CSE343/ECE343: Machine Learning
Assignment-2 Decision Trees, Random Forests and Perceptron
Max Marks: 25 (Programming: 15, Theory: 10) **Due Date**: 24/09/2023, 11:59 PM

Instructions

- Keep collaborations at high-level discussions. Copying/Plagiarism will be dealt with strictly.
- Late submission penalty: As per course policy.
- Your submission should be a single zip file **2020xxx_HW1.zip** (Where *2020xxx* is your roll number). Include **all the files (code and report with theory questions)** arranged with proper names. A single **.pdf report** explaining your codes with results, relevant graphs, visualization and solution to theory questions should be there. The structure of submission should follow:

  2020xxx_HW2
  |– code_rollno.py/.ipynb
  |– report_rollno.pdf
  |– (All other files for submission)
- Anything not in the report will **not** be graded.
- Remember to **turn in** after uploading on Google Classroom. No excuses or issues would be taken regarding this after the deadline.
- Start the assignment early. Resolve all your doubts from TAs in their office hours at least **two days before the deadline.**
- Your code should be neat and well-commented.
- **You have to do either Section B or C.**
- **Section A is mandatory.**

---

1. (4 points) **Section A (Theoretical)**

   (a) (1 marks) Discuss the trade-off between correlation and diversity in Random Forests. Why is it important for the trees to be correlated up to a certain extent while maintaining diversity?

   (b) (1 marks) When might the "curse of dimensionality" become an issue in Naive Bayes? What strategies can be employed to mitigate this problem in practice?

   (c) (1 marks) What kind of problems will the Naive Bayes classifier face if it encounters some value of attributes which was not present in the training dataset? Do you think this will affect the inference results? If yes, list some of the approaches to mitigate this problem, else give your reason why you think so that the results won't be affected. Explain with an example.

(d) (1 marks) Do you think that while splitting a decision tree node using Information Gain might be biased if some attributes have more cardinality than others? If yes, mention some other criterion for attribute selection with proper explanation, else give reason why decision trees are unaffected by attributes having different cardinality of attributes. Explain with an example.

2. (5 points) Rahul's decision-making process for playing a sport is based on the weather and certain conditions. Given that he prefers outdoor activities, if there is no rain, he will play outdoor. For outdoor activities, if more than 7 of his friends are playing, he will opt for football; otherwise, he will choose badminton. On the other hand, if it's indoor due to rain, he will play table tennis (TT) only if he can borrow TT rackets from a friend; otherwise, he will play pool.

   (a) (1 marks) Draw decision tree for the given scenario and write the all possible outcomes and their conditional probabilities expression.

   (b) (1 mark) Rahul decides to rely on a weather prediction app that claims to accurately forecast 'Rainy' and 'Clear' days. On any given day, the app predicts 'Rainy' with a probability of 0.3 (30) and it predicts 'Clear' with a probability of 0.7 (70percent)The app's accuracy for predicting 'Rainy' days is 80 percent, and its accuracy for predicting 'Clear' days is 90percent. What is the probability that it's going to rain on a day, given that the app predicts 'Rainy'?"

   (c) (1.5 marks) Rahul's friend Ram has asked Rahul to come along with him for gym (where he can go irrespective of weather conditions). But Rahul's likelihood of going to the gym depends on his mood. If he is in good mood then there is a 80 percent chance that he will go to the gym and 20 percent chance that that he will stick to his earlier plan whereas, if he is in bad mood there is a 40percent chance that he will go to the gym and 60 percent chance that he will stick to his earlier plan. At gym he will either do cardiological exercise or weight training, both of which are equally likely. Draw the new decision tree and write down all possible outcomes and their conditional probabilities (in expression form).

   (d) (1.5 marks) The probability of Rahul having good mood is 0.6 and bad mood is 0.4, and his mood is determined by factor F (amount of sleep he had last night). Assume that Rahul had 7 hours of sleep on the night before he is making this decision (Given $P(F = 7 \mid \text{Good mood}) = 0.7$ and $P(F = 7 \mid \text{Bad mood}) = 0.45$). Find the probability of all the possible outcomes and state the most likely outcome.

3. (15 points) **Section B (Library Implementation)**

   Decision Tree and Random Forests

   Perform classification task on the heart disease dataset using only the relevant attributes mentioned on the repository website.

   Dataset: **Heart Disease - UCI Machine Learning Repository**

   (a) (2 marks) Preprocess the dataset if required and perform Exploratory Data Analysis

(b) (1 marks) Split the dataset into train and test sets in the ratio 80:20

(c) (3 marks) Train decision trees using 'entropy' and 'gini impurity' as the splitting criterion and report the best criterion for attribute selection based on the accuracy scores.

(d) (4 marks) Now taking the best criterion for attribute selection in part c, perform hyperparameter search for the parameters min_samples_split and max_features using Grid Search. Select the best combination of the hyperparameters using the test data scores.

(e) (5 marks) Finally train a random forest classifier for the same dataset. Perform Grid Search for the parameters n_estimators, max_depth and min_samples_split. Report the best combination of hyperparameters and present the classification report on the test data.

4. (15 points) **Section C (Algorithm implementation using packages)**

1. Implement a Decision Tree from Scratch for Classification. Create a Decision Tree classifier from scratch using the NumPy and Pandas libraries. Design a class called MyDecisionTree (specifically for solving classification problems.)

(a) (10 marks)Include the following functions in MyDecisionTree class:

(a) cost_function(): Develop a cost function that could be either the Gini index or Information gain. This function should compute the impurity of a node or the gain achieved by a potential split.
Gini index: 1- sum(proportion)2
proportion = number of values/count of rows

(b) make_split(): Define the basic mechanism of splitting a node in the Decision Tree such that it selects the best feature and value to split on.

(c) max_depth(): Define the maximum depth of the tree.

(d) Pruning (optional): define a function to remove branches that doesn't contribute much for improving accuracy.

(e) predict()

(f) score(): Create a function to evaluate the DT (performance metric)
Provide proper documentation (well-commented codes)

(b) (5 marks) Additionally, you may demonstrate the effectiveness of your MyDecisionTree class by training and evaluating it on the given **dataset**.