Graded Lab Assignment 1: Clustering

Implement the *k*-means clustering algorithm from scratch. You can use any programming language of your choice. Using a library to perform your clustering does **not** count for any credit.

The goal of this lab assignment is to implement the *k*-means algorithm so that you have a deep understanding of the algorithm.

Please do go through the lecture material in detail before you start on your implementation to ensure that you fully understand the algorithm. Please run your program with different values of *k* and with progressively larger datasets and choose the *k* initial clusters in different ways to get a better perspective on how the *k*-means algorithm works.

Test the algorithm using any spatial dataset that is freely available in the public domain. Lots of spatial datasets are downloadable from websites such as:

- http://freegisdata.rtwilson.com/
- http://water.usgs.gov/lookup/getgislist
- http://www.diva-gis.org/Data
  [The above are just examples of sites where you can download spatial datasets. Please feel free to download any spatial data that is freely available in the public domain and is legal to download i.e., no copyright restrictions etc.]

Deliverable: Your k-means clustering program should be able to run with at least 10,000 spatial points. You are required to provide a demo of your program and then run it on a testing dataset provided. **Slots for program demos & testing will be communicated to you by your TA.**

Please note the following points:
1. This lab assignment will contribute to **5% of your grades** for the course.
2. Please write the Names and Roll Numbers of your group members in the Comments section at the beginning of your program file. All programs should be submitted before the stated deadline. Please note that your programs may be subjected to plagiarism checks.
3. The **deadline** for assignment submission is **March 1, 2021, 11.59 pm IST**.
4. This is a **HARD deadline** and no points will be awarded for the assignment if you submit after the deadline, unless there are extenuating circumstances.
5. You need to show the demo of your program running with at **least 10,000 points in 2-dimensions**. Of course, you should try to run your program with a much larger number of points, but 10,000 points is a bare minimum criterion for the completion of this assignment.
6. **While you're expected to build this using any publicly available dataset, a separate dataset will be used for the testing.** The TAs will share the schema of the dataset & clarify and questions about file format about the testing dataset well before the deadline so that you can adjust your model accordingly. **(Basically, be prepared to read in a dataset from a file and run your clustering model on it for different values of K)**

7. The grading criteria for this assignment will be based on effort, adherence to learning points from your previous non-graded assignment, code quality, visualization, results and scalability.

8. ***Any act of plagiarism will result in a zero for the entire assignment. Hence, please avoid any form of plagiarism.***

9. This is a group assignment, hence please do NOT collaborate with your fellow students in other groups towards the completion of this assignment. However, you are obviously required to collaborate effectively with members of your own group to ensure that you are able to function as a team player.

Graded Lab Assignment 2: Association Rule Mining

Implement the association rule mining algorithm discussed in Unit 2. Please do go through the lecture in detail before you start on your implementation to ensure that you fully understand the algorithm. Please run your program with different values of the input parameters and with progressively larger datasets to get a better perspective on how the association rule mining algorithm works.

Test the algorithm using a randomly generated input dataset.

Deliverable: You are required to provide a demo of your program on or before the deadline specified below in this document.

Please note the following points:

1. This lab assignment will contribute to **10% of your grades** for the course.

2. Please write the Names and Roll Numbers of your group members in the Comments section at the beginning of your program file. All programs should be submitted on or before the stated deadline. Please note that your programs may be subjected to plagiarism checks.

3. The **deadline** for assignment submission is **March 24, 2021 (**The deadline time is **11.59 pm IST**.)

4. This is a **HARD deadline** and no points will be awarded for the assignment if you submit after the deadline, unless there are extenuating circumstances.

5. You need to show the demo of your program. Of course, you should try to run your program by progressively increasing the size of the input dataset.

6. The grading criteria for this assignment will be based on effort, code quality, visualization, results and scalability.

7. Please avoid any form of plagiarism.

8. This is a group assignment, hence please do NOT collaborate with your fellow students in other groups towards the completion of this assignment. However, you are obviously required to collaborate effectively with members of your own group to ensure that you are able to function as a team player.

Data Mining & Warehousing

Lab Assignment – 3 (Mini-project on Visualization and Decision Support)

In this assignment, you will need to programmatically access any e-travel website (such as www.booking.com, agoda.com, yatra.com etc.) to mine data about hotel reviews, statistically analyze the data and visually present your results.

The goal of this assignment is to ensure that you are trained in handling real-world data. Of course, when you deal with real-world data, you will need to work on peripheral activities such as data cleaning, pre-processing the data etc., before you do data mining. This assignment will also train you to deal with unstructured text data. You will also be learning Weka while doing this assignment. (Materials about Weka will be provided and discussed in class.) You will also receive exposure to some important technologies such as Nominatum / MapBox, NLTK etc.

Visit any ONE of these websites and mine the listed data points:

1. The rating for a hotel (Overall Rating)
2. List of comments for a given hotel + rating provided (In reverse chronological order. i.e. newest first)
3. Be sure to mine at least 100 comments for each hotel.
4. Mine the information for 20 hotels in Bangkok, Singapore & Kuala Lumpur **each.**

   (Thus in total, you will need to mine: 20 (hotels) x100 (comments each) x3 (cities) = 6k comments)

   Save all your mined data in an appropriate format.

After you've completed mining your data, complete the following analysis:

1. Cluster the hotels that have similar ratings (overall rating)
   a. You can use WEKA (or any free software) for performing as well as visualizing the clustering
2. Perform sentiment analysis on the comments.
   a. You can use NLTK's inbuilt polarity score to assign a score to each comment.
   b. Check the co-relation between the Polarity score of each comment and the actual rating provided by the user (the rating linked with that comment).
   c. Aggregate (mean/median) the polarity score of all the comments and calculate an overall polarity score for each hotel. Then check for any co-relation between the overall polarity score and the overall rating of the hotel.
3. Visualize the above results. Be sure to highlight any particular anomalies in your visuals.
4. Use any mapping resource (like Nominatim or MapBox) to geospatially visualize the top 10 hotels in each of the 3 cities. Make 2 maps based on the two scoring mechanisms available:
   a. Top 10 hotels based on overall rating of the e-travel website
   b. Top 10 hotels based on overall calculated polarity score

For this assignment, submit:

- Your final mined dataset.

- A two-page report that contains your visualizations with interesting insights. PDF only. Times New Roman 11 point font size.

You will need to demo your code & display your visualizations in a code review session.

This assignment is worth 20 marks.

Deadline for submission: **11:59 PM IST on 17th April, 2021.**

Data Mining & Warehousing

Assignment on Star Schema

Required: Create a sample star schema design for organizing SALES data for the following scenarios:

- a fashion retail chain such as Inditex-Zara
- a fast-food restaurant such as KFC or McDonalds
- an automobile company such as Honda, Mahindra or Nissan

Note:

- This is an open-ended question, hence please feel free to make any reasonable assumptions.
- This is a group assignment.
- Please submit a single PDF file only.
- This assignment carries 10 marks.
- The upload link will be sent to you.

Deadline for submission: **11:59 PM IST on 20th April, 2021**

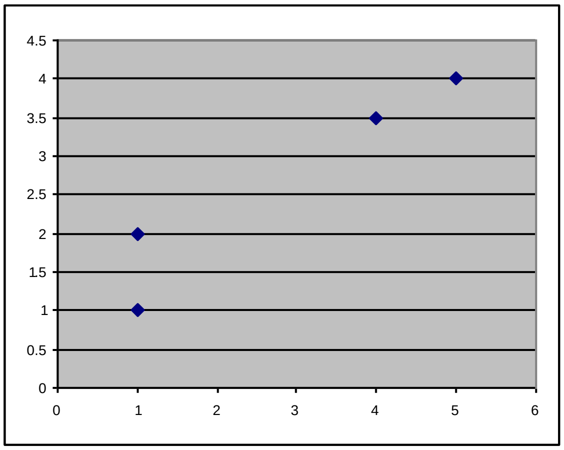
Data Mining & Data Warehousing Course -- Mid-sem Exam

Instructions

1. Answer ALL questions.
2. Please write your NAME and ROLL NUMBER clearly in your answer script.
3. Time limit for this exam is 70 minutes.
4. Once you have completed the exam, please upload your file (either a PDF file or a scanned image file) to the link, which has been sent by your TA. The marks for each question are indicated in brackets.
5. Your answers should be concise, but should cover ALL the key points. I will grade based on the content of your answers (not the length of your answers).
6. This is an open-book exam i.e., you can look at the lecture slides, text books, your own notes etc.
7. Please avoid any discussions among yourselves during the duration of the exam.

QUESTIONS

1. Show the execution of the k-means clustering algorithm on the following data, where k =2. Show the relevant steps clearly and in as much detail as you can. [6]



2. List any three scenarios where association rule mining would be ineffective i.e., in such scenarios, association rule mining would provide results that are not useful for decision-making by the concerned stakeholders. [6]

3. How would you use data mining and analytics to improve customer satisfaction at a brick-and-mortar retail store? First, define the context of the application scenario. You can make any reasonable assumptions. Discuss how you would proceed for each phase of data mining i.e., starting from data collection all the way to the interpretation of results. Clearly mention which data mining algorithms/techniques (e.g., association rule mining, clustering etc.) you would deploy and justify why your chosen algorithm is a good fit to the given application scenario. [8]

Note: This is an open-ended question. Please feel free to make any reasonable assumptions.

Data Mining & Data Warehousing Course -- End-sem Exam

Instructions

1. Answer ALL questions.
2. Please write your NAME and STUDENT ID clearly in your answer script.
3. Time limit for this exam is 3 hours.
4. Once you have completed the exam, please upload your file (either a PDF file or a scanned image file) to the link, which has been sent by your TA. The marks for each question are indicated in brackets.
5. Your answers should be concise, but should cover ALL the key points. I will grade based on the content of your answers (not the length of your answers).
6. This is an open-book exam i.e., you can look at the lecture slides, text books, your own notes etc.
7. Please avoid any discussions among yourselves during the duration of the exam.

QUESTIONS

1. Create a sample star schema design for organizing the data concerning placements for the training and placement division of a University: the training and placement division in a University is responsible for career development of students e.g., by connecting students and companies to each other. [This is an open-ended question, hence please feel free to make any reasonable assumptions.] [10]

2. How would you use data mining and analytics to improve some aspect of Covid-19 pandemic management during these unprecedented pandemic times?

   First, define your goals clearly in terms of which aspect(s) of Covid-19 management you would like to address. You can make any reasonable assumptions. Discuss how you would proceed for each phase of data mining i.e., starting from data collection all the way to the interpretation of results. Clearly mention which data mining algorithms/techniques (e.g., association rule mining, clustering etc.) you would deploy and justify why your chosen algorithm is a good fit to the given application scenario. [10] [**Note:** This is an open-ended question. Please feel free to make any reasonable assumptions.]

3. Consider the scenario of a large retail supermarket selling a large number of items. Each item has a price and a frequency of sales. These items differ in terms of their expiry deadlines e.g., dairy products expire within days, some products can take 1-2 weeks to expire and some products expire after multiple months. You can consider N categories of products, each category corresponding to a specific range of expiry deadline times. Furthermore, you can consider N different time periods, and each category of products expires in only one of these time periods.

   Design an algorithm to create itemsets in the afore-mentioned scenario and place the itemsets (in the retail store slots) such that the revenue of the retailer is maximized as far as possible, while ensuring that items closer to expiry deadlines are placed in the retail store slots before these items expire. [10] [**Note:** This is an open-ended question. Please feel free to make any reasonable assumptions.]