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

Lecture 11: Unsupervised Machine Learning

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

16TH FEB, 2021

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

Organization of Lecture 11

- Introduction Segment
  - Recap of Lecture 9-10
- Main Segment
  - Clustering: More methods
  - Distance metrics
  - Measuring cluster quality
  - Explaining / describing clusters
- Concluding Segment
  - About Next Lecture – Lecture 12
  - Ask me anything

Discussion

- Quiz 1: uploading score is not working

- Project
  - Reviewed most project plans
  - Dates are missing
  - Please discuss in office hour

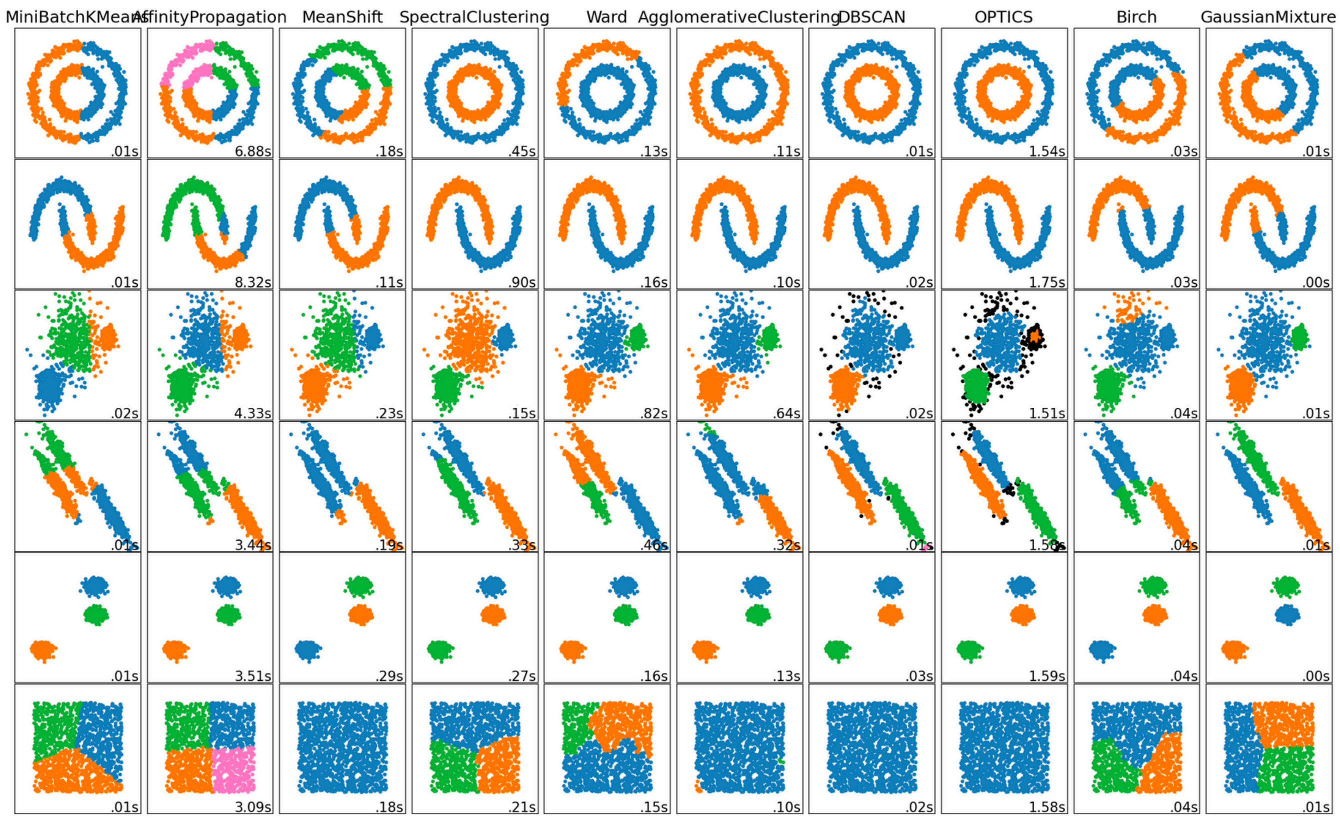
Introduction Segment

Recap of Lecture 10

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

Main Segment

Snapshot of Clustering Methods



A comparison of the clustering algorithms in scikit-learn

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

Conceptual Clustering

- Conceptual clustering
  - A form of clustering in machine learning
  - Produces a classification scheme for a set of unlabeled objects
  - Finds characteristic description for each concept (class)
- COBWEB (Fisher'87)
  - A popular a simple method of incremental conceptual learning
  - Creates a hierarchical clustering in the form of a classification tree
  - Each node refers to a concept and contains a probabilistic description of that concept

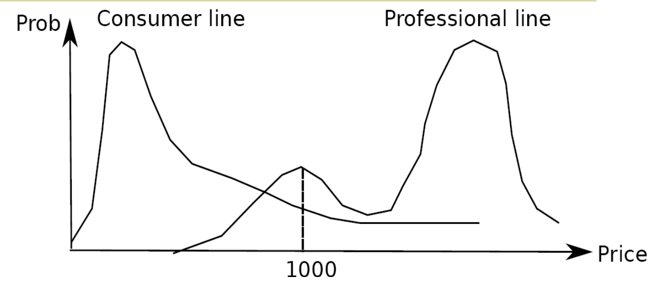
Code in Python
- https://github.com/cmaclell/concept_formation
- https://concept-formation.readthedocs.io/en/latest/examples.html
- https://concept-formation.readthedocs.io/en/latest/examples/cobweb_cluster_mushroom.html

Probabilistic Model-Based Clustering

Cluster analysis is to find hidden categories.

A hidden category (i.e., *probabilistic cluster)* is a distribution over the data space, which can be mathematically represented using a probability density function (or distribution function).

- Ex. 2 categories for digital cameras sold
  - consumer line vs. professional line
  - density functions f_1, f_2 for C_1, C_2
  - obtained by probabilistic clustering

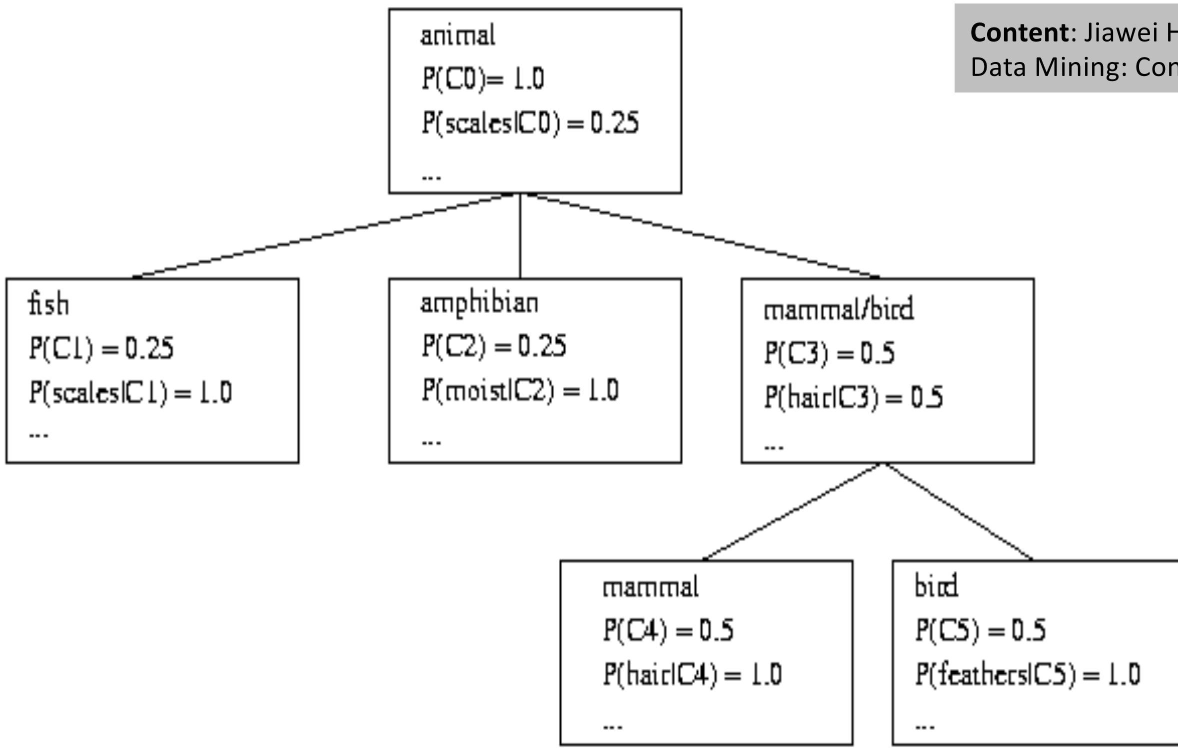


- A **mixture model** assumes that a set of observed objects is a mixture of instances from multiple probabilistic clusters, and conceptually each observed object is generated independently
- **Out task**: infer a set of *k* probabilistic clusters that is mostly likely to generate *D* using the above data generation process

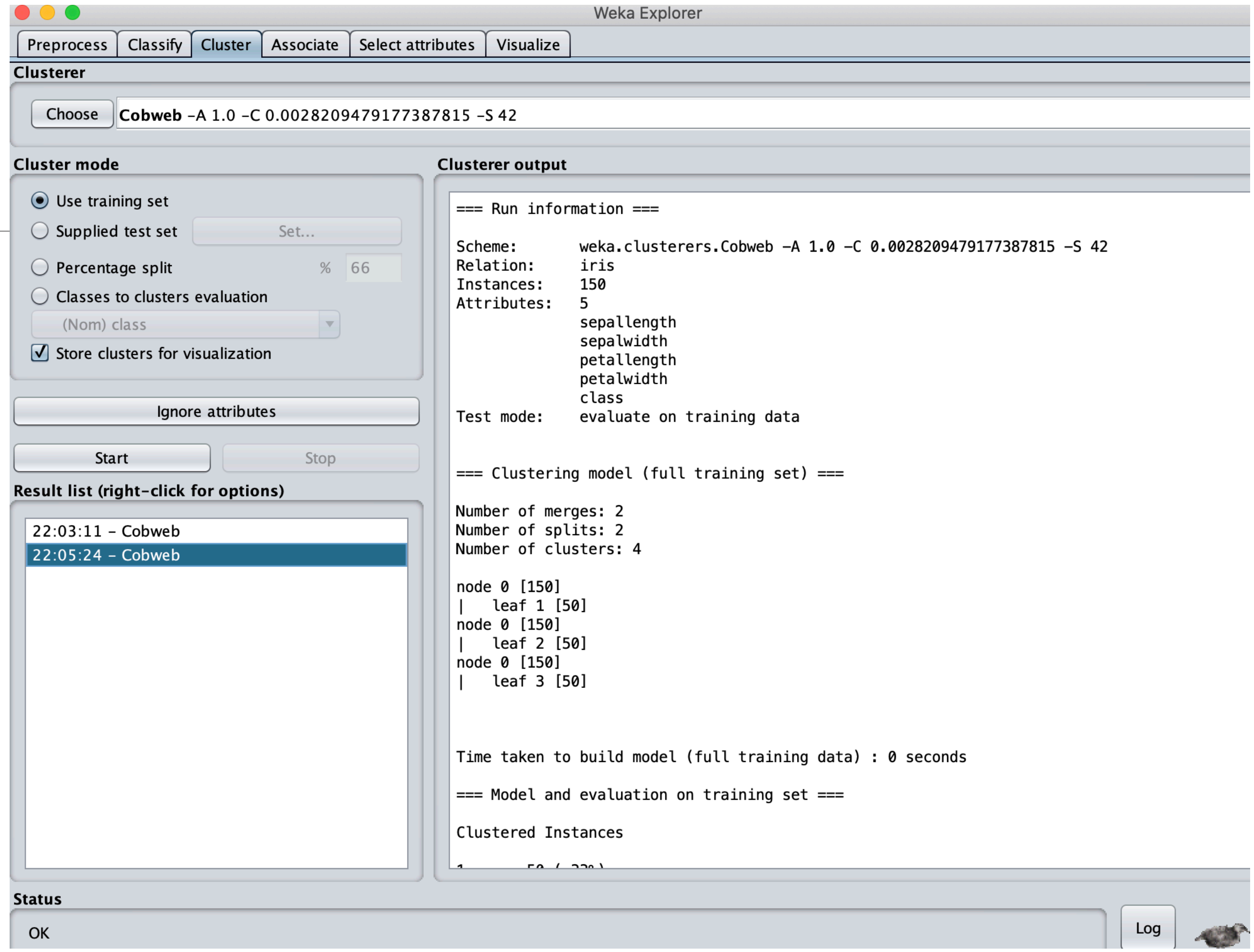
COBWEB Clustering Method

A classification tree

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



COBWEB in Weka



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

More on Conceptual Clustering

- Limitations of COBWEB
  - The assumption that the attributes are independent of each other is often too strong because correlation may exist
  - Not suitable for clustering large database data – skewed tree and expensive probability distributions
- CLASSIT
  - an extension of COBWEB for incremental clustering of continuous data
  - suffers similar problems as COBWEB
- AutoClass (Cheeseman and Stutz, 1996)
  - Uses Bayesian statistical analysis to estimate the number of clusters

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

Distance Metrics – Numeric Variables

- Numeric quantity
  - Interval-scaled variables: continuous measurements of a roughly linear scale.
- Standardize with mean absolute deviation
  - $s_f = (1 / n) * (|x_{1f} – m_f| + ... + |x_{1f} – m_f|)$
    - s_{nf} and m_f are measurements and mean, respectively
  - $z_{if} = (x_{if} – m_f) / s_f$

Examples: weight, height, latitude, longitude, temperature

- Distances for numbers
  - Euclidean: d(i,j) = square root ($|x_{i1} – x_{j1}|^2 + ... + |x_{ip} – x_{jp}|^2$) , for p-dimensional data
  - Manhattan: d(i,j) = $|x_{i1} – x_{j1}| + ... + |x_{ip} – x_{jp}|$, for p-dimensional data
  - Minlowski: 1/q root ($|x_{i1} – x_{j1}|^q + ... + |x_{ip} – x_{jp}|^q$) , for p-dimensional data

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

Distance Metrics – Binary Variables

		Object J		
		1	0	Sum
Object I	1	q	r	q+r
	0	s	t	s+t
	Sum	q+s	r+t	q+r+s+t

Contingency table for binary variables

- Notation
  - q: number of binary variables that equal 1 for both objects I and J
- Distance between objects by matching
- d(I, J) = (r + s) / (q + r + s + t)

Examples:
Smoker/ non-smoker,
electric v/s non-electric car

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

Distance Metrics – Nominal Variables

- Notation
  - m: number of matches in values of nominal variables between objects I and J
  - M: total number of variables

- Distance between objects defined by matching
- d(I, J) = (p - m) / (p)

Examples:
map-color - red, yellow, green, pink, blue

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

Distance Metrics – Ordinal Variables

- Conversion and notation
  - $z_{if} = (r_{if} - 1) / (M_{if} - 1)$
  - variable f of i-th object has $1..M_f$ states in that order

- Now reuse distances for numbers
  - Euclidean: $d(i,j) =$ square root ($|x_{i1} - x_{j1}|^2 + \ldots + |x_{ip} - x_{jp}|^2$) , for p-dimensional data
  - Manhattan: $d(i,j) = |x_{i1} - x_{j1}| + \ldots + |x_{ip} - x_{jp}|$, for p-dimensional data
  - Minlowski: 1/q root ($|x_{i1} - x_{j1}|^q + \ldots + |x_{ip} - x_{jp}|^q$) , for p-dimensional data

Examples:
professor ranks – assistant, associate, full
Medals – bronze, silver, gold
Military - …

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

Distance for Mixed Variable Types

- Keep separate and perform cluster analysis separately
  - Impractical
- Combine them into one scale between 0 to 1
- d(i,j) = $\frac{\sum_{f=1}^{p} \delta_{ij}{}^{(f)} d_{ij}{}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}{}^{(f)}}$
  - Where $\delta_{ij}{}^{(f)}$ is 0 if x_{if} or x_{jf} are missing, otherwise 1
  - $d_{ij}{}^{(f)}$ is distance between i and j for feature f and type
- There can be a weighted variation too

Exercise - 1

- Consider clustering of days
  - What are some possible groups?
  - What features make sense?
  - What distances make sense?

Exercise - 2

Consider clustering of documents, like resumes, into groups

- What are some possible groups?
  - By areas: Technology, finance, services, manufacturing, …

- What features make sense?
  - Syntactic: Words, sentiments, …
  - Semantic: qualification, experience, …

- What distances make sense?

Clustering Quality

Case A: Ground Truth is Known

- **homogeneity**: each cluster contains only members of a single class.
- **completeness**: all members of a given class are assigned to the same cluster

- Example:
  - true labels = [0, 0, 0, 1, 1, 1]
  - P1: Predicted labels = [0, 0, 1, 1, 2, 2]
  - P2: Predicted labels = [0, 0, 0, 2, 2, 2]
- In example P1, informally
  - Homogeneity - (Predicted) 1 has members of 0 and 1
  - Completeness – (Actual) 0 is assigned to 0 and 1, (Actual) 1 is assigned 1 and 2

Note: P2 is homogeneous and complete

Content acknowledgement: Sci-kit: https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

Case A: Ground Truth is Known

- **homogeneity**: each cluster contains only members of a single class.
- **completeness**: all members of a given class are assigned to the same cluster

- **v-measure**

$$v = \frac{(1 + \beta) \times \textbf{homogeneity} \times \textbf{completeness}}{(\beta \times \textbf{homogeneity} + \textbf{completeness})}$$

Content acknowledgement: Sci-kit: https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

Case B: Ground Truth is Unknown

Silhouette Coefficient

- **a**: The mean distance between a sample and all other points in the same class.
- **b**: The mean distance between a sample and all other points in the *next nearest cluster*.

The Silhouette Coefficient *s* for a single sample is then given as:

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

The Silhouette Coefficient for a set of samples is given as the mean of the Silhouette Coefficient for each sample.

-1: incorrect clustering
+1: highly dense clustering.
Scores around zero indicate overlapping clusters.

Question: can you calculate when all data is in one cluster?

Content acknowledgement: Sci-kit: https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

Case B: Ground Truth is Unknown

Davies-Bouldin Index

- s_i, the average distance between each point of cluster i and the centroid of that cluster – also know as cluster diameter.
- d_{ij}, the distance between cluster centroids i and j.

A simple choice to construct R_{ij} so that it is nonnegative and symmetric is:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

Then the Davies-Bouldin index is defined as:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij}$$

0: best
1: worst

Limitation: Needs euclidean distances

Content acknowledgement: Sci-kit: https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

Measuring Clustering Quality

- Two methods: extrinsic vs. intrinsic
- Extrinsic: supervised, i.e., the ground truth is available
  - Compare a clustering against the ground truth using certain clustering quality measure
  - Ex. Recall - precision and recall metrics in classification
- Intrinsic: unsupervised, i.e., the ground truth is unavailable
  - Evaluate the goodness of a clustering by considering how well the clusters are separated, and how compact the clusters are
  - Ex. Silhouette coefficient

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

Measuring Clustering Quality: Extrinsic Methods

- Clustering quality measure: $Q(C, C_g)$, for a clustering C given the ground truth C_g.
- Q is good if it satisfies the following **4** essential criteria
  - Cluster homogeneity: the purer, the better
  - Cluster completeness: should assign objects belong to the same category in the ground truth to the same cluster
  - Rag bag: putting a heterogeneous object into a pure cluster should be penalized more than putting it into a *rag bag* (i.e., "miscellaneous" or "other" category)
  - Small cluster preservation: splitting a small category into pieces is more harmful than splitting a large category into pieces

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

Summary

- Cluster analysis groups objects based on their similarity and has wide applications
- Measure of similarity can be computed for various types of data
- Clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods
- K-means and K-medoids algorithms are popular partitioning-based clustering algorithms
- Birch and Chameleon are interesting hierarchical clustering algorithms, and there are also probabilistic hierarchical clustering algorithms
- DBSCAN, OPTICS, and DENCLU are interesting density-based algorithms
- STING and CLIQUE are grid-based methods, where CLIQUE is also a subspace clustering algorithm
- Quality of clustering results can be evaluated in various ways

Code Examples

- Clustering quality
  - https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l10-11-unsupervised-ml/clustering-quality-measures.ipynb

- Clustering methods
  - https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l10-11-unsupervised-ml/Cluster-exploration-syntheticdata.ipynb

Exercise: Weka

- Pick a data-set with at least 5 attributes
- Cluster with 2 methods
- Review cluster quality

Explaining Clusters

- How to describe them ?
  - Centroid
  - Exemplars

- What name to give them ?
  - Using features of the members
  - Algorithm may produce (Concept Clustering)

- Explanations can be based on domain specific rules

Lecture 11: Concluding Comments

- Clustering: More method
- Distance metrics
- Measuring cluster quality
- Explaining / describing clusters

Concluding Segment

About Next Lecture – Lecture 12

Lecture 12: Unsupervised Learning

- Advanced ML topics
  - AutoAI – automating machine learning pipeline
  - Generating explanations
- Quiz 2
- Reading exercise: AutoAI paper
  - Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools; https://arxiv.org/abs/1908.05557, 2019
  - Discuss in class on Feb 23, 2020