Introduction to Artificial Intelligence
Unsupervised Learning

Janyl Jumadinova
October 21, 2016
Supervised learning vs. Unsupervised learning

- **Supervised learning**: discover patterns in the data that relate data attributes with a target (class) attribute.
  - These patterns are then utilized to predict the values of the target attribute in future data instances.
Supervised learning vs. Unsupervised learning

- **Supervised learning**: discover patterns in the data that relate data attributes with a target (class) attribute.
  - These patterns are then utilized to predict the values of the target attribute in future data instances.

- **Unsupervised learning**: the data has no target attribute.
  - We want to explore the data to find some intrinsic structures in them.
Clustering

- Organizing data into classes such that there is:
  - high intra-class similarity
  - low inter-class similarity
Clustering

- Organizing data into classes such that there is:
  - high intra-class similarity
  - low inter-class similarity
- Finding the class labels and the number of classes directly from the data (in contrast to classification).
Clustering

Organizing data into classes such that there is:
- high intra-class similarity
- low inter-class similarity

Finding the class labels and the number of classes directly from the data (in contrast to classification).

More informally, finding natural groupings among objects.
Clustering is one of the most utilized data mining techniques. It has a long history, and is used in almost every field, e.g., medicine, psychology, botany, sociology, biology, archeology, marketing, insurance, libraries, etc.

Ex.: Given a collection of text documents, we want to organize them according to their content similarities.

Ex.: In marketing, segment customers according to their similarities (to do targeted marketing).
Clustering

Clustering is one of the most utilized data mining techniques. It has a long history, and is used in almost every field, e.g., medicine, psychology, botany, sociology, biology, archeology, marketing, insurance, libraries, etc.

- Ex.: Given a collection of text documents, we want to organize them according to their content similarities.
Clustering is one of the most utilized data mining techniques. It has a long history and is used in almost every field, e.g., medicine, psychology, botany, sociology, biology, archeology, marketing, insurance, libraries, etc.

- Ex.: Given a collection of text documents, we want to organize them according to their content similarities.
- Ex.: In marketing, segment customers according to their similarities (to do targeted marketing).
What is a natural grouping among these objects?
What is a natural grouping among these objects?

Clustering is subjective

- Simpson's Family
- School Employees
- Females
- Males
What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features. Webster's Dictionary

Similarity is hard to define, but... "We know it when we see it"

The real meaning of similarity is a philosophical question. We will take a more pragmatic approach.
What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features. Webster’s Dictionary
What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features. Webster’s Dictionary
What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features. Webster’s Dictionary

Similarity is hard to define, but ... “We know it when we see it”

The real meaning of similarity is a philosophical question. We will take a more pragmatic approach.
**Definition:**

Let $O_1$ and $O_2$ be two objects from the universe of possible objects. The distance (dissimilarity) between $O_1$ and $O_2$ is a real number denoted by $D(O_1, O_2)$. 

![Image of objects and numbers]
What properties should a distance measure have?

- $D(A, B) = D(B, A)$  
  Symmetry  
  Otherwise you could claim “Greg looks like Oliver, but Oliver looks nothing like Greg.”

- $D(A, B) = 0$  
  Constancy of Self-Similarity  
  Otherwise you could claim “Greg looks more like Oliver, than Oliver does.”

- $D(A, B) = 0$ iff $A = B$  
  Positivity (Separation)  
  Otherwise there are objects in your world that are different, but you cannot tell apart.

- $D(A, B) \leq D(A, C) + D(B, C)$  
  Triangular Inequality  
  Otherwise you could claim “Greg is very like Bob, and Greg is very like Oliver, but Bob is very unlike Oliver.”
What properties should a distance measure have?

- \( D(A, B) = D(B, A) \)  \hspace{1cm} \text{Symmetry}
  
  Otherwise you could claim “Greg looks like Oliver, but Oliver looks nothing like Greg.”

- \( D(A, A) = 0 \)  \hspace{1cm} \text{Constancy of Self-Similarity}

  Otherwise you could claim “Greg looks more like Oliver, than Oliver does.”
What properties should a distance measure have?

- \( D(A, B) = D(B, A) \)  \hspace{1cm} \text{Symmetry}
  
  Otherwise you could claim “Greg looks like Oliver, but Oliver looks nothing like Greg.”

- \( D(A, A) = 0 \)  \hspace{1cm} \text{Constancy of Self-Similarity}
  
  Otherwise you could claim “Greg looks more like Oliver, than Oliver does.”

- \( D(A, B) = 0 \iff A = B \)  \hspace{1cm} \text{Positivity (Separation)}
  
  Otherwise there are objects in your world that are different, but you cannot tell apart.
What properties should a distance measure have?

- \( D(A, B) = D(B, A) \)  
  **Symmetry**
  
  Otherwise you could claim “Greg looks like Oliver, but Oliver looks nothing like Greg.”

- \( D(A, A) = 0 \)  
  **Constancy of Self-Similarity**
  
  Otherwise you could claim “Greg looks more like Oliver, than Oliver does.”

- \( D(A, B) = 0 \) iff \( A = B \)  
  **Positivity (Separation)**
  
  Otherwise there are objects in your world that are different, but you cannot tell apart.

- \( D(A, B) \leq D(A, C) + D(B, C) \)  
  **Triangular Inequality**
  
  Otherwise you could claim “Greg is very like Bob, and Greg is very like Oliver, but Bob is very unlike Oliver.”
How do we measure similarity?
How do we measure similarity?

To measure the similarity between two objects, transform one of the objects into the other, and measure how much effort it took. The measure of effort becomes the distance measure.

The distance between Patty and Selma.
Change dress color, 1 point
Change earring shape, 1 point
Change hair part, 1 point

\[ D(\text{Patty}, \text{Selma}) = 3 \]

The distance between Marge and Selma.
Change dress color, 1 point
Add earrings, 1 point
Decrease height, 1 point
Take up smoking, 1 point
Lose weight, 1 point

\[ D(\text{Marge}, \text{Selma}) = 5 \]

called the "edit distance" or the "transformation distance"
How do we measure similarity?

**Edit Distance Example**

It is possible to transform any string $Q$ into string $C$, using only *Substitution*, *Insertion* and *Deletion*. Assume that each of these operators has a cost associated with it.

The similarity between two strings can be defined as the cost of the cheapest transformation from $Q$ to $C$.

Note that for now we have ignored the issue of how we can find this cheapest transformation.

---

**How similar are the names “Peter” and “Piotr”?**

Assume the following cost function:

- *Substitution*: 1 Unit
- *Insertion*: 1 Unit
- *Deletion*: 1 Unit

$D(\text{Peter}, \text{Piotr})$ is 3

---

```
Peter
  ↓ Substitution (i for e)
Piter
  ↓ Insertion (o)
Pioter
  ↓ Deletion (e)
Piotr
```
Partitional Clustering

- Non-hierarchical, each instance is placed in exactly one of $K$ nonoverlapping clusters.
- Since only one set of clusters is output, the user normally has to input the desired number of clusters $K$. 
Minimize Squared Error

Distance of a point in cluster to the center of cluster:

$$s_{e K_i} = \sum_{j=1}^{m} ||t_{ij} - C_k||^2$$

$$s_{e K} = \sum_{j=1}^{k} s_{e K_j}$$

Objective Function
K-means clustering

- **K-means** is a partitional clustering algorithm.
- The k-means algorithm partitions the given data into *k* clusters.
  - Each cluster has a cluster center, called **centroid**.
  - *k* is specified by the user.
K-means Algorithm

1. Decide on a value for $k$. 
K-means Algorithm

1. Decide on a value for $k$.
2. Initialize the $k$ cluster centers (randomly, if necessary).
K-means Algorithm

1. Decide on a value for $k$.
2. Initialize the $k$ cluster centers (randomly, if necessary).
3. Decide the class memberships of the $N$ objects by assigning them to the nearest cluster center.
4. Re-estimate the $k$ cluster centers, by assuming the memberships found above are correct.
5. If none of the $N$ objects changed membership in the last iteration, exit. Otherwise goto 3.
K-means Algorithm

1. Decide on a value for $k$.
2. Initialize the $k$ cluster centers (randomly, if necessary).
3. Decide the class memberships of the $N$ objects by assigning them to the nearest cluster center.
4. Re-estimate the $k$ cluster centers, by assuming the memberships found above are correct.
K-means Algorithm

1. Decide on a value for \( k \).
2. Initialize the \( k \) cluster centers (randomly, if necessary).
3. Decide the class memberships of the \( N \) objects by assigning them to the nearest cluster center.
4. Re-estimate the \( k \) cluster centers, by assuming the memberships found above are correct.
5. If none of the \( N \) objects changed membership in the last iteration, exit. Otherwise goto 3.
K-Means Clustering: Step 1
K-Means Clustering: Step 2

Algorithm: k-means, Distance Metric: Euclidean Distance
K-Means Clustering: Step 3
K-Means Clustering: Step 4
K-Means Clustering: Step 5
How can we tell the right number of clusters?

- In general, this is an unsolved problem.
How can we tell the right number of clusters?

- In general, this is an unsolved problem.
How can we tell the right number of clusters?

- In general, this is an unsolved problem.

- We can use approximation methods!
When $k = 1$, the objective function is 873.0
When $k = 2$, the objective function is 173.1.
When $k = 3$, the objective function is 133.6
We can plot the objective function values for $k = 1...6$

- The abrupt change at $k = 2$, is highly suggestive of two clusters in the data.
- This technique for determining the number of clusters is known as “knee finding” or “elbow finding”.

![Graph showing objective function values for $k = 1...6$]
Strengths of K-Means

- Simple: easy to understand and to implement
- Efficient: Time complexity $O(tkn)$, where $n$ is the number of data points, $k$ is the number of clusters, and $t$ is the number of iterations.
  - Since both $k$ and $t$ are small, k-means is considered a linear algorithm.
- Often terminates at a local optimum.
  - The global optimum may be found using techniques such as: deterministic annealing and genetic algorithms
Weaknesses of K-Means

- The algorithm is only applicable if the mean is defined.
  - For categorical data - the centroid is represented by most frequent values.
  - The user needs to specify $k$. 
Weaknesses of K-Means

- The algorithm is only applicable if the mean is defined.
  - For categorical data - the centroid is represented by most frequent values.
  - The user needs to specify $k$.
- The algorithm is sensitive to outliers.
  - Outliers are data points that are very far away from other data points.
  - Outliers could be errors in the data recording or some special data points with very different values.
K-Means Summary

- Despite weaknesses, k-means is still the most popular algorithm due to its simplicity, efficiency and other clustering algorithms have their own lists of weaknesses.
- No clear evidence that any other clustering algorithm performs better in general, although they may be more suitable for some specific types of data or applications.
- Comparing different clustering algorithms is a difficult task. No one knows the correct clusters!