CSCE 590-1: From Data to Decisions with Open Data: A Practical Introduction to AI

Lecture 9: Paper Reading / Workshop
Lecture 10: Unsupervised Machine Learning

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

11TH FEB, 2021

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 9-10

- Introduction Segment
  - Quiz 1 evaluated
  - Recap of Lecture 8
- Main Segment
  - Discussion
    - Paper discussion: What Classification to use When
    - AAAI 2021 DEEP-DIAL Workshop
  - Unsupervised ML
    - Setting and characteristics
    - Method: k-means
    - Working with Weka
- Concluding Segment
  - About Next Lecture – Lecture 11
  - Ask me anything

Introduction Segment

Quiz 1 Evaluated

- Almost all did well
- pdf evaluated copy
  - Please use standard pdf writer
  - Not visible in some editing software

Recap of Lecture 8

- Supervised ML – looked at:
  - Naïve Bayes Method
  - Gradient Tree Boosting
  - Neural Network – MLP
  - Metrics: ROC/ AUC
- Paper discussion: 10 tips

Main Segment

Discussion: Which ML to Use

- Access: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5890912/

- Olson RS, Cava W, Mustahsan Z, Varik A, Moore JH. Data-driven advice for applying machine learning to bioinformatics problems. *Pac Symp Biocomput*. 2018;23:192-203.

Reading Group Allocation

a	Brendan	Curran
a	Sarayu	Das
b	Marilyn	Gartley
b	Vedant	Khandelwal
c	Vishal	Pallagani
c	Avineet Kumar	Singh
d	Ahad Hasan	Tanim
d	James	Thompson
e	Rohit	Naini
e	Terric	Taylor

AAAI 2021 Workshop

- Program: https://sites.google.com/view/deep-dial2021/program

- Feedback from students attending

Unsupervised Machine Learning

- Group data into clusters/ classes without supervision
  - Limited supervision

- What is a good cluster ?
  - Samples within a cluster should be “**near**” to each other (**cohesiveness**)
  - Samples in a cluster should be “**far**” from other samples in other clusters. (**distinctiveness**)

Data Representation

- Data matrix representation
  - N objects (data rows) x p attributes (columns)
  - Similar to classification

- Dissimilarity matrix
  - Object x Object structure
  - D(I, j) is difference or dissimilarity between (I, j), 0 means similar and 1 means dissimilar

Clustering for Data Understanding and Applications

- Biology: taxonomy of living things: kingdom, phylum, class, order, family, genus and species
- Information retrieval: document clustering
- Land use: Identification of areas of similar land use in an earth observation database
- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults
- Climate: understanding earth climate, find patterns of atmospheric and ocean
- Economic Science: market resarch

Content: Jiawei Han, Micheline Kamber and Jian Pei
Data Mining: Concepts and Techniques, 3rd ed.

Clustering as a Preprocessing Tool (Utility)

- Summarization:
  - Preprocessing for regression, PCA, classification, and association analysis
- Compression:
  - Image processing: vector quantization
- Finding K-nearest Neighbors
  - Localizing search to one or a small number of clusters
- Outlier detection
  - Outliers are often viewed as those “far away” from any cluster

Content: Jiawei Han, Micheline Kamber and Jian Pei
Data Mining: Concepts and Techniques, 3rd ed.

Considerations for a Clustering Algorithm

- Need a distance measure for ***far*** and ***near***
- Be able to explain what a cluster means
- Handle different types of attributes: numeric, categorical (nominal, ordinal), binary
- Detect different shapes of clusters
- Handle noisy data
- Scale
  - Size
  - Dimensions

Major Clustering Approaches (I)

Partitioning approach:

- Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
- Typical methods: **k-means**, k-medoids, CLARANS

Hierarchical approach:

- Create a hierarchical decomposition of the set of data (or objects) using some criterion
- Typical methods: Diana, Agnes, **BIRCH**, CAMELEON

Density-based approach:

- Based on connectivity and density functions
- Typical methods: **DBSACN**, OPTICS, DenClue

Grid-based approach:

- based on a multiple-level granularity structure
- Typical methods: STING, WaveCluster, CLIQUE

Content: Jiawei Han, Micheline Kamber and Jian Pei
Data Mining: Concepts and Techniques, 3rd ed.

Major Clustering Approaches (II)

Model-based:

- A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other
- Typical methods: **EM**, SOM, COBWEB

Frequent pattern-based:

- Based on the analysis of frequent patterns
- Typical methods: p-Cluster

User-guided or constraint-based:

- Clustering by considering user-specified or application-specific constraints
- Typical methods: COD (obstacles), constrained clustering

Link-based clustering:

- Objects are often linked together in various ways
- Massive links can be used to cluster objects: **SimRank**, LinkClus

Content: Jiawei Han, Micheline Kamber and Jian Pei
Data Mining: Concepts and Techniques, 3rd ed.

Content: Jiawei Han, Micheline Kamber and Jian Pei
Data Mining: Concepts and Techniques, 3rd ed.

Partitioning Algorithms: Basic Concept

Partitioning method: Partitioning a database ***D*** of ***n*** objects into a set of ***k*** clusters, such that the sum of squared distances is minimized (where c_i is the centroid or medoid of cluster C_i)

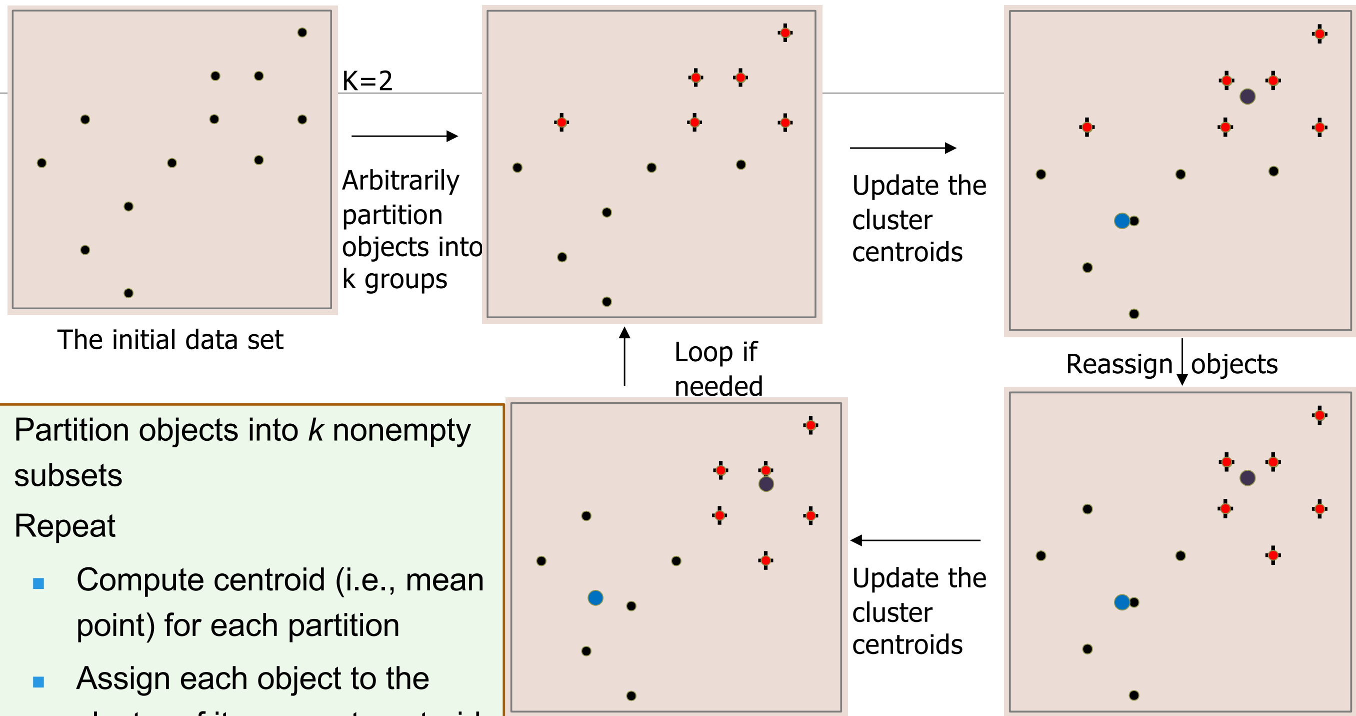
$$E = \Sigma_{i=1}^{k} \Sigma_{p \in C_i} (p - c_i)^2$$

Given *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion

- Global optimal: exhaustively enumerate all partitions
- Heuristic methods: *k-means* and *k-medoids* algorithms
- *k-means* (MacQueen'67, Lloyd'57/'82): Each cluster is represented by the center of the cluster
- *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster

Content: Jiawei Han, Micheline Kamber and Jian Pei
Data Mining: Concepts and Techniques, 3rd ed.

An Example of *K-Means* Clustering



- Partition objects into *k* nonempty subsets
- Repeat
  - Compute centroid (i.e., mean point) for each partition
  - Assign each object to the cluster of its nearest centroid
- Until no change

Content: Jiawei Han, Micheline Kamber and Jian Pei
Data Mining: Concepts and Techniques, 3rd ed.

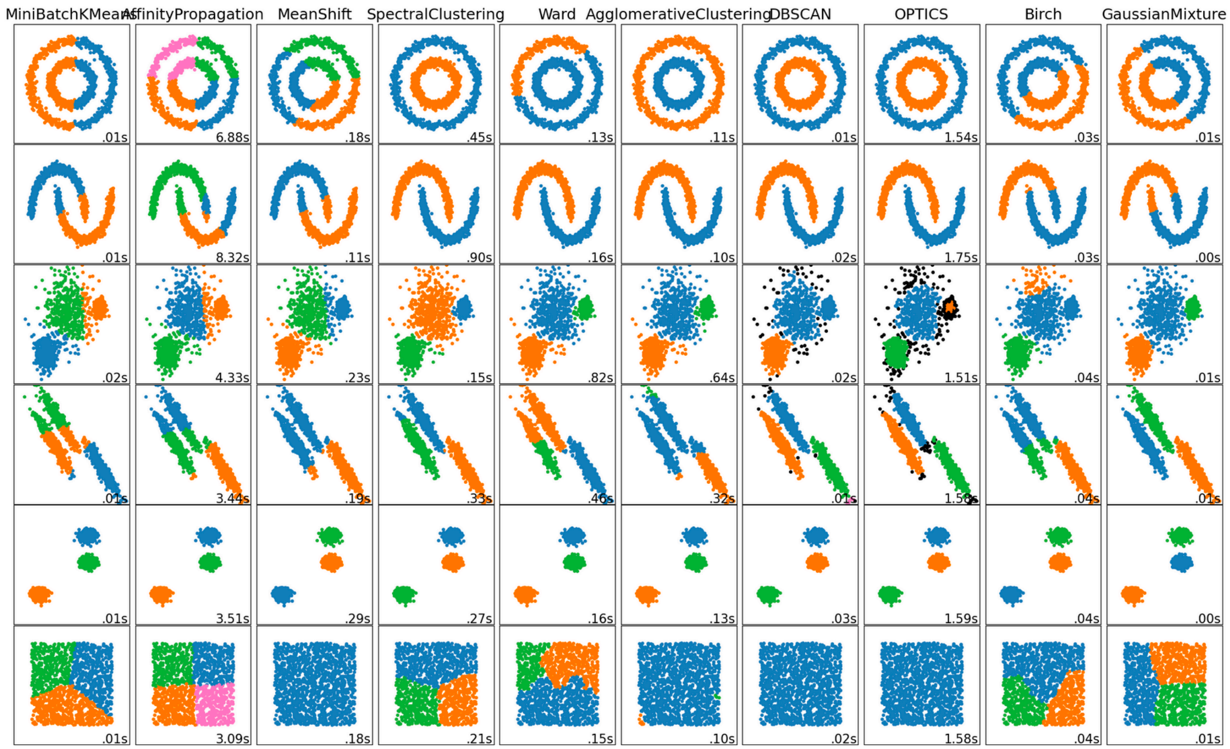
Comments on the *K-Means* Method

- **Strength**: *Efficient*: *O*(*tkn*), where *n* is # objects, *k* is # clusters, and *t* is # iterations. Normally, *k*, *t* << *n*.
  - Comparing: PAM: $O(k(n-k)^2)$, CLARA: $O(ks^2 + k(n-k))$
- **Comment**: Often terminates at a *local optimal*.
- **Weakness**
  - Applicable only to objects in a continuous n-dimensional space
    - Using the k-modes method for categorical data
    - In comparison, k-medoids can be applied to a wide range of data
  - Need to specify *k,* the *number* of clusters, in advance (there are ways to automatically determine the best k (see Hastie et al., 2009)
  - Sensitive to noisy data and *outliers*
  - Not suitable to discover clusters with *non-convex shapes*

Exercise: Weka

- Use K-means on weather.arff
- Vary k

Snapshot of Clustering Methods



A comparison of the clustering algorithms in scikit-learn

Lecture 10: Concluding Comments

- We looked at paper - (Which ML to use)
- Understood Clustering problem
- Understood k-means
- Explored with
  - Weka tool
  - Code sample

Concluding Segment

About Next Lecture – Lecture 11

Lecture 11: Unsupervised Learning

- Structured Data: Unsupervised Methods

- Methods: More methods
- Measuring cluster quality
- Explaining clusters
- Working with Weka