CAP 5768 MapReduce Project

1. Please submit the project report as well as upload the source code
2. Please write the names and pather IDs of the team members at the first page of the submission and also clearly specify the contributions of each member
3. For questions on the project, please contact Ms. Qing Wang (Email: wangqingemail@gmail.com and office ECS 251)

1 Optical Character Recognition

This is a programming exercise. You will implement a kNN classifier using programming model MapReduce for recognizing handwrittern numerals. The data used to train/evaluate your classifier is provided, but you are reponsible for implementing the algorithm. We suggest you to use Apache Hadoop, which you need to program using Java.

Two data files will be posted on the Moodle: zip.train and zip.test. As the names suggest, the firest should be used to train your classifier, the second to test its performance. Each contains 16×16-pixel grayscale images of the digits 0-9. There is one digit per row: the first column is the number itself, the next 256 the corresponding grayscale values.

For a given test point, the k training points with the smallest Euclidean distance to it ($\|x_{test} - x_{train}\|$) should be used as its nearest neighbors.

A) **On a single machine**, using only the first 1000 training points, list your classifier's error (fraction of mistakes) over the test set for all odd values of k from 1 to 25 in table. For what value(s) of k is the best performance attained? (10 points)

B) **On a single machine**, but now using the entire training set, list your classifier's error (fraction of mistakes) over the test set for all odd values of k from 1 to 25 in table. What value of k is the optimal? (10 points)

C) Design your own **MapReduce** including the configuration information (e.g. how many mappers nodes and reducers nodes). Draw the workflow of your MapReduce. (10 points)

D) The pseudocodes of the mapper and reducer function. (10 points)

E) **On your cluster**, using the entire training set, list your classifier's error

(fraction of mistakes) over the test set for all odd values of k from 1 to 25 in table. What value of k is the optimal? (10 points)

F) **Performance Comparison**. (50 points)

1) Compare the performance (e.g. time cost) between traditional kNN algorithm and your Map-Reduced kNN algorithm. Using the first 1000, 3000, 5000, and 7000 training points respectively, list your classifier's error (fraction of mistakes) and time cost over the test set for all odd values of k from 1 to 25 of both traditional kNN algorithm and Map-Reduced kNN algorithm in table and highlight the optimal k for each training set. (30 points)

2) Using the entire training set, for all odd values of k from 1 to 25, calculate the time cost of two algorithms respectively over the test set and draw the line chart. The x axis represents the k which is the odd value from 1 to 25, while y axis indicates the time cost (secs). (10 points)

3) Compared with the conventional kNN algorithm, list the advantages and disadvantages of kNN algorithm using MapReduce model. (10 points)