CSCI 4380/6380 – Data Mining Fall 2023

Assignment 4 Due date: November 20th, 11:59pm, EST

---

1. (20 points) Suppose that the data mining task is to cluster points (with (x,y) representing location) into three clusters, where the points are

A1(2,10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9).

The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show only

(a) The three cluster centers after the first round of execution.

(b) The final three clusters.

2. (10 points) Both k-means and k-medoids algorithms can perform effective clustering.

(a) Illustrate the strength and weakness of k-means in comparison with k-medoids.

(b) Illustrate the strength and weakness of these schemes in comparison with a hierarchical clustering scheme.

3. (30 points) Programming task.

The Banknote Dataset [the dataset is available here: https://archive.ics.uci.edu/dataset/267/banknote+authentication] involves predicting whether a given banknote is authentic given a number of measures taken from a photograph. There are 1,372 observations with 4 feature variables and 1 label variable.

Assuming we do not know the labels. Perform different clustering algorithms based on this dataset and perform evaluation.

(a) Implement k-means algorithm as a function. (10 points) Set k=2 and obtain the clustering results, evaluate the results with Purity and Normalized Mutual Information (NMI). (5 points)
(b) Using existing functions from sk-learn to obtain clustering results with DBSCAN, GMMs, and Spectral Clustering when k=2. (10 points). Compare their results with Purity and Normalized Mutual Information (NMI). (5 points)

4. (40 points) Open question:

ChatGPT (https://chat.openai.com/chat) is a large language model developed by OpenAI, based on the GPT-3.5/GPT-4 architecture. It is a machine learning model that has been trained on a vast amount of text data from the internet, allowing it to understand and generate human-like language.

ChatGPT can be used for a variety of natural language processing (NLP) tasks, such as language translation, text summarization, and text completion. Additionally, it can be used for conversational AI applications, where it can generate responses to user input in a way that mimics natural human conversation.

Prompt engineering is the practice of giving an AI model specific instructions to produce the results you want. A prompt is a sequence of text or a line of code that can trigger a response from an AI model.

Please show a demo of how ChatGPT can help you with a data mining project, i.e., collect the data, preprocess the data (optional), analyze the data (modeling), and present the results. During your exploration, is there any limitations for ChaGPT?