COL 341: Assignment 4

Notes:

- You are advised to use vector operations (wherever possible) for best performance.
- Include a report of maximum 5 pages which should be a brief description explaining what you did. Include any observations and/or plots required by the question in the report.
- You should use Python for all your programming solutions.
- Your assignments will be auto-graded, make sure you test your programs before submitting. We will use your code to train the model on training data and predict on test set.
- Input/output format, submission format and other details are included. Your programs should be modular enough to accept specified parameters.
- You should submit work of your own. You should cite the source, if you choose to use any external resource. You will be awarded F grade or DISCO in case of plagiarism.
- You can use total of 7 buffer days across all assignments.
- Data is available at this link

1. **Decision Tree (50 points, Release date: Oct.03, 2019, Due date: Oct.11, 2019)**
   In this problem, you will work with the Adult Dataset available on the UCI repository. Read about the dataset in detail from the link given above. For the purpose of this assignment, you have been provided with a subset of this dataset. Specifically, you have been provided with separate training, validation and testing sets. The first row in each file specifies the type of each attribute. The last entry in each row denotes the class label. You can encode the discrete non-numeric attributes to some numeric value. You have to implement the decision tree algorithm for predicting whether a person is rich (1) or not (0) based on various personal attributes.

   Construct a decision tree from first principle for the above prediction problem. You should not use any library available for decision tree. Experiment with different splitting criterions such as Gain Ratio, Information Gain etc. alongside different processing techniques for numerical attributes such as median splitting or threshold based splitting(refer to section 3.7.2 of Machine Learning by Tom Mitchell). In case of a tie, choose the attribute which appears first in the ordering as given in the training data. Plot the train, validation and test set accuracies against the number of nodes in the tree as you grow the tree. On X-axis you should plot the number of nodes in the tree and Y-axis should represent the accuracy.

   One of the ways to reduce over-fitting in decision trees is to grow the tree fully and then use post-pruning based on a validation set. In post-pruning, we greedily prune the nodes of the tree (and sub-tree below them) by iteratively picking a node to prune so that resultant tree gives maximum increase in accuracy on the validation set. In other words, among all the nodes in the tree, we prune the node such that pruning it(and sub-tree below it) results in maximum increase in accuracy over the validation set. This is repeated until any further pruning leads to decrease in accuracy over the validation set. Post prune the tree obtained in step (a) above using the validation set. Again plot the training, validation and test set accuracy against the number of nodes in the tree as you successively prune the tree.

Evaluation:

- Marking will be done in two parts: code (80%) and plot(20%).
- For code: you can get 0 (error), half (code runs fine but predictions are incorrect within some predefined threshold) and full (works as expected).
- For plot: you can get 0 (wrong plot),half (plot is partially correct) and full (plot is as expected).

Submission Instructions:
Decision Tree

Submit your code in 1 executable python files called dt.py

python dt.py trainfile.csv validfile.csv testfile.csv validpred.txt testpred.txt
Here you have to write the predictions (1 per line) and create 2 line aligned files validpred.txt and testpred.txt for validation and test files respectively.

Part (b)
Here you have to submit a pdf file with plots constructed alongside complete details of splitting criteria and numerical attribute processing used. There will be no demos or presentations for this part.

Coding Guidelines:

- Grading will be based on a relative ranks on a weighted score of testing and validation accuracy.
- Unfair means in this assignment include use of external libraries like sklearn or use of any other classifier other than decision tree. Furthermore the purpose of including validation accuracy while marking is to give post pruning more credit. Needless to say as validation dataset is going to be provided with labels there is chance of mischief where you can use provided labels to increase accuracy. Be assured your submissions will be carefully checked and any such attempt will be **severely punished** with appropriate disciplinary measures.
- Their will be another required file coined MoodleID where you simply have to fill in your moodle ID eg. me2110786. This is necessary for construction of a leaderboard wherein ranks based on best scores till a point can be seen.