Programming Assignment 02
Deadline : 11th September 2016, 11:59 PM

Guidelines for submitting the assignments :

- **Make separate scripts for each question and name them accordingly.**
- **Team Size : 3**
- **Languages allowed : Java, R**
- **Attach the screenshots of the output of each question.**
- **Explicitly mention the inbuilt functions/Libraries used.**
- **The final compressed file should have the following naming convention: FirstName_RollNo.zip.**
- **Plagiarism policy is applicable.**

Q1: Implement **decision tree algorithm ID3** for **multi split** on the following dataset using the following metrics:

Dataset Link:
https://forge.scilab.org/index.php/p/rdataset/source/tree/master/csv/datasets/Titanic.csv

A. Gini Index.
B. Information Gain.
C. Perform Chi-square pruning for both above mentioned metrics and report your results.

Note:

- Show the decision trees formed at each stage of the algorithm for all parts.
- Report the validation and test error at the end of the algorithm with different splits of dataset (66.7%-33.3%) on each iteration averaged over 20 iterations.

Reading Material:

- *Chi-square pruning:* http://select.cs.cmu.edu/class/10701-F09/recitations/recitation4_decision_tree.pdf
    1. Build complete tree.
    2. Consider each 'leaf' decision and perform the chi-square test (label vs. split variable).
- *Gini Index:* https://en.wikipedia.org/wiki/Decision_tree_learning
- *Id3 pseudocode:* https://en.wikipedia.org/wiki/ID3_algorithm

  1.Calculate the entropy of every attribute using the data set S.

  2. Split the set S into subsets using the attribute for which entropy is minimum (or, equivalently, information gain is maximum).

  3. Make a decision tree node containing that attribute.

  4. Recurse on subsets using remaining attributes.

Q2: Implement **K- means clustering** algorithm on the following dataset:
Dataset Link:
http://en.osdn.jp/projects/sfnet_irisdss/downloads/IRIS.csv/

A. For K=4. Also show the initial cluster, intermediate cluster(any 2) and the final cluster formed.
B. Find the optimal value of K (ranging from 2 to 12) for which the error is minimized and plot the graph showing the error curve obtained on different values of K.
C. Compare the output of EM algorithm(Expectation Maximization) with K-means algorithm at K=4.
D. Mark which of the following is correct/incorrect with reasons:
    1. EM and K-Means perform hard assignment of data points to a cluster. (By hard assignment it means that either a point belongs to a cluster or it does not).
    2. Both algorithms produce spherical clusters.
    3. Both algorithms are based on distance as metric to assign a point to a cluster.
    4. EM uses Poisson's distribution to define clusters.

Note:

- Error = Root Mean Squared error.

Reading Material:

- *Finding optimal value of K:* https://www.quora.com/How-can-we-choose-a-good-K-for-K-means-clustering (follow the elbow method).