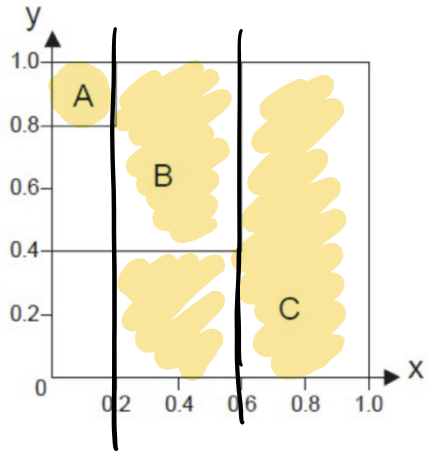
DS 310 Machine Learning
Fall 2021 / Amulya Yadav
HW #2: 50 points

Please submit one files on Canvas:

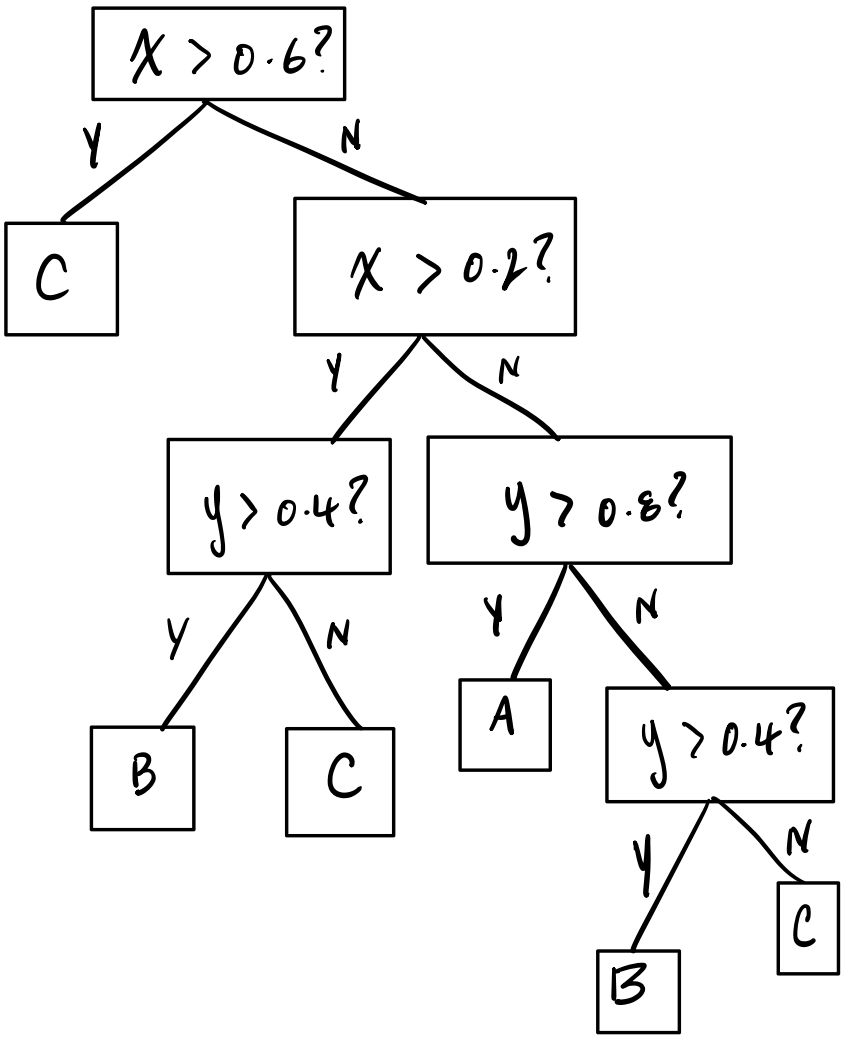
- *your answer to this homework;*

Please reach out to TA Hangzhi Guo (email: hangz@psu.edu) if there is any question about the assignment.

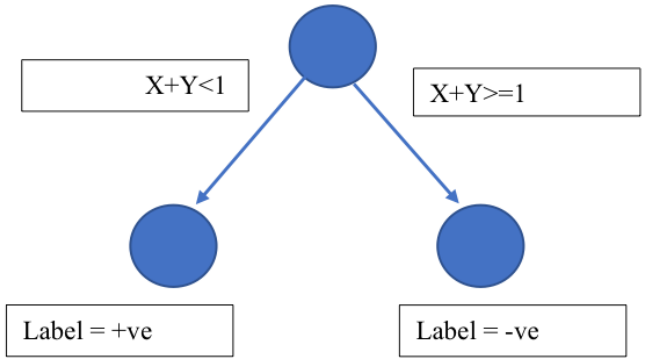
1. (5 points) Consider 2-dimensional data set shown on the right, where A, B, and C are the class labels for the respective regions. Draw the full decision tree that perfectly classifies each of the data sets given. You do not have to consider the impurity measure (e.g., entropy, Gini index) used by the decision tree algorithm. Ignore pre-pruning and post-pruning. Assume there are no noise and missing attribute values.



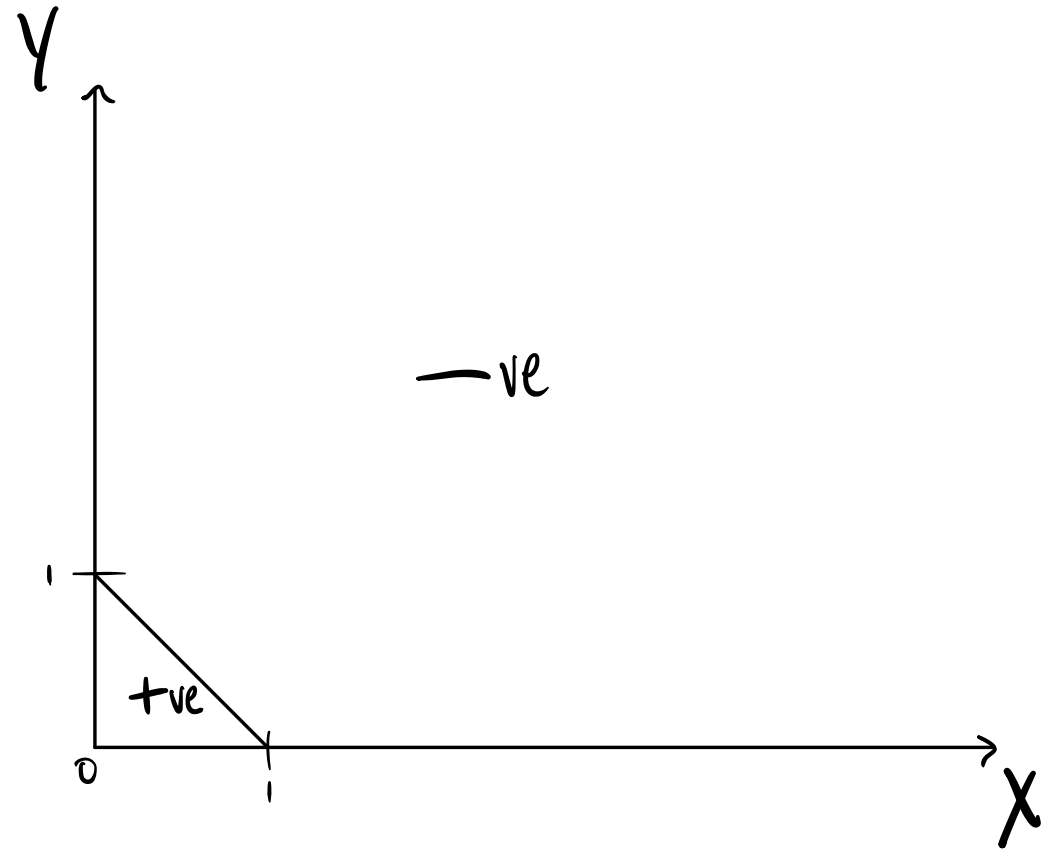
1.



2. (5 points) Consider a dataset which has two features: X and Y in your dataset. Also, unlike the example shown above, our dataset has only two kinds of labels: +ve and -ve. I construct a decision tree which looks like the one shown below. Plot on an X-Y plane (as the one shown above) the decision boundaries corresponding to this decision tree. Note that this kind of a decision tree (which involve splits on multiple features as part of the same split) is called an oblique decision tree.



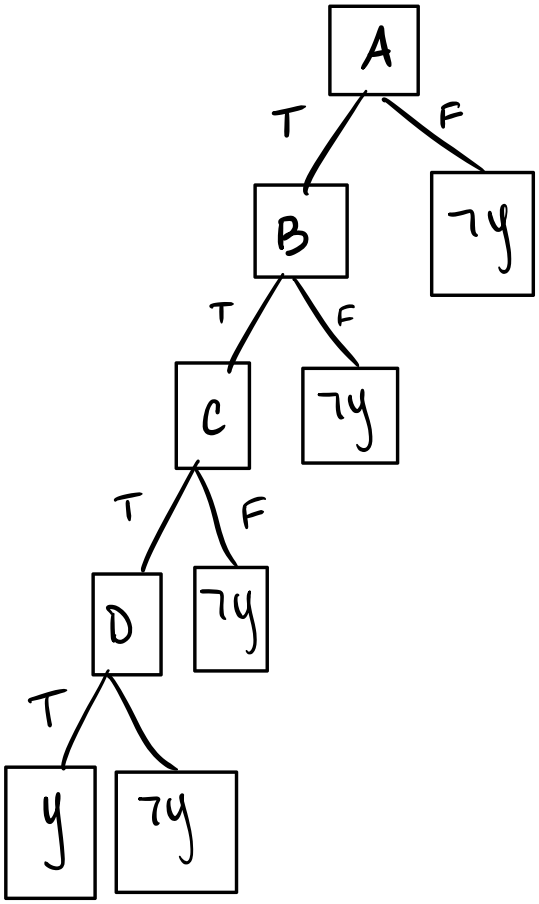
2.



3. (5 points) Consider a data set that contains 4 Boolean attributes A, B, C, and D, and a Boolean class y. For the Boolean expression below (between the class y and the rest of the attributes), it is possible to construct a smaller decision tree that perfectly classifies the data without generating the complete tree (i.e., the number of leave nodes is less than 16). Then, draw such a tree.

$$y = A \wedge B \wedge C \wedge D$$

3.



4. (5 points) Consider the problem of predicting how well a baseball player will bat against a particular pitcher. The training set contains ten positive and ten negative examples. Assume there are two candidate attributes for splitting the data—i.e., ID (which is unique for every player) and Handedness (left or right). Among the left-handed players, nine of them are from the positive class and one from the negative class. On the other hand, among the right-handed players, only one of them is from the positive class, while the remaining nine are from the negative class. Compute the information gain if we use Handedness as the splitting attribute.

4. $\text{original} = -\left(\frac{10}{20}\log_2\left(\frac{10}{20}\right) + \frac{10}{40}\log_2\left(\frac{10}{20}\right)\right) = 1$

$\text{left} = -\left(\frac{9}{10}\log_2\left(\frac{9}{10}\right) + \frac{1}{10}\log_2\left(\frac{1}{10}\right)\right) = 0.4689$

$\text{right} = -\left(\frac{1}{10}\log_2\left(\frac{1}{10}\right) + \frac{9}{10}\log_2\left(\frac{9}{10}\right)\right) = 0.4689$

$\text{Information Gain} = 1 - 0.4689 = \boxed{0.531}$

5. (5 points) Explain the following questions using your own words.
(a) What is an ensemble method?
(b) What is the difference between bagging and boosting in ensemble learning?

a) An ensemble method is using multiple machine learning models in the prediction of a single feature.

b) Bagging, or bootstrap aggregation, takes a dataset and makes subsets of the dataset. Then, a single ML model is trained using these different subsets. On the other hand, boosting doesn't make subsets of the dataset, but prioritizes the mistakes it makes during training and re-trains the ML model on the same dataset.

6. (5 points) Explain if the evaluation measure accuracy, $(TP + TN)/(TP + FP + FN + TN)$, is a good one or not in evaluating the classifiers to detect rare diseases that affect only a small percentage of the population at any given time.

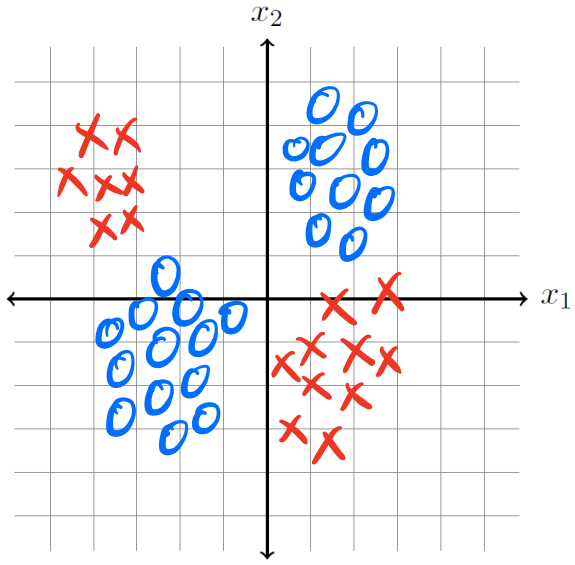
6.

The evaluation measure accuracy, $\frac{TP + TN}{TP + FP + FN + TN}$,

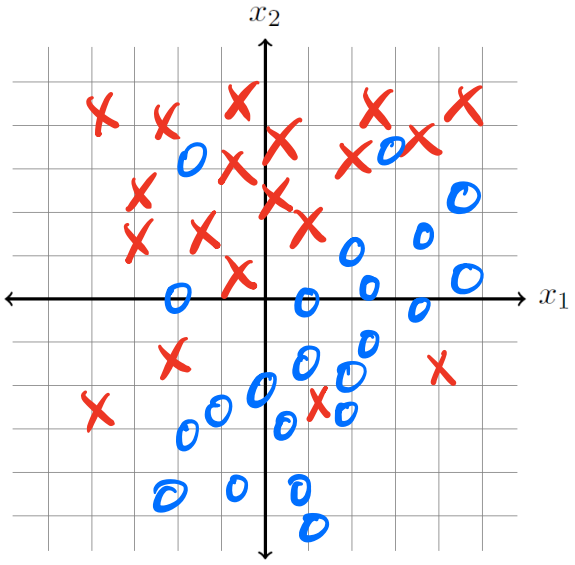
is not a good one in evaluating the classifiers to detect rare diseases that affect only a small percentage of the population at any given time. It is more important to evaluate the incorrect predictions rather than the correct ones since it could be the difference in recieving treatment or not.

7. (10 points) Consider trying to classify two classes "o" and "x" with two features x_1 and x_2.
(a) Using "o" to denote samples from one class, and "x" to denote samples from another class. Draw a scenario where the *random forests* will work well, but the *logistic regression* model will work poorly.

(a) Using "o" to denote samples from one class, and "x" to denote samples from another class. Draw a scenario where the *logistic regression* will work well, but the *random forests* model will work poorly.



(a)



(b)

8. (10 points) Consider the following training set by PSU for predicting whether there is traffic congestion in the morning on the Atherton street for a particular day. There are 100 examples in the training set, with 40% positive (i.e., congestion) and 60% negative (i.e., no-congestion) examples.

Accident	Weather	Construction	Number of positive training examples	Number of negative training examples
no	good	no	5	30
no	good	yes	10	20
yes	good	no	10	5
yes	bad	no	10	5
yes	bad	yes	5	0

no bad yes

Then, by applying the Naïve Bayes classifier, what would be predicted as the class label of the feature set: (Accident = no, weather = bad, construction = yes)?
(NOTE: Do not program Python or use Weka to solve this question. You need to show detailed calculation of Naïve Bayes classification to reach to the prediction).

A = attributes, C = congestion (positive), NC = no-congestion (negative)

$$P(A|C) = \frac{15}{40} \times \frac{15}{40} \times \frac{15}{40} = 0.0527$$

$$P(A|NC) = \frac{50}{60} \times \frac{5}{60} \times \frac{20}{60} = 0.0231$$

$$P(A|C)P(C) = 0.0527 \times 0.4 = 0.0211$$

$$P(A|NC)P(NC) = 0.0231 \times 0.6 = 0.0139$$

$$P(A|C)P(C) > P(A|NC)P(NC)$$

$\therefore$ THERE WILL BE CONGESTION