COL870: Clustering Algorithms

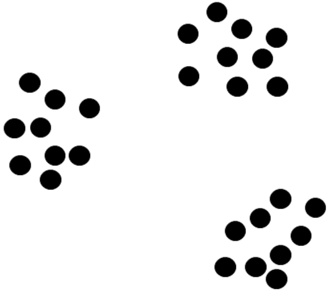
Ragesh Jaiswal, CSE, IIT Delhi

Introduction

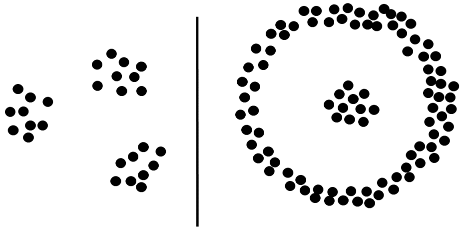
- What is data clustering?
  - Given a *representation* of n objects, find k groups based on a measure of *similarity* (dissimilarity) such that the similarities between objects in the same group are high while similarities between objects in different groups are low.
- Suppose the given objects to be clustered can be *represented* as points in two-dimensional space (i.e., $\mathbb{R}^2$).
  - What is a reasonable notion of *similarity* between objects?

Introduction

- What is data clustering?
  - Given a *representation* of n objects, find k groups based on a measure of *similarity* (dissimilarity) such that the similarities between objects in the same group are high while similarities between objects in different groups are low.
- Suppose the given objects to be clustered can be *represented* as points in two-dimensional space (i.e., $\mathbb{R}^2$).
  - What is a reasonable notion of *similarity* between objects?
    - Distance between points.



- What is data clustering?
  - Given a *representation* of n objects, find k groups based on a measure of *similarity* (dissimilarity) such that the similarities between objects in the same group are high while similarities between objects in different groups are low.
- Suppose the given objects to be clustered can be *represented* as points in two-dimensional space (i.e., $\mathbb{R}^2$).
  - What is a reasonable notion of *similarity* between objects?
    - Distance between points.
    - The notion of similarity/dissimilarity has to be defined carefully.



Data Representation

Data Representation

- What is data clustering?
  - Given a *representation* of n objects, find k groups based on a measure of *similarity* (dissimilarity) such that the similarities between objects in the same group are high while similarities between objects in different groups are low.
- The "data" for cluster analysis can be described by two standard formats:
  - Pattern Matrix
  - Proximity Matrix

Data Representation

- The "data" for cluster analysis can be described by two standard formats:
  - Pattern Matrix
    - Data is represented as an $n \times d$ matrix where each row corresponds to a "pattern/item/object" and each column denotes a "feature/measurement".
    - Example: For patient records in a hospital each row corresponds to a patient and each column denotes a feature such as age, weight, height, measurement for certain medical tests etc.
    - The d features are usually visualised as a set of orthogonal axes. Given this, the items then are points in a d-dimensional space called pattern space.
  - Proximity Matrix

Data Representation

- The "data" for cluster analysis can be described by two standard formats:
  - Pattern Matrix
  - Proximity Matrix
    - This is an $n \times n$ matrix where n denotes the number of items/patterns. The entries in this matrix is called proximity indices. The $(i,j)^{th}$ entry in this matrix denotes the proximity between the i^{th} and j^{th} item.
    - Proximity could indicate similarity or dissimilarity. For example for dissimilarity $D(i,i) = 0$ and for similarity $D(i,i) \geq \max_k D(i,k)$.

Data Representation

- The "data" for cluster analysis can be described by two standard formats:
  - *Pattern Matrix*: $n \times d$ matrix denoting the data points.
  - *Proximity Matrix*: $n \times n$ matrix denoting the pairwise proximity between data points.
- How to interpret the numbers in the above matrices? The numbers can be of the following nature
  - *Nominal*: The numbers are used as names. For example, a yes/no response can be encoded as 0/1 or 500/1000 etc.
  - *Ordinal*: The numbers have meaning with respect to each other. That is, column entries $1, 2, 3$ is equivalent to the column entries $1, 20, 300$.
  - *Ratio scale*: The numbers have absolute meaning. For example, distance between two cities, temperature etc.

Data Representation

- The "data" for cluster analysis can be described by two standard formats:
  - *Pattern Matrix*: $n \times d$ matrix denoting the data points.
  - *Proximity Matrix*: $n \times n$ matrix denoting the pairwise proximity between data points.
- How to interpret the numbers in the above matrices? The numbers can be of the following nature
  - *Nominal*:
  - *Ordinal*:
  - *Ratio scale*:
- Can we obtain the proximity matrix from the pattern matrix?
  - The most common dissimilarity measure using the pattern matrix (with ratio-scaled data) is the *Minkowski* metric which is defined as follows: Let x denote the pattern matrix. Then we have

$$D(i,k) = \left(\sum_{j=1}^{d} |x(i,j) - x(k,j)|^r \right)^{1/r} \quad \text{where } r \geq 1$$

Data Representation

- Can we obtain the proximity matrix from the pattern matrix?
  - The most common dissimilarity measure using the pattern matrix (with ratio-scaled data) is the *Minkowski* metric which is defined as follows: Let x denote the pattern matrix. Then we have

$$D(i,j) = \left(\sum_{j=1}^{d} |x(i,j) - x(k,j)|^r \right)^{1/r} \quad \text{where } r \geq 1$$

- Specific instances of Minkowski metric is given below:
  - *Euclidean distance*: $r = 2$
  - *Manhattan distance*: $r = 1$
  - *Sup distance*: $r \to \infty$. This means that $D(i,k) = \max_{1 \leq j \leq d} |x(i,j) - x(k,j)|$
- The Euclidean distance is the most commonly used distance measure in Engineering.

Data Representation

- Can we obtain the proximity matrix from the pattern matrix?
  - In a number of settings the pattern matrix contains binary nominal values (i.e., 0/1 indicating yes/no)
  - In such cases the proximity index $D(i, j)$ is calculated in the following manner:
    - Let $a_{00} = |\{k : x(i, k) = 0 \text{ and } x(j, k) = 0\}|$
    - Let $a_{01} = |\{k : x(i, k) = 0 \text{ and } x(j, k) = 1\}|$
    - Let $a_{10} = |\{k : x(i, k) = 1 \text{ and } x(j, k) = 0\}|$
    - Let $a_{11} = |\{k : x(i, k) = 1 \text{ and } x(j, k) = 1\}|$
  - *Simple Matching Coefficient*: $D(i, j) = \frac{a_{00}+a_{11}}{a_{00}+a_{11}+a_{01}+a_{10}}$
  - *Jaccard Coefficient*: $D(i, j) = \frac{a_{11}}{a_{11}+a_{01}+a_{10}}$

Data Representation

- What can we do for "missing data"?
  - In a number of settings some data entries might be missing. For example, missing medical test for some individual etc.
  - Missing data is handled in the following manner:
    - Delete the items or features that have missing entries.
    - Suppose the j^{th} entry of the i^{th} item is missing. Find the k "nearest neighbours" of the i^{th} item and replace the the missing entry with the average value of the j^{th} feature of these nearest items.
    - Skip the missing features while calculating the distance between pair of items.
    - Try computing the missing entries by assuming certain reasonable properties of the pattern matrix.

Data Representation

- How do we normalize across different features?
  - Some features might be recorded using a larger range of numbers (e.g. distance in inches) compared other features (e.g., distance in miles). When calculating dissimilarity using Minkowski metric, one feature might dominate the dissimilarity.
  - Here is the standard way to normalise the $n \times d$ pattern matrix x. Let y denote the normalised matrix.
    - For all j, let $m_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$
    - For all j, let $s_j^2 = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - m_j)^2$
    - $y_{ij} = \frac{x_{ij} - m_j}{s_j}$
  - All features in y have 0 mean and unit variance.
- Is such normalisation always desirable when used for clustering purposes?
  - Can you think of an example?

Data Representation

- Can we reduce the dimensionality of the data?
    - Discussed later

The k-means Problem

k-means

Given a set of points $S \subset \mathbb{R}^d$ in a d dimensional Euclidean space, and an integer k, output a set $T \subset \mathbb{R}^d$ of points (called *centers*) such that $|T| = k$ and the following cost function is minimised:

$$cost(S, T) = \sum_{x \in S} \min_{z \in T} ||x - z||^2.$$

Figure : What is the solution for the 2-means problem for the above 2-D dataset?

The k-means Problem

k-means

Given a set of points $S \subset \mathbb{R}^d$ in a d dimensional Euclidean space, and an integer k, output a set $T \subset \mathbb{R}^d$ of points (called *centers*) such that $|T| = k$ and the following cost function is minimised:

$$cost(S, T) = \sum_{x \in S} \min_{z \in T} ||x - z||^2.$$

- How hard is the 1-means problem?
  - Given $x_1, ..., x_n \in \mathbb{R}^d$ find a point $z \in \mathbb{R}^d$ such that $f(z) = \sum_i ||x_i - z||^2$ is minimized.

The k-means Problem

k-means

Given a set of points $S \subset \mathbb{R}^d$ in a d dimensional Euclidean space, and an integer k, output a set $T \subset \mathbb{R}^d$ of points (called *centers*) such that $|T| = k$ and the following cost function is minimised:

$$cost(S, T) = \sum_{x \in S} \min_{z \in T} ||x - z||^2.$$

- How hard is the 1-means problem?
  - Given $x_1, ..., x_n \in \mathbb{R}^d$ find a point $z \in \mathbb{R}^d$ such that $f(z) = \sum_i ||x_i - z||^2$ is minimized.
  - What is $\frac{\partial f(z)}{\partial z_i}$?
  - So, for what z, $\sum_i ||x_i - z||^2$ gets minimized?

The k-means Problem

k-means

Given a set of points $S \subset \mathbb{R}^d$ in a d dimensional Euclidean space, and an integer k, output a set $T \subset \mathbb{R}^d$ of points (called *centers*) such that $|T| = k$ and the following cost function is minimised:

$$cost(S, T) = \sum_{x \in S} \min_{z \in T} ||x - z||^2.$$

- How hard is the 1-means problem?
  - Given $x_1, ..., x_n \in \mathbb{R}^d$ find a point $z \in \mathbb{R}^d$ such that $f(z) = \sum_i ||x_i - z||^2$ is minimized.
  - What is $\frac{\partial f(z)}{\partial z_i}$? $\frac{\partial f(z)}{\partial z_i} = nz_i - \sum_j x_{ji}$
  - So, for what z, $\sum_i ||x_i - z||^2$ gets minimized? $z = \frac{\sum_i x_i}{n}$
  - $\frac{\sum_i x_i}{n}$ is called the *centroid* of the points $x_1, ..., x_n$.

The k-means Problem

k-means

Given a set of points $S \subset \mathbb{R}^d$ in a d dimensional Euclidean space, and an integer k, output a set $T \subset \mathbb{R}^d$ of points (called *centers*) such that $|T| = k$ and the following cost function is minimised:

$$cost(S, T) = \sum_{x \in S} \min_{z \in T} ||x - z||^2.$$

- How hard is the k-means problem for $k > 1$ when $d = 1$? In other words, how hard is the 1-dimensional k-means problem?

The k-means Problem

k-means

Given a set of points $S \subset \mathbb{R}^d$ in a d dimensional Euclidean space, and an integer k, output a set $T \subset \mathbb{R}^d$ of points (called *centers*) such that $|T| = k$ and the following cost function is minimised:

$$cost(S, T) = \sum_{x \in S} \min_{z \in T} ||x - z||^2.$$

- How hard is the k-means problem for $k > 1$ when $d = 1$? In other words, how hard is the 1-dimensional k-means problem?
  - There is a simple Dynamic Programming algorithm for this problem!

The k-means Problem

k-means

Given a set of points $S \subset \mathbb{R}^d$ in a d dimensional Euclidean space, and an integer k, output a set $T \subset \mathbb{R}^d$ of points (called *centers*) such that $|T| = k$ and the following cost function is minimised:

$$cost(S, T) = \sum_{x \in S} \min_{z \in T} ||x - z||^2.$$

- How hard is the k-means problem for $k > 1$ and $d > 1$?
  - NP-hard.

End