Homework 4

Q1

100 Points

Description:

Overview and Assignment Goals:

The objectives of this assignment are the following: Implement the K-Means Algorithm Deal with Image data (processed and stored in vector format) Think about Best Metrics for Evaluating Clustering Solutions

Detailed Description:

- There are 2 leaderboard submissions for this homework (they will appear as separate parts on Miner). HW3 - Iris is an easier dataset where K-Means can be tested quickly. The second part involves dealing with image data*

For this assignment, you are required to implement the K-Means algorithm. Please do not use libraries for this assignment except for pre-processing the datasets.

Part 1

The famous Iris dataset serves as an easy benchmark for evaluation. Test your K-Means Algorithm on this easy dataset with 4 features:

sepal length in cm sepal width in cm petal length in cm petal width in cm and 150 instances.

Assign the 150 instances in the test file to 3 cluster ids given by 1, 2 or 3. The leaderboard will output the V-measure and this benchmark can be used as an easy step for the main HW3.

The training data is a NULL FILE.

The file iris_new_data.txt, under "Test data," contains the data you use for clustering.

The format example is given by iris_format.txt.

Part 2

Input Data (provided as new_test.txt) consists of 10,000 images of handwritten digits (0-9). The images were scanned and scaled into 28x28 pixels. For every digit, each pixel can be represented as an integer in the range [0, 255] where 0 corresponds to the pixel being completely white, and 255 corresponds to the pixel being completely black. This gives us a 28x28 matrix of integers for each digit. We can then flatten each matrix into a 1x784 vector. No labels are provided.

Format of the input data: Each row is a record (image), which contains 784 comma-delimited integers. Examples of digit images can be found at
http://cs.gmu.edu/~sanmay/ImageExamples.png
For evaluation purposes (Leaderboard Ranking), we will use the V-measure in the sci-kit learn library that is considered an external index metric to evaluate clustering. Essentially your task is to assign each of the instances in the input data to K clusters identified from 1 to K.

For the leaderboard evaluation set K to 10. Submit your best results. The leaderboard will report the V-measure on 50% of sampled dataset.

Some things to note:

-- As usual, the public leaderboard shows results for 50% of randomly chosen test instances only. This is a standard practice in data mining challenges to avoid gaming of the system.

-- In a 24-hour cycle you are allowed to submit a clustering solution 10 times only. Only one account per student is allowed.

-- The final ranking will be based on the last submission.

-- format.txt shows an example file containing 10,000 rows with random class assignments from 1 to 10.

Rules: -- This is an individual assignment. Discussion of broad level strategies is allowed but any copying of submission files and source code will result in an honor code violation. Similarly, it's not acceptable to copy code from the internet, even if you cite the source. Doing so will result in an honor code violation.

-- Feel free to use the programming language of your choice for this assignment.

-- While you can use libraries and templates for dealing with input data you should implement your own K-Means algorithm.

Deliverables: -- Valid submissions to the Miner Website -- Gradescope submission of source code and report: Create a folder called HW4_LastName. Create a subfolder called src and put all the source code there. Create a subfolder called Report and place a maximum 4 page, single-spaced report describing details regarding the steps you followed for developing the clustering solution for image data. Be sure to include the following in the report: -- Name registered on Miner website. -- Ranks & V scores for your submissions for HW4-Iris and HW4-Image (at the time of writing the report). You will be graded on the ranking for both datasets, but the Image data will most likely have more weight. -- Implement your choice of internal evaluation metric and plot this metric on y-axis with value of K increasing from 2 to 20 in steps of 2 for the data. -- Description of your approach (Pseudocode, how you choose or deal with the initial centers, how many runs etc) -- Tables and/or graphs that report your results. -- Any feature selection/reduction you used in this study. -- To ensure correctness, also submit results of evaluation on the standard Iris dataset provided as part of HW4-iris.