Amirkabir University of Technology
(Tehran Polytechnic)

Assignment 2
Data come with high speed as stream of a storm!

Homeworks Guidelines and Policies

- It is expected that the students submit an assignment report (HW2_[student_id].pdf) as well as required source codes (.m or .py) into an archive file (HW2_[student_id].zip). Please combine all your Persian reports just into a single .pdf file by problems order. Code without report has an exact zero point.
  Please do not use implementation tools when it is asked to solve the problem by hand, otherwise you will be penalized and lose some points.
- You are free to solve by-hand problems on a paper and include their pictures in your report. Here, cleanness and readability are of high importance. Images should also have appropriate quality.
- Your work will be evaluated mostly by the quality of your report. Do not forget to explain your answers clearly, and provide enough discussions when needed.
- In each homework, 4 points (out of a possible 100) belong to compactness, expressiveness, and neatness of your report and codes.
- By default, we assume you implement your codes in Python. If you are using MATLAB or R, you have to use the equivalent functions when it is asked to use specific Python functions.
- Your codes must be separated for each question.
- Make sure you have access to Courses, because that is where all assignments as well as course announcements are posted. Homework submissions are only made through Courses.

- Please submit your work **before the end of May 16th**.
- There will be a 10% penalty for every late day.
- You are not allowed to share codes/answers or use works from the past semesters. Violators will receive a zero for that particular problem.
- If there is any question, please do not hesitate to contact us through the Telegram group chat or following email addresses: **m.ebadpour@aut.ac.ir**, **faramarz.aghajani@aut.ac.ir**, and **danial.alizadeh@aut.ac.ir**.

Sec. 1) Locality Sensitive Hashing (LSH) (32 Pts.)

Three questions are asked in this section. For the programming part, a data set has been provided to you, which is actually a continuation of the data set related to the first exercise. Please read the questions carefully and do the requested things for each one; You may need to explain more about some parts.
Note that in this question you should not use the library ready to implement LSH and you should write its functions yourself; but you can write the functions as you like and you don't need to follow the solution mentioned in the booklet.

1. Name the steps of LSH and briefly explain each one.
2. Name one of the disadvantages of the LSH method and state its cause.
3. In this question, we are going to use a huge dataset that contains information about customer purchases from an e-commerce platform. Each row in the dataset represents a single purchase and includes details such as customer ID, product category, product name purchased, price, and timestamp.
    a) Print the groups and names of goods in this data set and bring them in your report (avoid bringing duplicate names).
    b) Find 10 of the most similar customers to the desired 10 customers and bring them in your report as follows (you can use any criterion to find similarity, but you must explain why you chose that criterion).

        Top 10 similar customers for customer # :
        Customer ID : #, Similarity : #
        ….

    c) For every 10 desired customers of part b, offer products and state according to which customer similar to this customer, you have offered this product.

        Recommended products for customer #:

        Product: Clothing_Suit
        Customers who purchased this product:
        Customer ID: #
        Customer ID: #
        Customer ID: #

        Product: Sports & Outdoors_Boxing Gloves
        Customers who purchased this product:
        Customer ID: #
        Customer ID: #
        Customer ID: #

Corresponding TA: Alizadeh

Sec. 2) Streams are more than words! (32 Pts.)

1. In this part, we want to use the DGIM algorithm. Below is a data stream with a window size of 25 and current bucketing.

   a) The following bits enter the window, one at a time: 1 1 0 1 1 0 1. What is the bucket configuration in the window after this sequence of bits has been processed by DGIM?
   b) After having processed the bits from (a), what is now the estimate of the number of 1's in the latest 20 bits of the window?

2. Suppose the stream is 1,3,2,4,2,4,1,3,1,1,3,4,2,2,1:
   a) Compute zeroth moment for this stream.
   b) Compute first moment for this stream.
   c) Compute second moment for this stream.
   d) Suppose we keep three variables X1 , X2 and X3. Also assume that at random we pick the 2nd, 7th and 10th positions to define these three variables. Estimate the second moment of this data stream using the AMS algorithm.

3. In this exercise, we want to use the Flajolet-Martin method in estimating the number of distinct elements. Suppose our stream consists of the integers 5, 1, 2, 3, 3, 4, 2, 5, 1. You should treat the result as a 4-bit binary integer. Determine the tail length for each stream element and the resulting estimate of the number of distinct elements if the hash function is:
   a) $h(x) = 4x + 1 \; mod \; 16$
   b) $h(x) = 3x + 6 \; mod \; 16$

4. In this exercise, we ask you to implement the DGIM algorithm and use it to estimate the number of 1-bits in the window. The binary data stream is provided in the 'data_stream.txt' file.
   a) Set the window size to 2000 and count the number of 1-bits within the current window.
   b) Write a function that accurately calculates the number of 1-bits in the current window.
   c) At each moment, calculate the accuracy of the DGIM algorithm estimation. What were the highest and lowest accuracy values observed during the processing of this data stream?
   d) Generate three binary data streams of 100000 bits randomly (One stream should have twice as many 1s as 0s, another should have twice as many 0s as 1s, and the third should have an equal number of 0s and 1s).
   e) Process each of the data streams in part "d" with the DGIM algorithm and exact method, with a window size of 10000. determine the accuracy and execution time in each one.

Corresponding TA: Aghajani

Sec. 3) Rhythms in Data: Exploring Music Clustering Techniques with Spotify (32 Pts.)

Clustering in the music market is pivotal for understanding user preferences, organizing vast collections of songs, and enhancing recommendation systems. By grouping similar songs together, clustering algorithms enable you, as the project lead, to provide personalized recommendations, improve user experience, and optimize resource allocation. In this project, you will delve into the realm of music clustering using a dataset from Spotify. The goal is to explore different clustering techniques, preprocess the data for analysis, and evaluate the performance of various algorithms.

a. **Dataset Exploration and Preprocessing**: At the project's outset, you will delve into the Spotify dataset to thoroughly understand its structure and contents. Through this exploration, you will identify columns containing redundant or irrelevant information that can be safely removed to streamline the dataset. By removing such columns, your aim is to reduce noise and enhance the efficiency of subsequent analysis. Additionally, you will perform preprocessing steps such as feature normalization to ensure that the data is in a suitable format for clustering algorithms.
b. **Comparison of BFR and CURE:** In this section, you will apply two popular clustering algorithms, BFR and CURE, to the preprocessed Spotify dataset. You will then evaluate the performance of these algorithms in terms of accuracy and runtime. By comparing the results obtained from both algorithms, you aim to gain insights into their effectiveness and efficiency in clustering music data. Your implementation should be from scratch, and the target column for clustering is 'genre.'
c. **MapReduce-Based K-Means Clustering:** Now, let's consider a scenario where multiple data centers collaborate to perform K-means clustering on the Spotify dataset while ensuring privacy and data security. You will propose a MapReduce-based approach to implement K-means clustering, where each song belongs to a specific data center accessible only by authorized personnel. Explain how centroids can be updated using MapReduce structures and implement this approach on the given dataset. Then, compare the accuracy and complexity of this method with BFR and CURE, shedding light on its effectiveness in privacy-preserving music clustering. 'Key' indicates the data center identification ID.
d. **Choice of Distance Measure:** The choice of distance measure plays a pivotal role in clustering algorithms as it determines the similarity between data points. In this project, you will carefully select a suitable distance measure based on the characteristics of the Spotify dataset and the clustering task at hand. Justify your choice of distance measure and highlight its importance in ensuring the accuracy and effectiveness of the clustering process.

Note 1: Assumptions are made regarding the exact value of K and the use of MapReduce for privacy-preserving clustering. Initialization of centroids is performed independently for each data center.
Note 2: Various strategies can be employed, such as hierarchical, loop, or recursive forms of MapReduce, to initialize centroids on each data center.
Note 3: You may use built-in visualization libraries and techniques like t-SNE to compare the quality of clustering. You can color each point by its ground truth.

Corresponding TA: Ebadpour