CS8803 BDS / CS4365

Homework Assignment 2

(Programming Category)

Student Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Student Session: cs8803 or CS4365 (circle one)

You are given three types of programming problems in the second homework assignment. You only need to choose one of them as your first homework. Feel free to choose any of your favorite programming language Java, C, Perl, Python. Also this assignment help you understand and use R/Mahout/…

Post Date: Tuesday of Week 6 (Sept. 22)

Due Date: midnight on Monday of Week 8 (Oct. 5)

**Problem 1: Hand-on Experiences with k-means clustering**

**Problem 1.1 Understanding k-means clustering**

1. Install Mahout on your laptop, and run the k-means algorithm (http://mahout.apache.org/users/clustering/k-means-clustering.html) or k-means commandline (<http://mahout.apache.org/users/clustering/k-means-commandline.htmlon>);
2. Select two datasets from the UCI repository (<http://archive.ics.uci.edu/ml/datasets.html>) or use datasets of your own, and
3. Report the runtime for two different datasets and the quality of clustering. Some reference on clustering evaluation can be found at <http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html>.
4. You may use excel file to generate your runtime statistics plot or organize the performance measurement data in a tabular format.
5. You are encouraged to learn by setting different k and varying the initial points and report the quality and runtime performance of Mahout K-means.

Deliverable.

1. Source code (if available)
2. screen shots of your execution process.
3. Runtime statistics in excel plots or tabular format.

**Problem 1.2: Hand-on Experimentation with k-Means Clustering**

You are asked to select and evaluate clustering algorithms using Mahout, R or Weka (http://www.cs.waikato.ac.nz/ml/weka/).  You are also welcome to implement your own k-means clustering algorithm and evaluate it.

1. Select two datasets from the UCI repository (<http://archive.ics.uci.edu/ml/datasets.html>) or use datasets of your own.
2. Determine how you will measure the quality of the clusters produced. Some reference on clustering evaluation can be found at <http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html>.
3. Select two algorithms for each dataset (e.g., K-means, canpopy or your own implementation) and compare their results using your quality metrics. Also you are encouraged to learn the performance impact of setting different k and varying the initial points and reporting the quality and runtime performance of Mahout K-means

4. Write a brief report to:

* Describe the datasets and your quality metrics.
* Describe your experiment setup such as how you preprocessed the data (if any), how you chose the parameters for the selected algorithms (if any), and why.
* Present the experiment results.  They should not be a simple copy-and-paste from Mahour/R/Weka output, but rather presented in a tabular or chart format for easy comparison.
* Discuss the insights and conclusions from your experiments.  For example, do different clustering methods make a difference in terms of quality or performance for the particular datasets you selected?  And why?  How data preprocessing might help?

5. Deliverable.

* Source code (if available)
* screen shots of your execution process.
* Runtime statistics in excel plots or tabular format.
* Report.

**Problem 2. Hand-on Experimentation with Classification**

Your task for this assignment is to explore run and evaluate the C4.5 decision tree classifier or Naïve Bayesian classifier using Mahout (<http://mahout.apache.org/users/classification/>) or R or Weka. You can implement your own classifier using any programming language that you are familiar with to compare with those from Mahour, R or Weka.

1. The program should be executable with at least 3 parameters: the name of the training dataset file, the name of the test dataset file and the name of the output file.

2. The program should output a file that contains the class labels for all the records in the test dataset and the classification accuracy computed as the percentage of correctly classified records in the test dataset.

3. Evaluate your implementation using the provided mushroom dataset [(mushroom.training](http://www.mathcs.emory.edu/~lxiong/cs570_s11/share/assignments/assignment3/datasets/mushroom.training), [mushroom.test](http://www.mathcs.emory.edu/~lxiong/cs570_s11/share/assignments/assignment3/datasets/mushroom.test)).  You will use the training set to build your classifier and the test set to evaluate its accuracy.  The provided two datasets were created using the [original mushroom dataset](http://archive.ics.uci.edu/ml/datasets/Mushroom) from UCI repository but with one attribute with missing values removed. The training dataset contains 7423 records and the test dataset 701 records.  The first attribute is the class of each record and the rest 21 attributes are categorical attributes.

4. Write a brief report that include the following:

* Present and discuss the results of your experiments or your implementation on the provided dataset and other datasets (if any).
* Discuss, if any, the experiences and lessons you have learned from the implementation and experimentation.

5. Deliverable.

* One tar or zip file that contains your source files, the executable, a readme file explaining how to compile/run your program, the output file for the test dataset screen shots of your execution process.
* Runtime statistics in excel plots or tabular format.
* Report in pdf/word/ppt.

**Problem 3. Hand on Experimentation with Collaborative Filtering**

Your task for this assignment is to compare two alternative CF algorithms such as user-based collaborative filtering algorithm and item based CF algorithm provided by Mahout (http://mahout.apache.org/users/recommender/) on two datasets of your choice. Feel free to introduce optimization.

 Write a brief report that include the following:

* Present and discuss the results of your experiments or your optimization on the datasets.
* Discuss the experiences and lessons you have learned from the implementation and experimentation.

Deliverable.

* One tar or zip file that contains your source files, the executable, a readme file explaining how to compile/run your program, the output file for the test dataset screen shots of your execution process.
* Runtime statistics in excel plots or tabular format.
* Report in pdf/word/ppt.