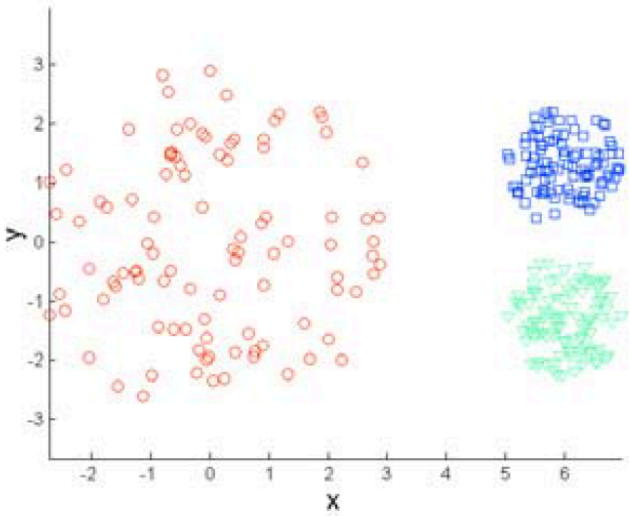
Original Points



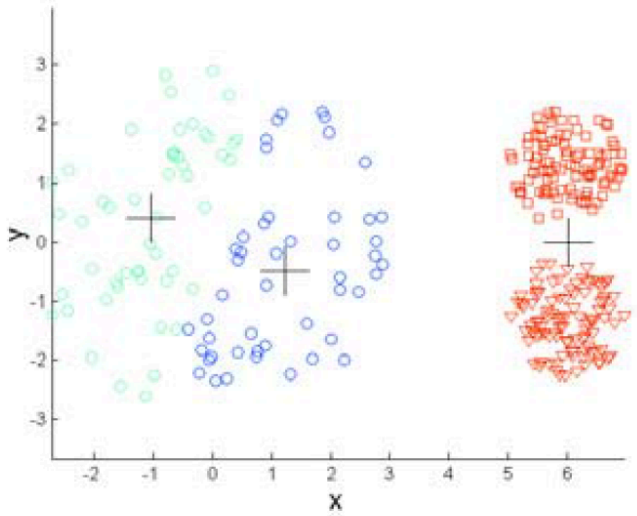
K-means (3 Clusters)

Limitations of K-means

- Clusters with different densities



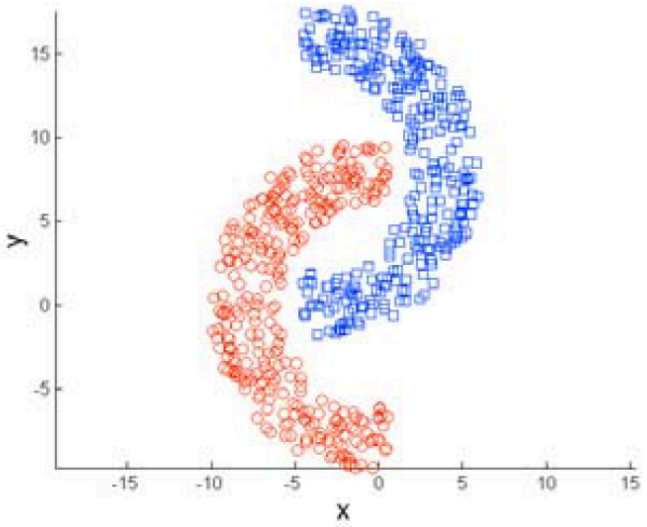
Original Points



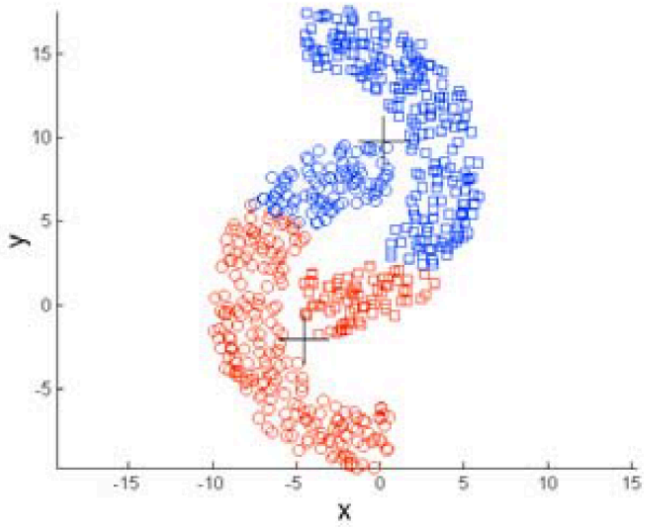
K-means (3 Clusters)

Limitations of K-means

- Non-globular shapes



Original Points



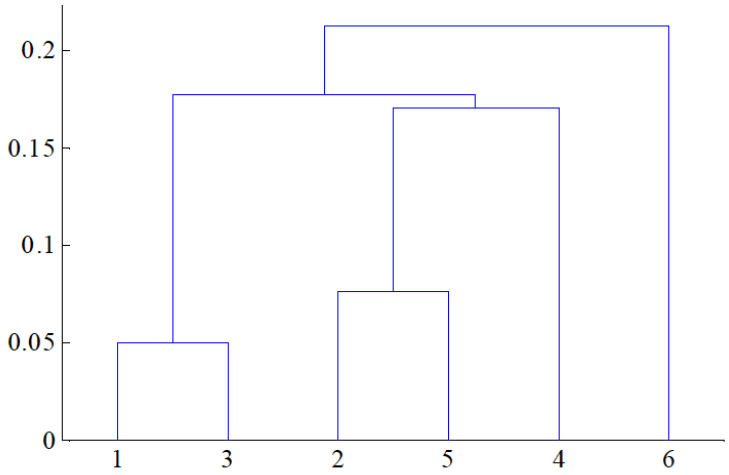
K-means (2 Clusters)

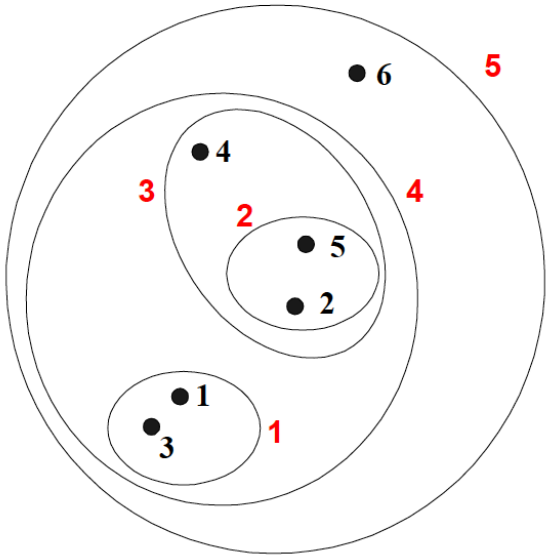
Topics

- Definition of clustering
- Types of clustering
- Types of clusters
- Measures of (Dis)Similarity
- K-means algorithm
- Hierarchical clustering
- Divisive and agglomerative clustering
- Inter-cluster similarity measures
- Evaluation of clustering
- Applications of text clustering
- Some examples of text clustering
- Conclusion

Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram





Strengths of Hierarchical Clustering

- Do not have to assume any particular number of clusters
- Any desired number of clusters can be obtained by "cutting" the dendrogram at the proper level
- The hierarchy may correspond to meaningful taxonomies
  - Like in biological sciences

Types of Hierarchical Clustering

- Two main types of hierarchical clustering
  - Agglomerative: bottom-up
  - Divisive: top-down
- Traditional hierarchical algorithms use a similarity or distance matrix in order to merge or split one cluster at a time
- Agglomerative approach is a more popular hierarchical clustering technique

Topics

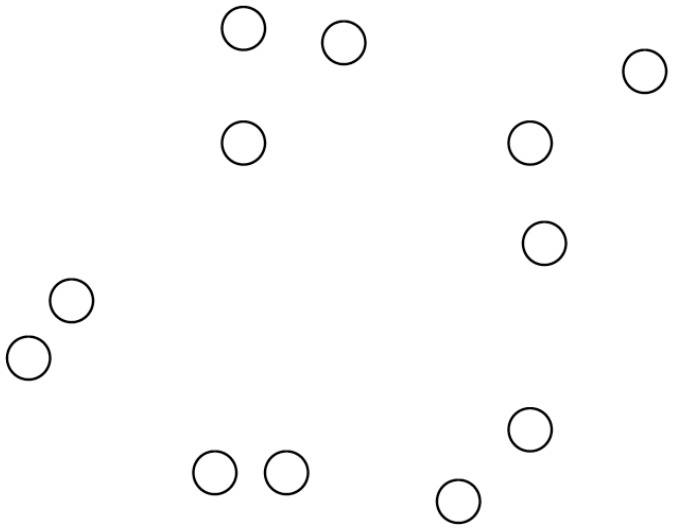
- Definition of clustering
- Types of clustering
- Types of clusters
- Measures of (Dis)Similarity
- K-means algorithm
- Hierarchical clustering
- Divisive and agglomerative clustering
- Inter-cluster similarity measures
- Evaluation of clustering
- Applications of text clustering
- Some examples of text clustering
- Conclusion

Agglomerative Clustering

- Basic algorithm is
  1. Compute the proximity matrix
  2. Let each data point be a cluster
  3. **Repeat**
  4. Merge the two closest clusters
  5. Update the proximity matrix
  6. **Until** only a single cluster remains
- Key operation is the computation of the proximity of two clusters
- Different approaches to defining the distance between clusters distinguish the different algorithms

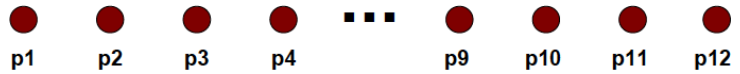
Agglomerative Clustering

- Start with cluster of individual points and a proximity matrix



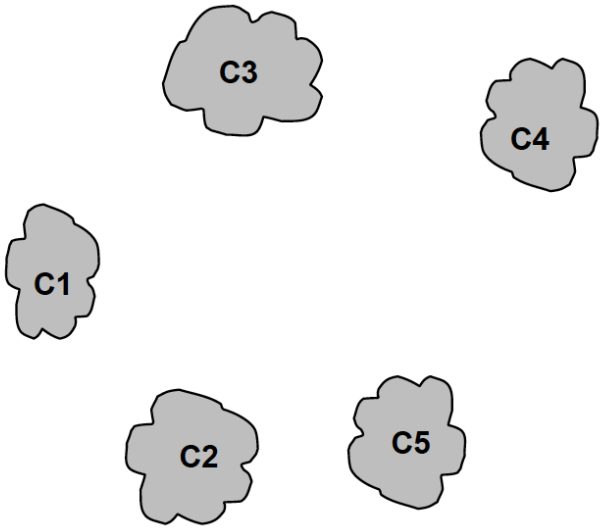
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix



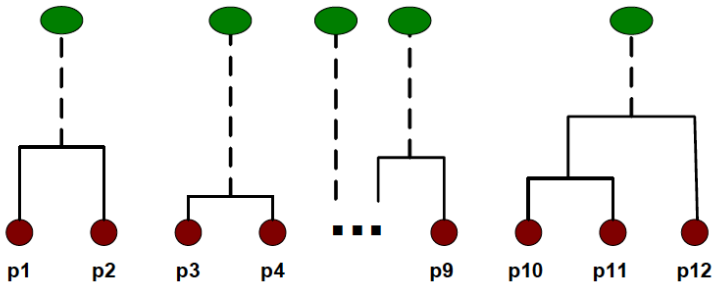
Agglomerative Clustering

- After some merging steps, we have some clusters



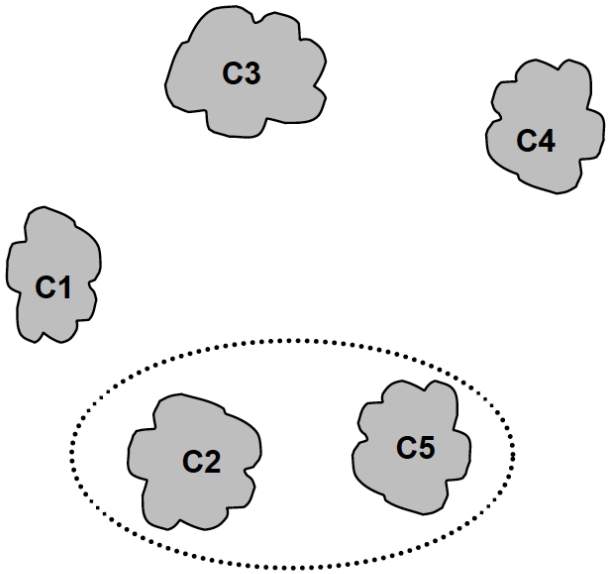
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



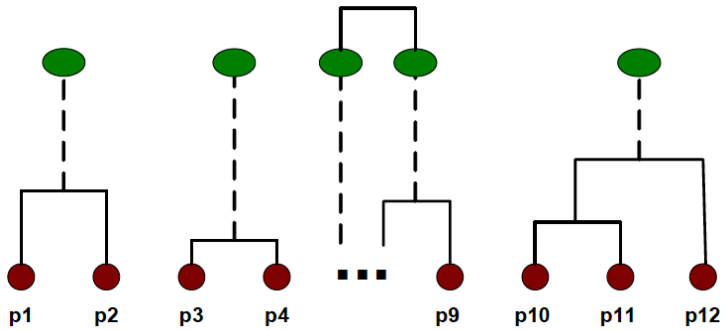
Agglomerative Clustering

- We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.



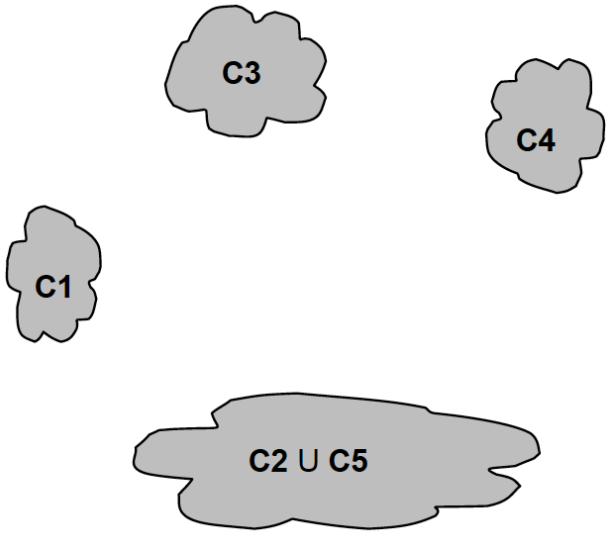
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



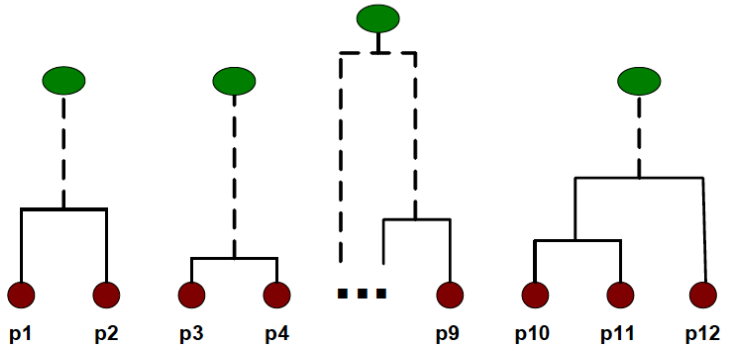
Agglomerative Clustering

- The question is “How do we update the proximity matrix?”



	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



Topics

- Definition of clustering
- Types of clustering
- Types of clusters
- Measures of (Dis)Similarity
- K-means algorithm
- Hierarchical clustering
- Divisive and agglomerative clustering
- Inter-cluster similarity measures
- Evaluation of clustering
- Applications of text clustering
- Some examples of text clustering
- Conclusion

Inter-Cluster Similarity

- MIN or Single Link: two most similar (nearest) points
  - Good: can handle non-elliptical shapes
  - Bad: sensitive to noise and outliers
- MAX or Complete Link: two most distant points
  - Good: less susceptible to noise and outliers
  - Good: trends to break large clusters
  - Bad: biased towards globular clusters

Inter-Cluster Similarity

- Group Average
  - Average of pairwise proximity between points in the two clusters

$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| * |\text{Cluster}_j|}$$

  - Compromise between Single and Complete Link
  - Good: less susceptible to noise and outliers
  - Bad: biased toward globular clusters

Inter-Cluster Similarity

- Ward's method
  - the proximity between two clusters is defined as the **increase in the squared error** that results when two clusters are **merged**.
  - Hierarchical analogue of K-means (can be used to initialize K-means)
  - Good: less susceptible to noise and outliers
  - Bad: biased towards globular clusters

Hierarchical Clustering Problems

- Once a decision is made to combine two clusters, it cannot be undone
  - prevents a **local optimization** criterion from becoming a **global optimization** criterion
- No objective function is directly minimized
  - use various criteria to decide “locally” at each step which clusters should be joined
- Different schemes have some of these weaknesses:
  - Sensitivity to noise and outliers
  - Difficulty handling different sized clusters and convex shapes
  - Breaking large clusters

Topics

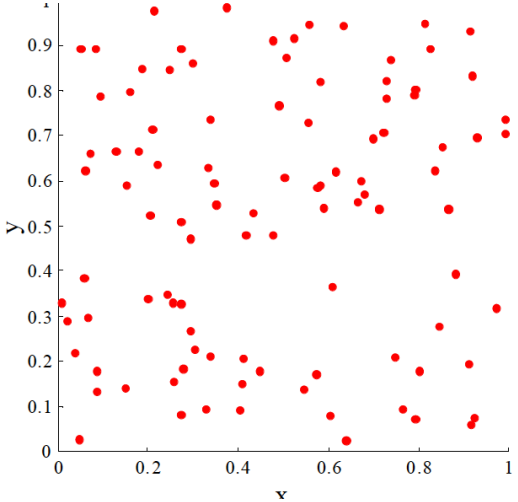
- Definition of clustering
- Types of clustering
- Types of clusters
- Measures of (Dis)Similarity
- K-means algorithm
- Hierarchical clustering
- Divisive and agglomerative clustering
- Inter-cluster similarity measures
- Evaluation of clustering
- Applications of text clustering
- Some examples of text clustering
- Conclusion

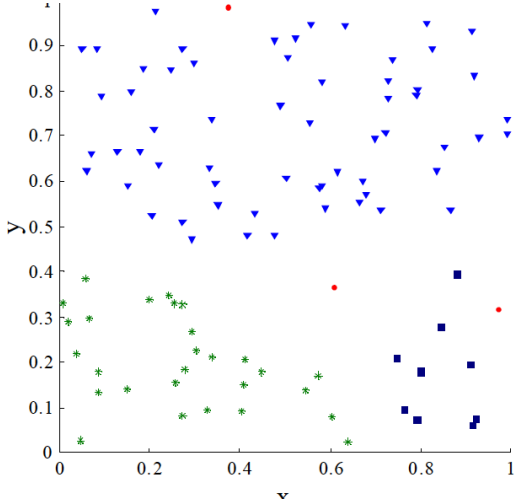
Cluster Validity

- For supervised classification we have a variety of measures to evaluate how good a model is
  - Accuracy, precision, recall, …
- Analogous question for cluster analysis: how to evaluate the “goodness” of the resulting clusters?
- Why do we want to evaluate them?
- To avoid finding patterns in noise
- To compare clustering algorithms
- To compare two sets of clusters
- To compare two clusters

Cluster Validity

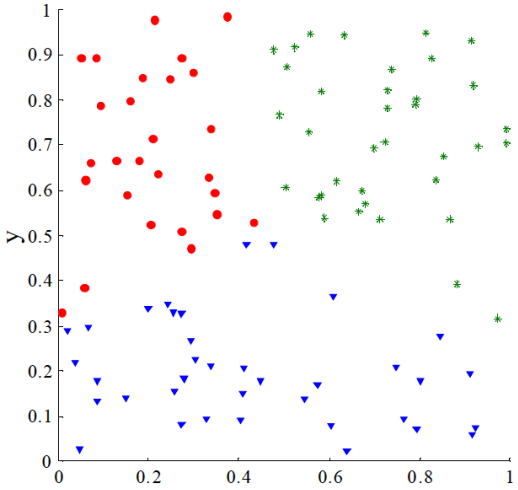
Random Points

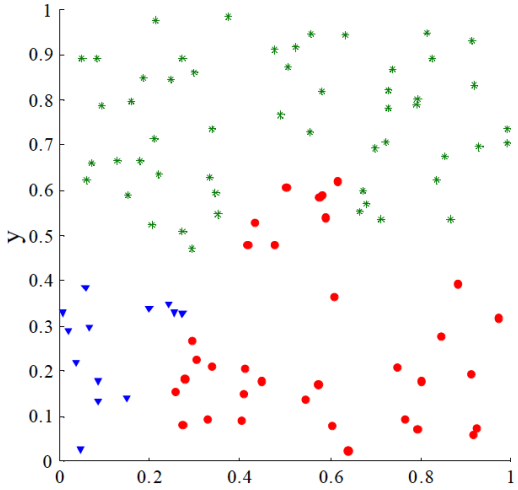




DBSCAN

K-means





Complete Link

Cluster Validity

- Three type of numerical measures to judge cluster validity
- External Index
  - Used to measure the extent to which cluster labels match externally supplied class labels: Entropy
- Internal Index
  - Used to measure the goodness of a clustering structure without respect to external information: Sum of Squared Error (SSE)
- Relative Index
  - Used to compare two different clustering methods or clusters.
  - Often an external or internal index is used for this function

Cluster Validity

- Some examples of internal measure
  - Correlation
  - Similarity matrix
  - SSE
  - Cluster Cohesion
  - Cluster Separation
  - Silhouette Coefficient
- Some examples of external measure
  - Entropy
  - Purity

Cluster Validity: External Measure

K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

Cluster Validity: Entropy

- Calculating the class distribution of the data for each cluster:
- for cluster j we compute p_{ij}
  - the probability that a member of cluster j belongs to class i
- $p_{ij} = m_{ij} / m_j$
- Where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j
- Using these class distributions the entropy of each cluster j is calculated using the standard formula.
  - L is number of classes

$$e_j = -\sum_{i=1}^{L} p_{ij} log_2 p_{ij}$$

Cluster Validity: Entropy

- The total entropy of a set of clusters is calculated as the sum of the entropies of each cluster weighted by its size
  - m_j is size of cluster j, K is number of clusters, m is total number of data points

$$e = \sum_{i=1}^{K} \frac{m_j}{m} e_j$$

- The more the clustering is similar to predefined classes, the smaller the entropy is.

Cluster Validity: Purity

- Using the definition for entropy, the purity of cluster j is given by $purity_j = \max p_{ij}$
- Overall purity of a clustering is

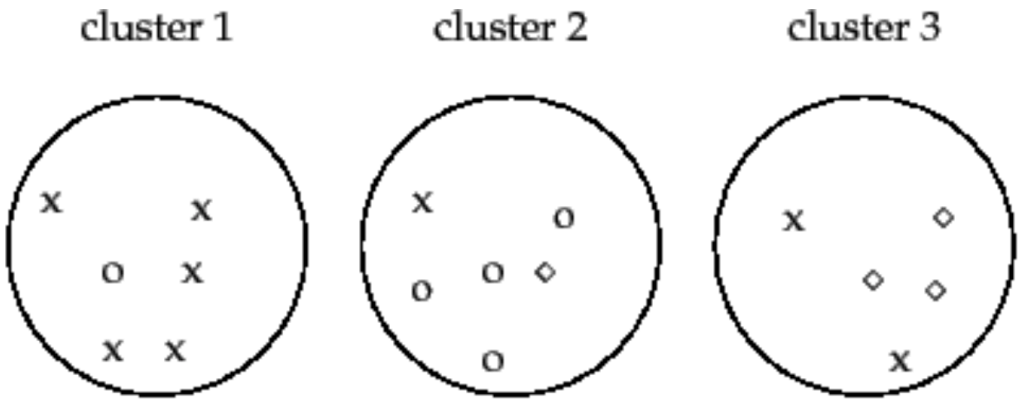
$$purity = \sum_{i=1}^{K} \frac{m_j}{m} purity_j$$

More intuition

- Each cluster is assigned to the class which is most frequent in the cluster
- Then the accuracy of this assignment is measured by counting the number of correctly assigned points and dividing by number of all data points

Cluster Validity: Purity

Example



- Majority class in clusters
  - Cluster 1: x,5
  - Cluster 2: o,4
  - Cluster 3: ‹›,3
- Purity = (1/17) × (5 + 4 + 3) ≈ 0.71

Topics

- Definition of clustering
- Types of clustering
- Types of clusters
- Measures of (Dis)Similarity
- K-means algorithm
- Hierarchical clustering
- Divisive and agglomerative clustering
- Inter-cluster similarity measures
- Evaluation of clustering
- Applications of text clustering
- Some examples of text clustering
- Conclusion

Applications of Text Clustering

- Document organization and browsing
  - The hierarchical organization of documents into coherent categories can be very useful for systematic browsing of the document collection
  - A classical example: Scatter/Gather method (1992)

Applications of Text Clustering

- Corpus summarization
  - Clustering techniques provide a coherent summary of the collection in the form of *cluster-digests* or *word-clusters*, which can be used in order to provide summary insights into the overall content of the underlying corpus.
- Variants of such methods, especially sentence clustering, can also be used for document summarization.

Applications of Text Clustering

- Document Classification
  - While clustering is inherently an unsupervised learning method, it can be leveraged to improve the quality of the results in supervised methods like classification.
  - In particular, *word-clusters* and *co-training methods* can be used in order to improve the classification accuracy of supervised applications with the use of clustering techniques.

Topics

- Definition of clustering
- Types of clustering
- Types of clusters
- Measures of (Dis)Similarity
- K-means algorithm
- Hierarchical clustering
- Divisive and agglomerative clustering
- Inter-cluster similarity measures
- Evaluation of clustering
- Applications of text clustering
- Some examples of text clustering
- Conclusion

Text Clustering Example

Document clustering

- Documents are represented using the vector-space model
- Each document considered to be a vector d in the term-space
- $d_{tf} = (tf_1, tf_2, \ldots, tf_n)$, tf_i is the weight of each term in document (normalized TF/IDF weights)
- The similarity between two documents can be measured by cosine measure (most common measure)

$$cosine(\boldsymbol{d}_1, \boldsymbol{d}_2) = (\boldsymbol{d}_1 \bullet \boldsymbol{d}_2) / \|\boldsymbol{d}_1\| \|\boldsymbol{d}_2\|$$

- One of the clustering algorithms like K-means then, can be used for clustering documents

Topics

- Definition of clustering
- Types of clustering
- Types of clusters
- Measures of (Dis)Similarity
- K-means algorithm
- Hierarchical clustering
- Divisive and agglomerative clustering
- Inter-cluster similarity measures
- Evaluation of clustering
- Applications of text clustering
- Some examples of text clustering
- Conclusion

Conclusion

- There are lots of clustering algorithms that are studied and improved by research.
- These algorithms are widely used in different fields for various data mining tasks.
- Most of these algorithms are applicable to text data
- Some optimizations required for processing sparse vectors that are common in text data
- There are lot more to cover but our time was limited!

Thank You!

Questions …