Instructions

- This homework assignment is worth 100 points.
- Please submit a **.ipynb** file to Blackboard.
- **Please strive for clarity and organization.**
- **Due Date: November 3, 2023 by 11:59 pm.**

Exercise 1

A data scientist is running an `AdaBoost` classifier on a dataset with 100 observations. Answer the following:

(a) (3 points) What is the weight initial weight of observation 72th in the training dataset? Be specific.

(b) (3 points) The 72th observation in the training dataset is misclassified by the first weak learner chosen by the data scientist. Is the new weight of the 72th observation in the training dataset (i.e., the weight assigned to the 72th observation after choosing the first weak classifier) larger or smaller than the weight assigned to that observation initially? Be specific.

Exercise 2

(4 points) Explain why `AbaBoost` is an ensemble learning algorithm? Be specific.

Exercise 3

(10 points) Suppose you are running `AdaBoost` with 4 training examples. At the start of the current iteration, the four examples have the weights shown in the following table. Another column says if the weak classifier got them correct or incorrect. Determine the new weights for these four examples, and fill in the corresponding entries in the table. Show all your work.

Observation	Old Weight	Correct?	New Weight
1	0.16	Correct	
2	0.64	Correct	
3	0.08	Incorrect	
4	0.12	Incorrect	

Exercise 4

(4 points) If your `AdaBoost` ensemble under-fits the training dataset, what would you do to fix that? That is, which hyper-parameters should you tweak?

Exercise 5

(4 points) For binary classification, which of the following statements are **TRUE** of AdaBoost with decision trees as learners?

(a) It usually has lower bias than a single decision tree.

(b) It is popular because it usually works well even before any hyper-parameter tuning.

(c) It assigns higher weights to observations that have been misclassified.

(d) It can train multiple decision trees in parallel.

(e) All of the above.

(f) None of the above.

Exercise 6

(4 points) Which of the following is/are **TRUE** about gradient boosting trees?

(a) In gradient boosting trees, the decision trees are independent of each other.

(b) In gradient boosting trees, the decision trees are dependent of each other.

(c) It is a method for improving the performance by aggregating the results of several decision trees.

(d) (a) and (b)

(e) (a) and (c)

(f) (b) and (c)

(g) None of the above.

Exercise 7

(6 points) In this course have covered two boosting frameworks. What is the main difference between `AdaBoost` and Gradient Boosting? Be specific.

Exercise 8

Consider the `framingham.csv` data file. The dataset is available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). The dataset provides the patients? information. It includes over 4,000 records and 15 attributes. Each attribute is a potential risk factor. There are both demographic, behavioral and medical risk factors.

- Demographic:
  - Sex: male or female (Nominal)
  - Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
- Behavioral
  - Current Smoker: whether or not the patient is a current smoker (Nominal)
  - Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)
- Medical (history)
  - BP Meds: whether or not the patient was on blood pressure medication (Nominal)
  - Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
  - Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
  - Diabetes: whether or not the patient had diabetes (Nominal)
- Medical (current)
  - Tot Chol: total cholesterol level (Continuous)
  - Sys BP: systolic blood pressure (Continuous)
  - Dia BP: diastolic blood pressure (Continuous)
  - BMI: Body Mass Index (Continuous)
  - Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
  - Glucose: glucose level (Continuous)
- Predict variable (desired target)
  - 10 year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

In Python, answer the following:

(a) (5 points) Using the `pandas` library, read the csv data file and create a data-frame called `heart`. Remove the observations with missing values.

(b) (50 points) Based on the different interactions that we have so far with this data file, it seems that `age`, `totChol`, `sysBP`, `BMI`, `heartRate`, and `glucose` are the most important predictor variables when it comes to predict the likelihood of `TenYearCHD`. Then do the following:

(i) Using `TenYearCHD` as the target variable and the other variables as the input variables, split the data into two data-frames (taking into account the proportion of 0s and 1s): `train` (80%) and `test` (20%).

(ii) Using the `train` data-frame, build a `RandomForestClassifier` model (with `n_estimators = 500` and `max_depth = 5`)) in which: `age`, `totChol`, `sysBP`,`BMI`, `heartRate`, and `glucose` are the predictor variables, and `TenYearCHD` is the target variable. Using this model, predict the risk of coronary disease of the patients in the `test` data-frame. Using 10% as cutoff value, report the accuracy and recall.

(iii) Using the `train` data-frame, build a `ExtraTreesClassifier` model (with `n_estimators = 500` and `max_depth = 5`)) in which: `age`, `totChol`, `sysBP`,`BMI`, `heartRate`, and `glucose` are the predictor variables, and `TenYearCHD` is the target variable. Using this model, predict the risk of coronary disease of the patients in the `test` data-frame. Using 10% as cutoff value, report the accuracy and recall.

(iv) Using the `train` data-frame, build an `AdaBoostClassifier` model (with `n_estimators = 500` and `max_depth = 3`, and `learning_rate = 0.01`) in which: `age`, `totChol`, `sysBP`,`BMI`, `heartRate`, and `glucose` are the predictor variables, and `TenYearCHD` is the target variable. Using this model, predict the risk of coronary disease of the patients in the `test` data-frame. Using 10% as cutoff value, report the accuracy and recall.

(iv) Using the `train` data-frame, build an gradient boosting model (with `n_estimators = 500` and `max_depth = 3`, and `learning_rate = 0.01`) in which: `age`, `totChol`, `sysBP`,`BMI`, `heartRate`, and `glucose` are the predictor variables, and `TenYearCHD` is the target variable. Using this model, predict the risk of coronary disease of the patients in the `test` data-frame. Using 10% as cutoff value, report the accuracy and recall.

Repeat (i)-(iv) 100 times and compute the average accuracy and average recall for each of the model. What model would you use to predict `TenYearCHD` based on the accuracy and recall?

(c) (7 points) Assuming that the ideal model needs to have a minimum accuracy and recall equal to 80%. Does the best model from part (b) meet this requirement? If not, provide recommendations how you tweak the model to reach the minimum requirements.