096224: Distributed Database Management

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

Recall K-Means clustering, which partitions N observations into k clusters.

Dropout Variation

Dropout variation of K-Means utilize a dropping-out parameter, q, which ignores each q^{th} observation within the cluster, during centroid calculation, while the observations are sorted in ascending order, using Euclidean distance from the centroid (i.e. the closest observations is first).

Spark Implementation (35 Points)

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

kmeans_fit(data, k, max_iter, q, init), where:

- data : PySpark DataFrame, which contains N real records of size d (i.e. shape of (N, d)).
- k : Number of clusters.
- max_iter : Maximum number of iterations.
- q : Dropout parameter (see Dropout Parameter).
- init : A list of k PySpark Rows, containing initial centroids (same dimension as data).

The function should return a PySpark DataFrame containing the clusters centroids.

Compute objective function using Euclidean Distance as distance metric.

Your algorithm should stop when reached convergence (i.e. the assignments no longer change) or maximum iterations exceeded.

Competitive Benchmarks (15 Points)

Your algorithm will be measured in time complexity.

Guidelines

- Use *Spark 3* and above.
- Data dimension shouldn't be fixed.
- Write clean code and document functions when necessary.
- Sample output will be given in the next days.

Part 2 (50 Points)

Please see the attached file: HW2_P2_S2022.pdf

General Guidelines

- Your final submission should be composed of two files:
  1. Python file (.py) named HW2_KMEANS_[ID1]_[ID2].py
  2. Pdf file (.pdf) named HW2_THEORY_[ID1]_[ID2].pdf
  No zip file.
- Questions related to this assignment will be answered in the forum, via Moodle, exclusively.
- Only one of the team members need to submit.
- Submission is due to June 15^{th}, any delay in submission will result in a reduction of 20 points from the final score of this assignment.

Good luck,
Course staff.