**Big Data Analytics – CS7070**

**Spring 2019**

**Homework Assignment #1**

**Due Date: March 12th, 2019, 9PM**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | A1 | A2 | A3 | Class |
| T1 | 0 | 0 | 0 | 0 |
| T2 | 0 | 0 | 1 | 1 |
| T3 | 0 | 1 | 0 | 1 |
| T4 | 0 | 1 | 1 | 0 |
| T5 | 1 | 0 | 0 | 1 |
| T6 | 1 | 1 | 0 | 0 |
| T7 | 1 | 1 | 1 | 1 |

Consider the data shown in the table. This data is to be used to construct a decision tree. We discussed in class today the outline of a MapReduce algorithm that will build a decision tree from a very large dataset stored in in HDFS across multiple nodes. We want to use the ID3 algorithm for decision tree induction that uses information gain to select the best attribute at test node of a decision tree. Assume that there is a controller program that wants to build the decision tree by launching various MapReduce jobs, and using and saving requisite results after each MapReduce iteration. In this context answer the following questions.

1. The controller launches a MapReduce iteration to compute the basic entropy of this database.
   1. Describe the structure of key-value pairs to be generated by the Mapper.
   2. Describe the computation performed by the Reducer.
   3. Describe the information that will be computed and saved by the controller module. How will the reducer output be used to do this computation.
   4. Show all the key-value pairs generated by the Mapper for the shown dataset.
   5. Show the results produced by the Reducer from the Mapper’s output.
2. The controller launches a MapReduce iteration to determine the best test attribute, from among the three attributes of the dataset. We want to achieve this with only one iteration of MapReduce.
   1. Describe the structure of key-value pairs to be generated by the Mapper.
   2. Describe the computation performed by the Reducer.
   3. Describe the information that will be computed and saved by the controller module. How will the reducer output be used to do this computation.
   4. Show all the key-value pairs generated by the Mapper for the shown dataset.
   5. Show the results produced by the Reducer from the Mapper’s output.