

CSCI 5622 Machine Learning

ML K-Means Clustering

DATE	READ	DUE
Today, Oct 21	<i>K-means</i>	Project Questions
Mon, Oct 26	Mixture of Gaussians	Peer Fdbk Grades
Wed, Oct 28	EM	Project Questions

www.RodneyNielsen.com/teaching/CSCI5622-F09/

Instructor: Rodney Nielsen

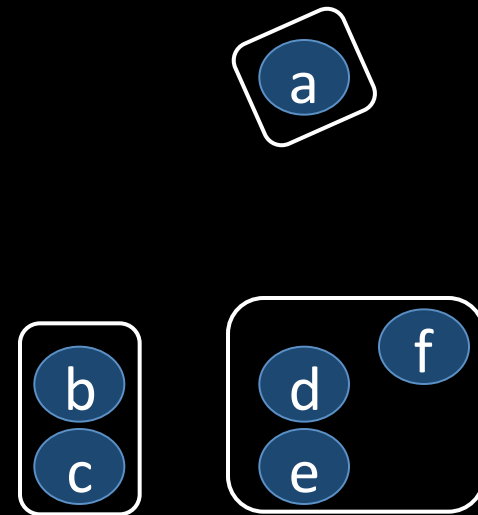
Assistant Professor Adjunct, CU Dept. of Computer Science

Research Assistant Professor, DU, Dept. of Electrical & Computer Engr.

Research Scientist, Boulder Language Technologies

ML Clustering / Unsupervised Learning

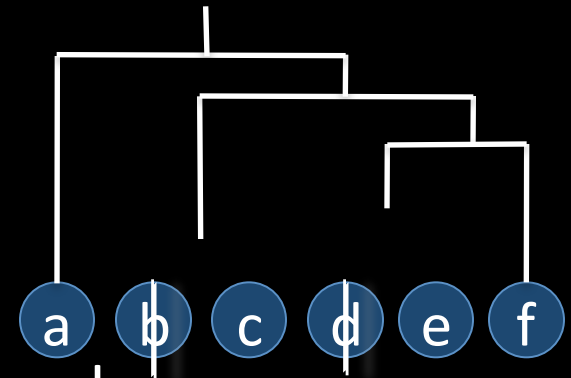
- Group similar instances into the same cluster
- Place unlike instances into different clusters
- Uses
 - Exploratory data analysis
 - Generalization / Learning



ML Types of Clustering

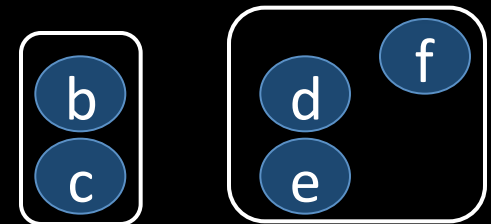
- Hierarchical clustering

- Relation between clusters is expressed in the clustering and represents similarity
- Tree where each node represents a cluster and the children represent subclasses of the parent

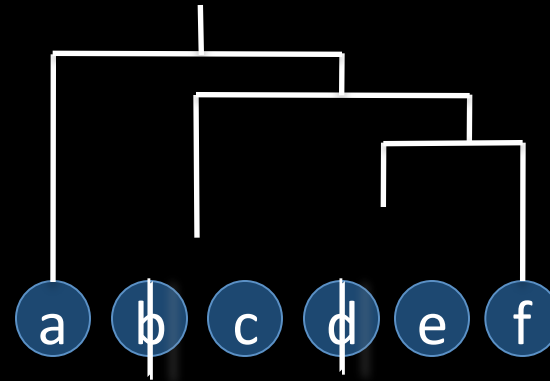
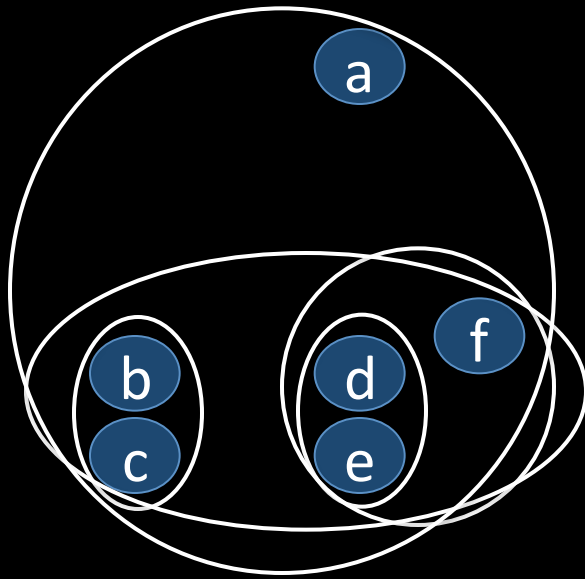


- Flat or Non-hierarchical clustering

- Typically pre-specified number of clusters
- Relation between clusters is unknown
- Typically iterative algorithms
 - Start with one set and iteratively improve clusters



ML Bottom-up Hierarchical Clustering



ML Types of Clustering

- Hard clustering
 - Each instance is placed in exactly one cluster
 - Hierarchical clustering is hard clustering
 - Drawback: consider clustering words by part-of-speech
- Soft clustering
 - Instances are placed in multiple clusters, typically all clusters probabilistically
 - Usually soft clustering indicates uncertainty
 - In disjunctive clustering, exs belong to multiple clusters
 - Flat clustering can be either hard or soft clustering

ML Hierarchical vs. Non-hierarchical

- Hierarchical clustering
 - Good for in-depth analysis
 - More informative than flat clustering
 - Best algorithm is problem-dependent
 - Computationally intensive vs. most flat clustering
- Non-hierarchical clustering
 - K-means is generally a good algorithm (or EM also)
 - More efficient
 - Many assume a Euclidian space

ML Hierarchical Clustering

- Bottom-up tree building
 - By iteratively clustering the most similar nodes
 - In a greedy search
 - Until the top most node contains all instances
 - Known as agglomerative clustering
- Top-down tree building
 - By starting with one group and iteratively dividing into groups with strong intra-group similarity (cohesion)

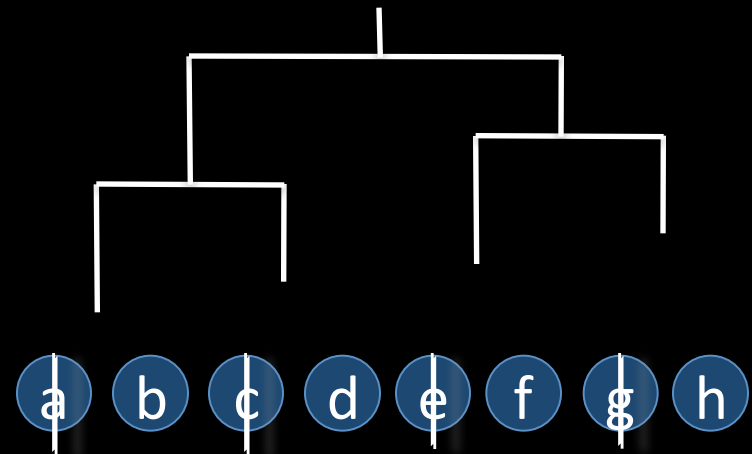
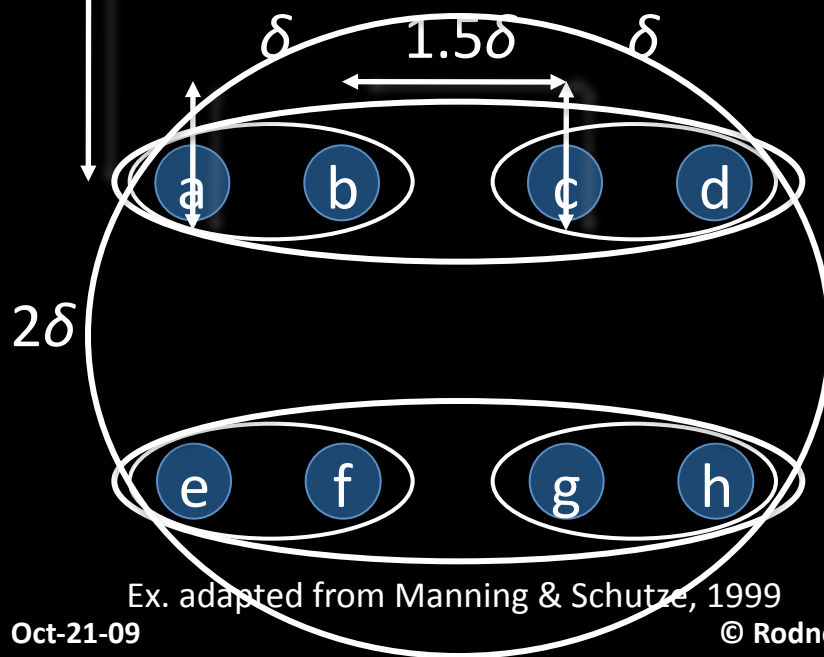
ML Bottom-up Hierarchical Clustering

- Single-link: similarity of 2 most similar members
- Complete-link: sim of 2 least similar members
- Group-average: ave similarity between members

ML Single-link Hierarchical Clustering

- Single-link: similarity of 2 most similar members
 - Search over all inter-member instance pairs
 - Good local similarity, but possibly poor global similarity
- Chaining effect

– Related to Minimum Spanning Tree (MST)



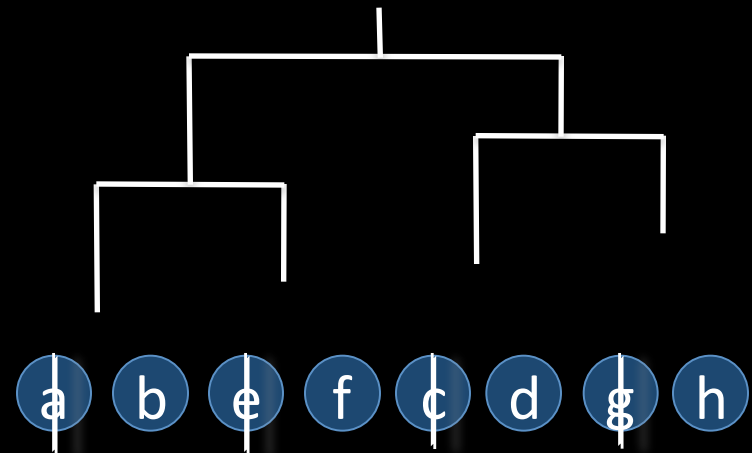
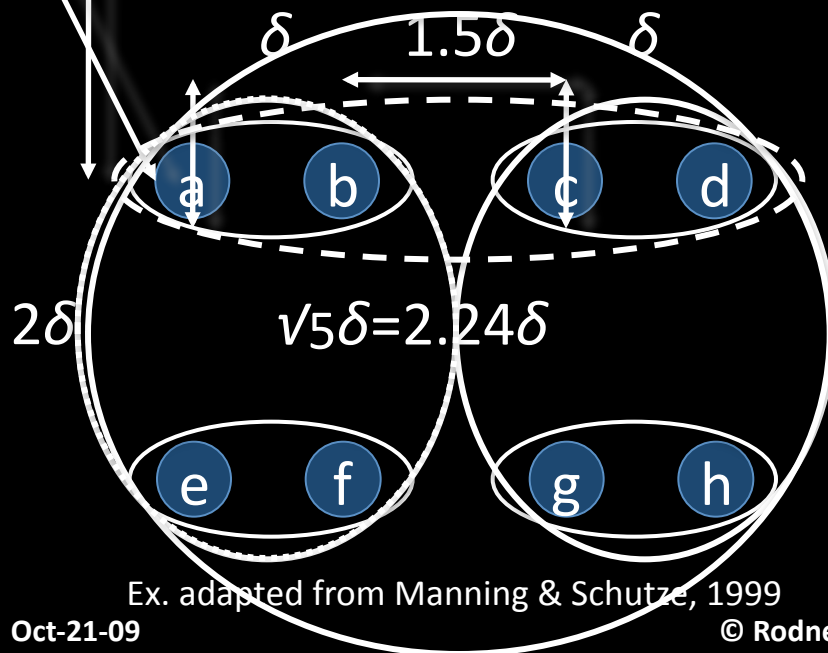
Ex. adapted from Manning & Schütze, 1999

ML Group-average Agglomerative Clustering

- Merge clusters by average similarity of instances
 - Avoids long stringy clusters

$$\text{ave}(s(a,b),s(a,c),s(a,d),s(b,c),s(b,d),s(c,d)) = (\delta+2.5\delta+3.5\delta+1.5\delta+2.5\delta+\delta)/6 = 12\delta/6$$

$$\text{ave}(s(a,b),s(a,e),s(a,f),s(b,e),s(b,f),s(e,f)) = (\delta+2\delta+\sqrt{5}\delta+2\delta+\sqrt{5}\delta+\delta)/6 = 10.5\delta/6$$



Ex. adapted from Manning & Schütze, 1999

ML Top-down Clustering Questions

- Using the same metrics (single-link, complete-link, group-average) and then the same bottom-up decision process...?
- Using different metrics, but the same bottom-up decision process...?
- Using a non-hierarchical clustering algorithm to partition the data at each node...?
 - This also allows for non-binary trees

ML Questions?

- Questions???

ML Non-Hierarchical Clustering

- Typically initialize the k clusters based on a different random seed for each cluster
- Then iteratively improve on this solution by moving instances from one cluster to another
- Hierarchical Clustering required just one pass

ML When to Stop Refinement

- When to stop?
 - Group average similarity
 - Mutual Information
 - Likelihood of the data given the clustering model

ML How Many Clusters

- How many clusters to create
 - Often know an appropriate number
 - Can use the same “goodness” measures, but these often increase with the number of clusters k
 - However, the goodness can jump by a larger amount from $k-1$ to k and a smaller amount from k to $k+1$, than seen on other transitions, which indicates the best k
 - MDL: Minimum Description Length

ML MDL Goodness Criteria

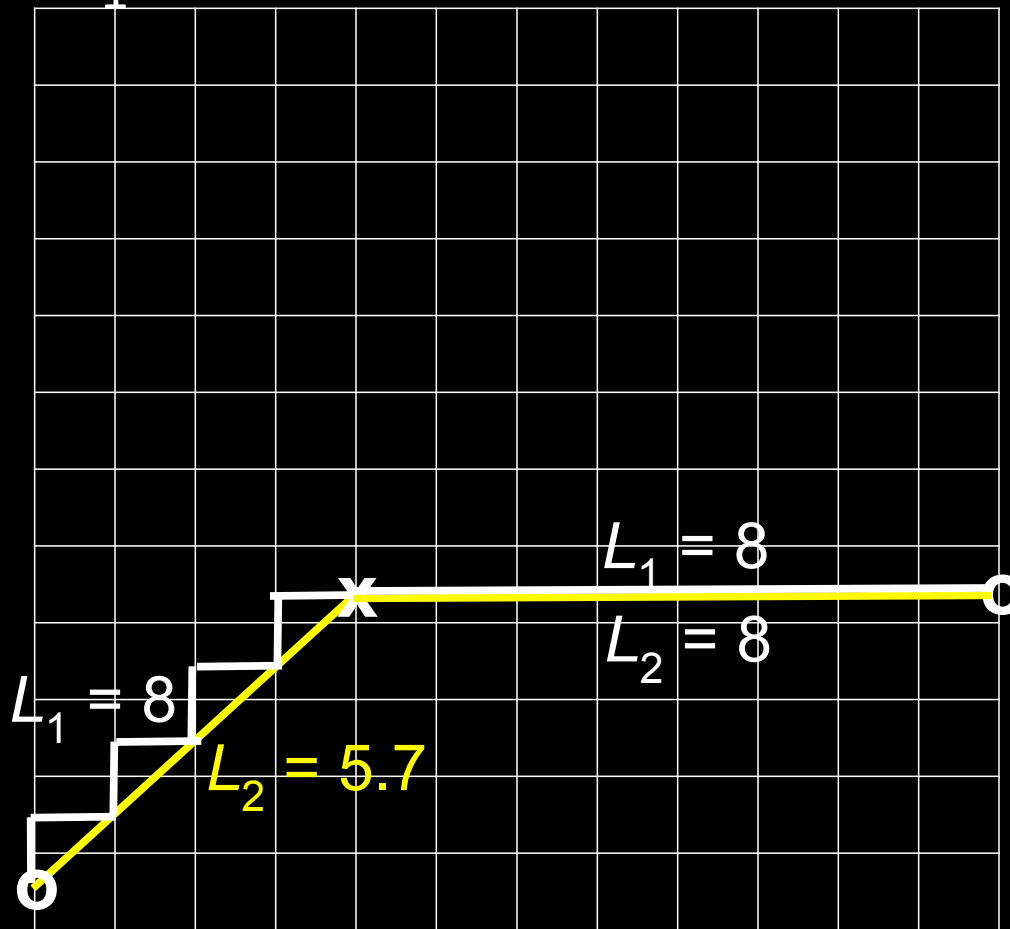
- All clusters *and* all instances are described in a binary code
- The length of this description determines the goodness of the clusters
- More clusters (higher k) requires that portion of the description be longer
- Instances are represented by their difference from the cluster
- More clusters \rightarrow smaller difference \rightarrow smaller *instance* description length, but
- More clusters \rightarrow longer *cluster* description length

ML What is the “Nearest” Cluster

- Compute a norm value to the centroid of each cluster
- The norm is typically the Euclidean distance (the L_2 norm)
 - $\sqrt{[\sum_j (x_j - c_j)^2]}$
- Sometimes Taxicab (aka Manhattan) distance is used (the L_1 norm)
 - $\sum_j |x_j - c_j|$
- This might be less sensitive to outliers

ML Euclidean vs. Manhattan Distance

- L_2 versus L_1 norm



ML Centroid vs. Medoid

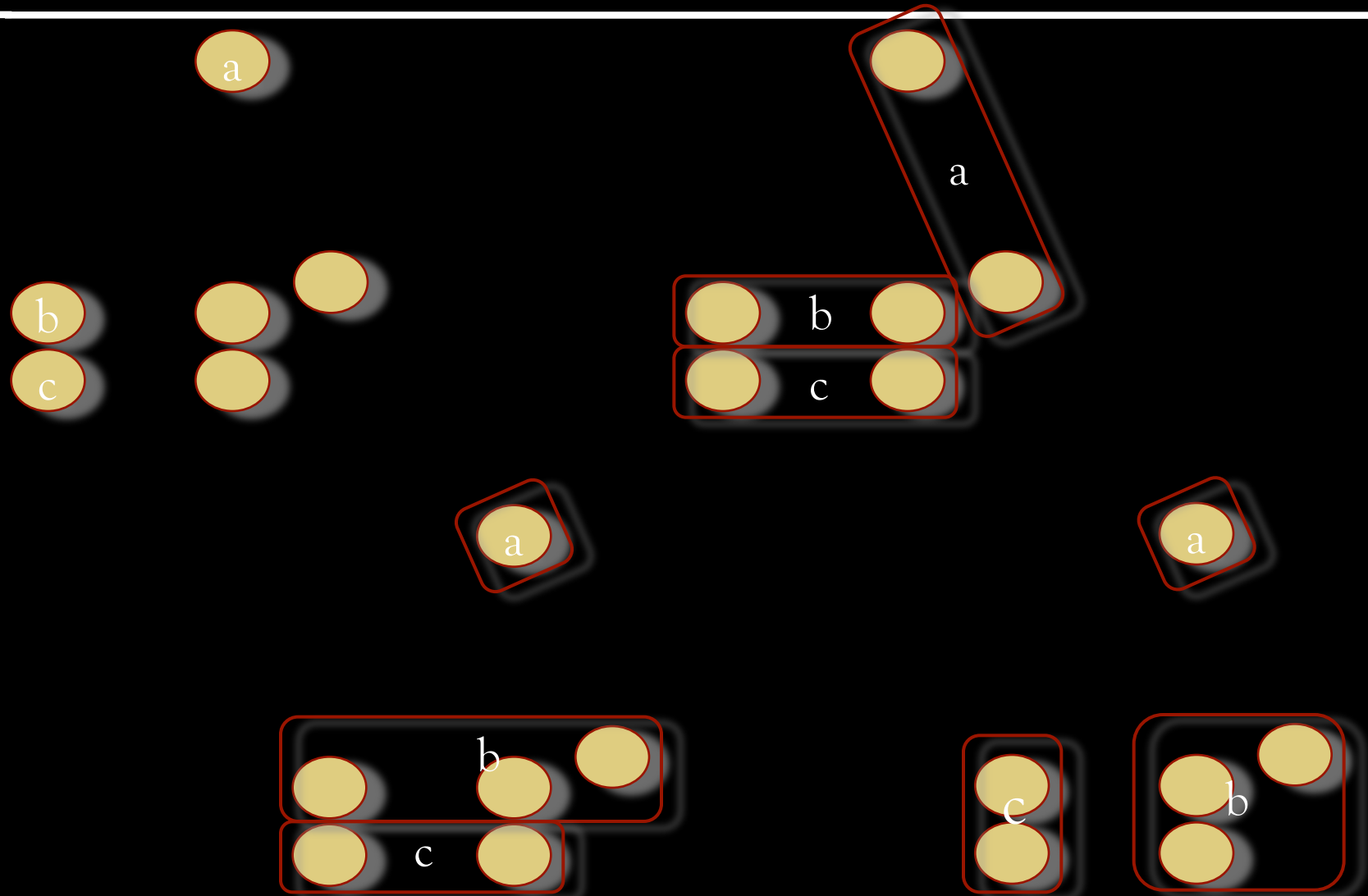
- Sometimes you might prefer to use a medoid rather than compute a centroid
- The medoid is a real instance near the centroid (a prototypical instance)

ML Complexity

- k -means is $O(kdN)$
- Until stopping criteria
 - Assign instances to the cluster whose centroid is “nearest”
 - Recalculate centroid = mean of cluster instances

ML

k-means Clustering



ML Clustering Example

Cluster	Members
1	ballot (0.28), polls (0.28), Gov (0.30) seats (0.32)
2	profit (0.21), finance (0.21), payments (0.22)
3	NFL (0.36), Reds (0.28), Sox (0.31), inning (0.33), quarterback (0.30), scored (0.33)
4	researchers (0.23), science (0.23)
5	Scott (0.28), Mary (0.27), Barbara (0.27), Edward (0.29)

- k -means is susceptible to local minima
 - Restart several time with different random cluster centroid seeds
 - Use hierarchical clustering to determine an initial set of clusters

- Assign instances to all clusters with some probability
- Expectation Maximization

ML Evaluating Clustering Results

- Purity
- ...

ML Questions?

- Questions???

ML Presentations

- Wed, Oct 28: Mark Lewis-Prazen
- Mon, Nov 2: Hansu Gu
- Volunteer
 - Wed, Nov 4: Paul Madden

ML Projects

- Office hours: will stay until all questions are addressed
- Or make an appointment for Friday