CSCE 580: Introduction to AI

Lecture 14: Unsupervised Machine Learning

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

3RD OCTOBER, 2024

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

Organization of Lecture 14

- Introduction Segment
  - Recap of Lecture 13
- Main Segment
  - Unsupervised ML
    - Setting and characteristics
    - Method: k-means
    - Working with Weka
- Concluding Segment
  - About Next Lecture – Lecture 15
  - Ask me anything

Introduction Segment

Recap of Lecture 13

- We talked about
  - The variety of methods for classification
  - Logistic Regression
  - Decision trees
  - Random forest
  - Naïve Bayes
  - Boosting
  - Metrics – AUC / ROC
  - Discussion: Choosing a method that works

Main Segment

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)

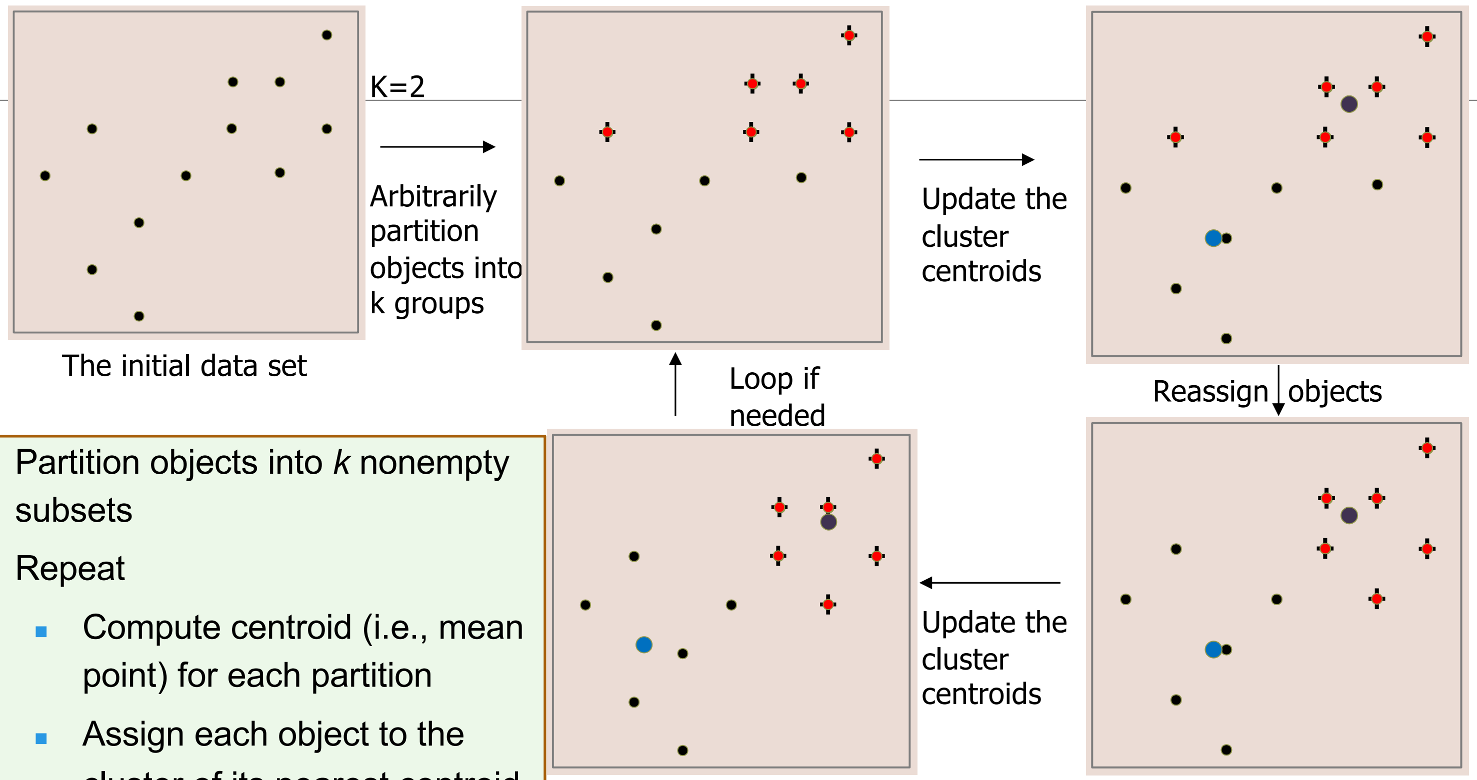
$$E = \Sigma_{i=1}^{k} \Sigma_{p \in C_i} (p - c_i)^2$$

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.

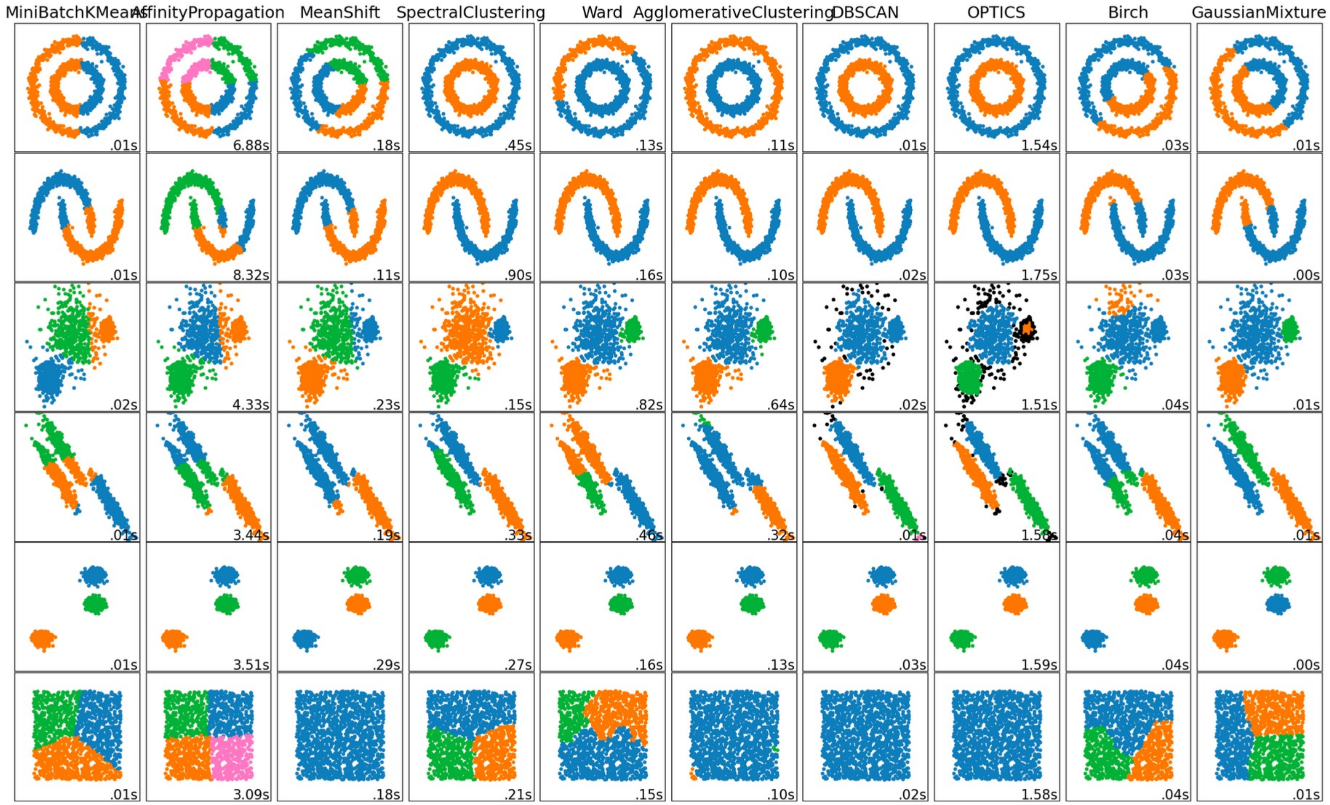
Comments on the *K-Means* Method

- **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

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
  - $s_f = (1 / n) * (|x_{1f} - m_f| + \ldots + |x_{1f} - m_f|)$
    - 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 ($|x_{i1} - x_{j1}|^2 + \ldots + |x_{ip} - x_{jp}|^2$) , 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

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**

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

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

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

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 14: Concluding Comments

- Understood Clustering problem
- Understood k-means
- A range of clustering methods
- Measuring cluster quality
- Explaining clusters
- Working with Weka, scikit and python code samples

Concluding Section

Course Project

Discussion: Projects

- New: two projects
  - Project 1: model assignment
  - Project 2: single problem/ llm based solving / fine-tuning/ presenting result

Project Discussion

1. Go to Google spreadsheet against your name
2. Enter model assignment name and link from (http://modelai.gettysburg.edu/)

1. Create a private Github repository called "CSCE58x-Fall2024-<studentname>-Repo". Share with Instructor (biplav-s) and TA (vishalpallagani)
2. Create Google folder called "CSCE58x-Fall2024-<studentname>-SharedInfo". Share with Instructor (prof.biplav@gmail.com) and TA (vishal.pallagani@gmail.com)
3. Create a Google doc in your Google repo called "Project Plan" and have the following by next class (Sep 5, 2024)

Timeline

1. Title:
2. Key idea: (2-3 lines)
3. Data need:
4. Methods:
5. Evaluation:
6. Milestones
    1. // Create your own
7. Oct 3, 2024

Reference: Project 1 Rubric (30% of Course)

Assume total for Project-1 as 100

- **Project results** – 60%
  - Working system ? – 30%
  - Evaluation with results superior to baseline? – 20%
  - Went through project tasks completely ? – 10%
- **Project efforts** – 40%
  - Project report – 20%
  - Project presentation (updates, final) – 20%
- **Bonus**
  - Challenge level of problem – 10%
  - Instructor discretion – 10%
- **Penalty**
  - Lack of timeliness as per your milestones policy (right) - up to 30%

Milestones and **Penalties**

- Project plan due by Sep 5, 2024 **[-10%]**
- Project deliverables due by Oct 3, 2024 **[-10%]**
- Project presentation on Oct 8, 2024 **[-10%]**

Report Format

1. Title:
2. Key idea: (2-3 lines)
3. Data need:
4. Methods:
5. Screen shot (as applicable)
6. Evaluation:
7. Experience: *what learnt, anything special to discuss with class*

Presentation Format

2 minute video

Screen Shot

1. Title:
2. Key idea: 1 line summary
3. Data need:
4. Effort and Result
   1. What was done (scope)
   2. What was not done (decided not to, couldn't)
   3. Result

Experience

About Next Lecture – Lecture 15

Lecture 15: Student Presentations

- Project-1 presentations
  - 1-2 minute video from uploaded presentation
  - 1 minute Q/A

9	Sep 17 (Tu)	Local search
10	Sep 19 (Th)	Adversarial games and search
11	Sep 24 (Tu)	Constraints & optimization
12	Sep 26 (Th)	Machine Learning - Basics
13	Oct 1 (Tu)	Machine Learning – Classification – Decision Trees, Random Forest, NBC, Gradient Boosting, ML-Text
14	Oct 3 (Th)	ML – Unsupervised / Clustering
15	Oct 8 (Tu)	Student presentations - project
16	Oct 10 (Th)	ML – NN, Deep Learning