COL870: Clustering Algorithms

Ragesh Jaiswal, CSE, IIT Delhi

Administrative Information

Administrative Information

- Instructor
  - Ragesh Jaiswal
  - Office: 403, SIT Building
  - *Email*: `rjaiswal@cse.iitd.ac.in`
- Please send email to set up a meeting with me.

Administrative Information

- Grading Scheme
  1. *Homework*: 25 (12.5 Theory and 12.5 programming)
  2. *Minor 1 and 2*: 15 points each.
  3. *Major*: 25 points.
  4. *Project*: 18 points.
  5. *Attendance*: 2 points.
- Policy on cheating:
  - Anyone found using unfair means in the course will receive an **F** grade.

Administrative Information

- Textbook: There are no textbooks for this course. This is an advanced level course. The reference material will include book chapters, course notes, research papers, and other material available online.

- Course webpage: `http://www.cse.iitd.ac.in/~rjaiswal/2015/col870`.
    - The site will contain course information, references, and announcements. Please check this page regularly.

Introduction

Introduction

- What is data clustering?

Introduction

- What is data clustering?
  - The task of grouping a set of objects such that objects in the same group (called *cluster*) are more similar than objects in different groups.
  - 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.

Introduction

- Why study data clustering?
  - Custering is usually the first step when trying to make sense of Big Data.
  - Common tool across many different fields such as image analysis, pattern recognition, bioinformatics, information retrieval etc.
- Data clustering has been used for the following three main purposes [Jain'09]:
  - *Underlying structure*: to gain insight into data, generate hypotheses, detect anomalies, and identify salient features.
  - *Natural classification*: to identify the degree of similarity among forms or organisms.
  - *Compression*: as a method for organizing the data and summarizing it through cluster prototypes.

Introduction

- How is clustering different from “classification”?

Introduction

- How is clustering different from "classification"?

Classification	**Clustering**
The task of assigning objects to predefined "classes".	The task of grouping objects without any knowledge of the "classes".

- The more general terms used are *supervised* learning versus *unsupervised* learning.

Supervised Learning	**Unsupervised Learning**
- Classes/labels known	- Classes/labels unknown
- Expert available	- No experts
- Training data available	- No training data

Introduction

Frequently Asked Questions

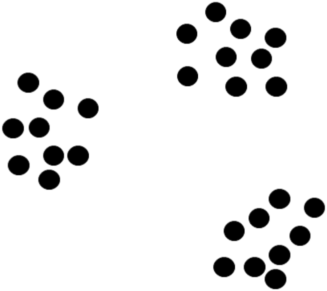
- What are the pre-requisites for the course?
  - Answer: Data Structures and Algorithms. I will assume that you are familiar with Algorithm design and analysis techniques and intractability concepts like NP-completeness etc.
- What flavour will this course have, algorithms or machine learning?
  - Answer: The course will have the rigour of typical algorithms courses. That is, we will spend time proving things. However, techniques we learn in the course will have applications in many different areas.
- Does the course involve programming?
  - Answer: Depends on the interest of the class. Our focus will be to understand clustering techniques. However, if the class is interested, I can set up programming assignments.
- Does the course involve a project?
  - Answer: Yes. You will be asked to read and present research papers on clustering topics towards the later part of the course.
- Do we become more *marketable* after doing all this?
  - Answer: Perhaps yes!

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?

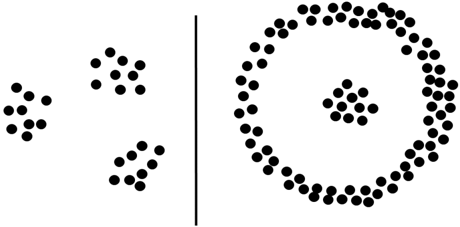
Introduction

- 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.



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.
    - The notion of similarity/dissimilarity has to be defined carefully.



End