Technion – Israel Institute of Technology Spring 2024
The Faculty of Data and Decision Sciences Homework 2

096224: Distributed Database Management

Part 1: K-Means Algorithm Assignment (50 Points)

Recall K-Means clustering, which partitions N observations into K clusters. Reminder of the algorithm:

1. Initialization - use K random or given points as the initial centroids.
2. Training Loop -
   - Assign each point in the data to the closest centroid's cluster, using Euclidean distance.
   - Recalculate each cluster's centroid - replace each centroid with the average of the points in the cluster it represents.
   - Iterate until convergence (no change in cluster assignments) or until max iterations number reached (hyperparameter)
3. Termination - output the final **K** centroids

Spark Implementation (50 Points)

Write your own implementation of K-Means algorithm, using PySpark. Your algorithm should utilize the following function declaration:

$$kmeans\_fit(data, init, k, max\_iter)$$

Input:

- data : PySpark DataFrame, which contains N real records of size d (i.e. shape of (N, d)).
- init : A PySpark Dataframe containing K rows, each entry is an initial centroid (same dimension as data).
- K : Number of clusters.
- max_iter : Maximum number of iterations.

Output:

- The function should return a PySpark DataFrame containing the final clusters centroids.
- We expect your output dataframe to contain a single column named 'centroids', and each row to contain a single dense Vector representing a centroid of a single cluster.

Use Euclidean Distance as your distance metric. Your algorithm should stop when reached convergence (i.e. the assignments no longer change) or maximum iterations exceeded. Notice that the algorithm is deterministic, and we **expect your results to be the same as given in the example for at least 3 decimal numbers** after the point.

Competitive Benchmarks (Bonus up to 8 Points)

Your algorithm will be measured in time complexity. This is a competition, the top fastest teams will get bonus points according to their results.

Part 2 (50 Points)

Please see the attached file: HW2_P2_S2024.pdf

General Guidelines

- Your final submission should be composed of a total of five files:
  1. **The filled Starter notebook** (see below) with your function implementation and outputs, named HW2_STARTER_[ID1]_[ID2].

     You should submit **3 versions:** a **.ipynb, PDF and HTML INCLUDING your outputs and results** from running your code on the **whole/large** data.
  2. **Python file (.py) named HW2_WET_[ID1]_[ID2].py with your function**, named exactly as requested in this file and as can be seen in the Starter notebook.
  3. **PDF file (.pdf) named HW2_DRY_[ID1]_[ID2].pdf** with your answers to the theoretical questions from the other file.

  No zip file.
- Questions related to this assignment will be answered in the forum, via Moodle, exclusively.
- Only one of the team members needs to submit.
- Submission is due to date stated in Moodle.
- There are 48 hours of automatic extension to the due date on Moodle. Extension requests that will be sent after the official due-date will be ignored. This automatic extension is nullified in case of a group-specific extension. Rules are detailed in Moodle forum.
- Any further delay in submission will result in a reduction of 20 points from the final score of this assignment.
- The use of generative AI tools such as ChatGPT is permitted. However, we recommend to use it sporadically and only after you have given the code a try yourself. Keep in mind that ChatGPT is not a legitimate helper during your final exam. Should you choose to use generative AI, please indicate clearly where it was used, in what manner did you decide to use it (e.g., which prompts were used) and to what extent was it helpful (e.g., what was the amount of work you needed to invest to correct it). Note that reporting on the use of said tool will not affect your grade in any way.

Link to the starter Colab notebook -

`https://colab.research.google.com/drive/1XFLmS-gSUzsN96EfrfjQ6CdvDodIREoj?usp=sharing`

Good luck,
Course staff.