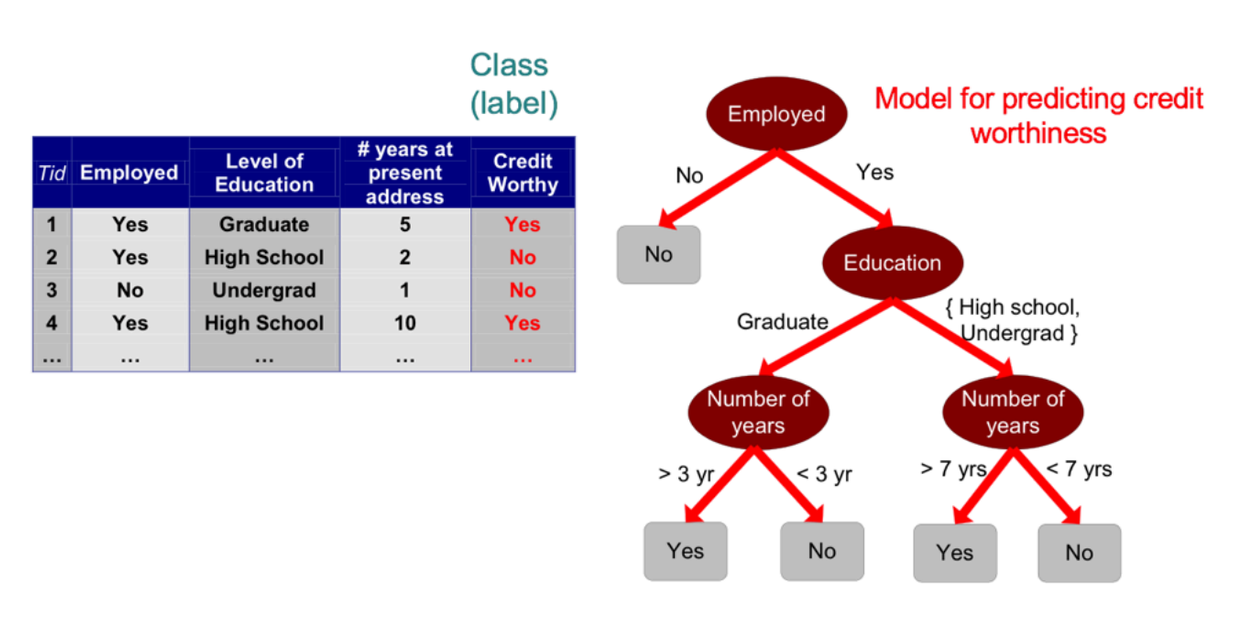
EE6435 online test 1

Question 1 — 2 pts

Given the training data in the table and the derived decision tree, **provide two records** with classification error 50% using this tree.



Record1: Employed= ______ , Level_of_education= ______ , Years_at_present_addres= ______ , Credit_Worthy= ______

Record2: Employed= ______ , Level_of_education= ______ , Years_at_present_addres= ______ , Credit_Worthy= ______

Question 2 — 5 pts

Given the integer array of size 20: <9, 3, 1, 4, 1, 10, 1, 1, 3, 5, 2, 0, 20, 50, 15, 8, 3, 0, 4, 1>, answer the following questions.

mode= ______

mean= ______

55% percentile = ______ ;

median= ______

range= ______

Question 3 — 6 pts

This is a multi-answer problem about the decision tree construction. It is possible that more than one choice is correct. **Mark all that are correct**.

- [ ] Given the same probability distribution vector for two classes, its entropy >= its Gini index
- [ ] Build an optimal decision tree is very difficult
- [ ] Using the Gini Index or Entropy as the impurity function will always lead to the same decision tree.
- [ ] Node impurity function must reach minimum for distributions (1, 0, 0), (0, 1, 0), and (0, 0, 1)
- [ ] For a training set with just two binary attributes (A and B), and a class C with label "Yes" and "No", P(C=Yes|A=1,B=1) + P(C=No|A=1,B=1)=1.0.
- [ ] In k-fold cross validation, 1/k of the training samples will be used as the validation set

Question 4 — 4 pts

The following table shows the results of a classification model. The training data has 10 records; each has three attributes. There are two classes (+ and -). Column "Ground Truth" is the ground truth. Column "Prediction" is the prediction. You need to summarize the model's performance using confusion matrix. The four boxes in the confusion matrix are named from a.1 to a.4. Fill out their values in the provided blanks.

a.1 = ______ ? a.2= ______ ? a.3= ______ ? a.4= ______ ?

Record	X1	X2	X3	Ground Truth	Prediction
1	0	2	1	+	-
2	1	1	0	-	-
3	1	2	1	-	-
4	0	2	0	+	+
5	1	1	1	-	-
6	1	0	1	-	-
7	0	2	0	-	-
8	1	1	0	-	-
9	0	1	1	-	-
10	0	0	0	-	-

	PREDICTED CLASS		
ACTUAL CLASS		Class=+	Class=-
	Class=+	a.1	a.2
	Class=-	a.3	a.4

Question 5 — 20 pts

The following figure shows the histograms of two classes (i.e. **series**) of data points (i.e. records). Each record only has **one attribute X** (X-axis) that takes integer values between 1 and 16. Y-axis shows the number of data points with the given attribute value. **Class 1 and Class 2 are represented by series 1 and series 2, respectively.** The frequency of each value at X-axis in each series is shown inside the circle. For example, when X=5, it occurs 8 times in series 1. We will consider to use three classifiers: Naive Bayes, decision tree, and kNN.

Fill out the blanks for each question.

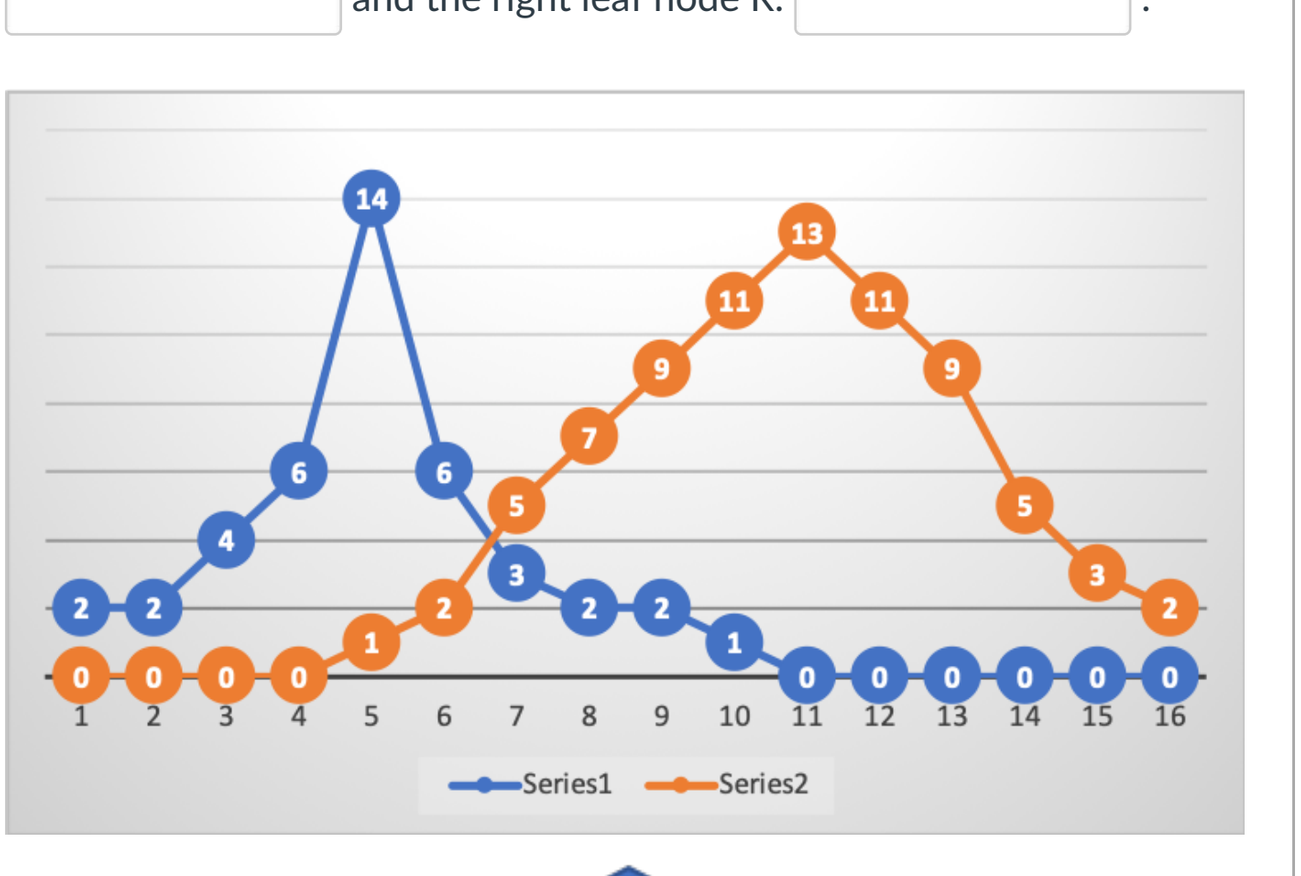
1. What is the prior probability of class 1: ______

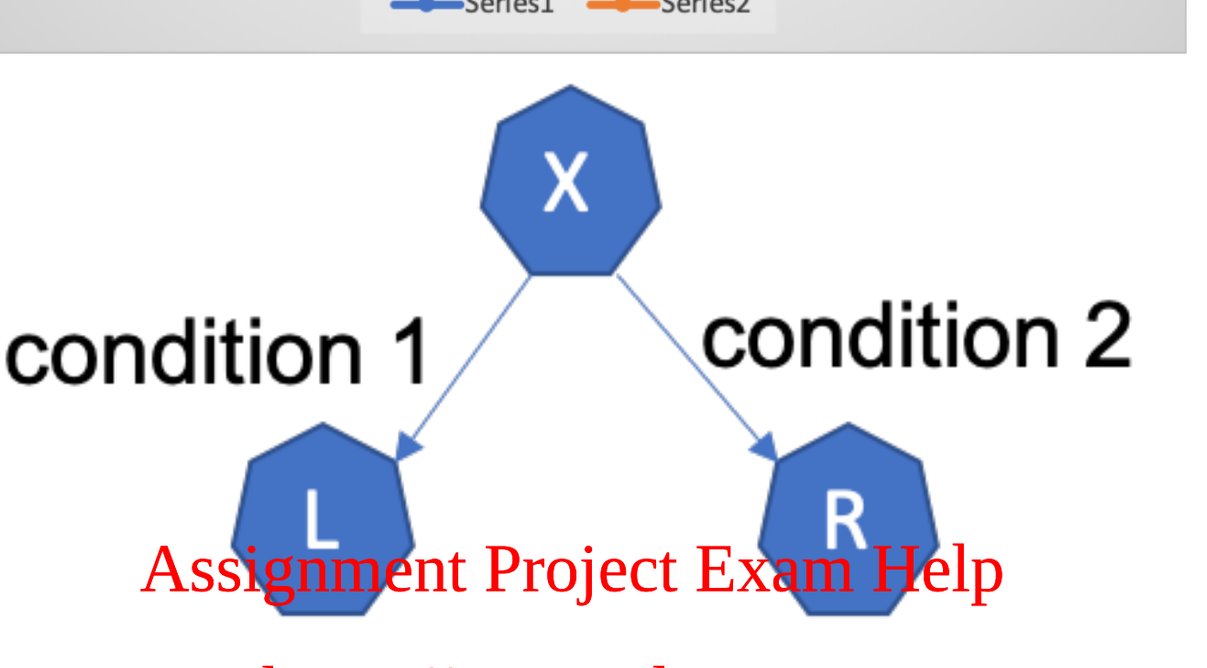
2. What is the prior probability of class 2: ______

3. What is the posterior probability P(class 1|X=7): ______ and P(class 2|X=7): ______ .

4. What is the label of input X=5 if we use kNN with k=5 ______ . Use the euclidean distance on X-axis.

5. Now we will build a decision tree with topology shown below the line plot. Attribute X has integer value from 1 to 16. Consider three values for the root node: X=6, 7, or 8. Which of them can lead to the optimal 1-level decision tree using misclassification error as the metric? Each node and edge in the tree has a label. Following the label, answer the following questions. Specify the value of X here: ______ ; specify the condition 1 here: ______ and condition 2 here: ______ ; what is the classification errors of the leaf node L: ______ and the right leaf node R: ______ .





Question 6 — 2 pts

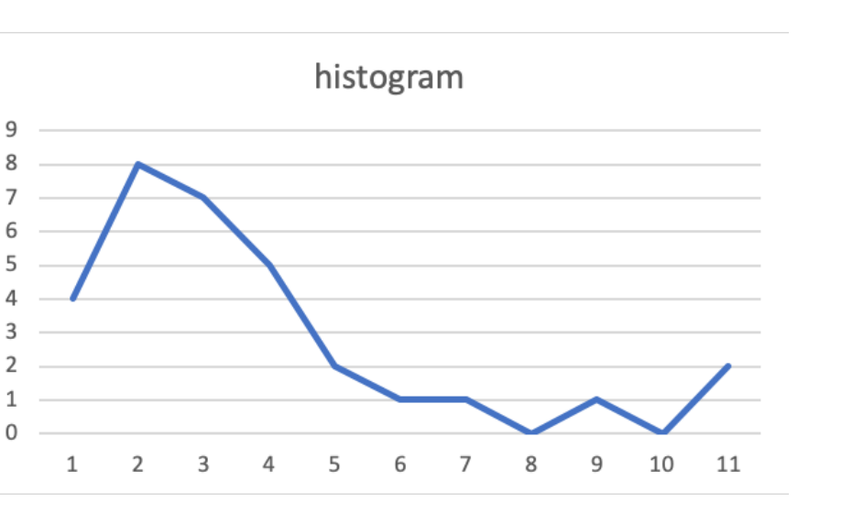
______ In the following histogram, what is the relationship between mean and median? Note that y-axis shows the occurrence times of the value on the x-axis.

A. Mean > median

B. Median > mean

C. Mean = median

D. Cannot be determined



Question 7 — 5 pts

Given the following training examples, derive the decision tree using Hunt's algorithm and the Entropy function. There are three attributes: X_1, X_2, and X_3. Y represents the class labels: + or -. This table contains **all the possible values for X_1, X_2, and X_3**. The splitting should stop when the leaf node becomes pure (has only one class).

Entropy = sum of (-p_i log p_i), where p_i is the probability of class i.

For each leaf node, mark the label. For all the other nodes, give your computation details of the Information Gain for the final tree only. For example, if your final tree uses X2 as the root attribute, you only need to show the steps of computing the information gain for using X2 as the attribute and its splitting conditions.

X1	X2	X3	Y
0	1	0	+
1	-1	0	+
1	-1	0	+
0	-1	1	-
0	1	1	-
1	1	1	-
1	0	1	+
1	-1	0	+

HTML Editor

12pt

0 words

Question 8 — 4 pts

Consider the training data set. There are three attributes A, B, and C. The class label is in column Y. Predict the class label for a test sample (A=0, B=2, C=0) using the naïve Bayes classifier. The answer can be +, -, or cannot decide. Show the intermediate steps of comparing P(Y=+|A=0, B=2, C=0) and P(Y=-|A=0, B=2, C=0).

Record ID	attribute A	attribute B	attribute C	Y
1	1	0	1	-
2	0	2	0	+
3	1	1	0	+
4	0	1	1	-
5	0	0	0	-
6	0	2	1	+
7	1	1	0	-
8	1	2	1	-
9	0	2	0	+
10	1	1	1	-

HTML Editor

12pt

0 words

Question 9 — 7 pts

Mr. X is trying to design a classification model that can automatically detect spam emails. In order to do this, he collected many spam emails and found that these emails often contain some keywords with higher frequency than expected in a regular email. He thus would like to design a classification model based on this observation. Of the following models, choose the best model to **incorporate this observation**: Naïve Bayes classifier, kNN, and SVM.

Choose the best model for this task and describe your model by answering the following questions:

1. What is the best model for this classification problem?
2. What are your training data?
3. Describe the unknown parameters in your model.
4. Describe how to derive these unknown parameters of your model
5. How to use your model to detect spam emails?

HTML Editor

12pt

0 words

Question 10 — 0 pts

If you need to make any clarifications, use this space.

HTML Editor

12pt

0 words

Saving... Submit Quiz