SMAI (CSE 471)
Spring-2019
Assignment-1 (50 points)
Posted on: 11/1/2019
Due on: 20/1/2019, 11:55 PM

- Questions can involve a mix of writing code/scripts and answering questions or analyzing results.
- Code: Your scripts should be of the form q-x-y.py where x is the main question, y is the sub-question. For e.g., q-1-2.py is Python script for sub-question 2 within question 1. If you are submitting Jupyter notebook file (.ipynb), make sure that it is properly formatted and documented with question part numbers (Part-1, Part-2 etc.).
- Ensure that submitted assignment is your original work. Please do not copy any part from any source including your friends, seniors and/or the internet. If any such attempt is caught then serious action will be taken.
- Use suitable train-validation split for your training and validation (20% of data).
- Numpy, pandas/csvReader(for data processing) are allowed. No inbuilt library allowed for decision tree. Write your own decision tree algorithm from scratch.
- Sample test file is given containing header and one sample row. Evaluation will be done based on your understanding, report and accuracy on purely unseen test data (provided at the time of assignment evaluation).
- Report your precision, recall, F1 score and accuracy on validation data in your report.
- Report should contain details of algorithm implementation, results and observations.

1 Question

1. (100 points) Design a Decision Tree classifier to predict which valuable employees will leave next. You are tasked with helping in reducing the number of senior employees leaving the company by predicting the next bunch. The fate of the company rests in your hands. Download dataset here(`http://researchweb.iiit.ac.in/~murtuza.bohra/decision_Tree.zip`).

   For each employee you are provided the following attributes:

   1. Satisfaction Level
   2. Last evaluation
   3. Number of projects

4. Average monthly hours
5. Time spent at the company
6. Whether they have had a work accident
7. Whether they have had a promotion in the last 5 years
8. Departments
9. Salary
10. Whether the employee has left the company

Note: Class Labels(column name left) : 1 for employee left the company, 0 for not. All parts of the question are compulsory, read them all before you design your algorithm.

1. **Part-1**: (30 points) Train decision tree only on categorical data. Report precision, recall, f1 score and accuracy.

2. **Part-2**: (30 points) Train the decision tree with categorical and numerical features. Report precision, recall, f1 score and accuracy.

3. **Part-3** (10 points) Contrast the effectiveness of Misclassification rate, Gini, Entropy as impurity measures in terms of precision, recall and accuracy.

4. **Part-4**: (10 points) Visualise training data on a 2-dimensional plot taking one feature (attribute) on one axis and other feature on another axis. Take two suitable features to visualise decision tree boundary (Hint: use scatter plot with different colors for each label).

5. **Part-5**: (10 points) Plot a graph of training and validation error with respect to depth of your decision tree. Also plot the training and validation error with respect to number of nodes in the decision tree.
6. **Part-6**: (10 points) Explain how decision tree is suitable handle missing values(few attributes missing in test samples) in data.