Homework Problem Set 2: Basic Classification

Due Thursday, Jan. 26 at 11:59 PM
Upload a pdf to Canvas

Question 1:
Consider the training set given below for predicting lung cancer in patients based on their symptoms (chronic cough and weight loss) and other lifestyle and environmental attributes (tobacco smoking and exposure to radon). Draw a two-level decision tree obtained using entropy as the impurity measure. Show your steps clearly (i.e., the computation of information gain for every candidate attribute at the first and second levels of the decision tree must be shown). Compute the training error of the decision tree.

Tobacco Smoking	Radon Exposure	Chronic Cough	Weight Loss	Lung Cancer
Yes	Yes	Yes	No	Yes
Yes	No	Yes	No	Yes
Yes	No	Yes	Yes	Yes
Yes	No	Yes	Yes	Yes
No	Yes	No	Yes	Yes
Yes	No	No	No	No
No	No	Yes	No	No
No	No	Yes	Yes	No
No	No	Yes	No	No
No	No	No	Yes	No

Question 2:
Consider a training set sampled uniformly from the two-dimensional space shown in Figure 1. Assume that the training set size is large enough so that the probabilities can be calculated accurately based on the areas of the selected regions.

The space is divided into three classes A, B, and C. In this exercise, you will build a decision tree from the training set.

(a) Compute the entropy for the overall data.

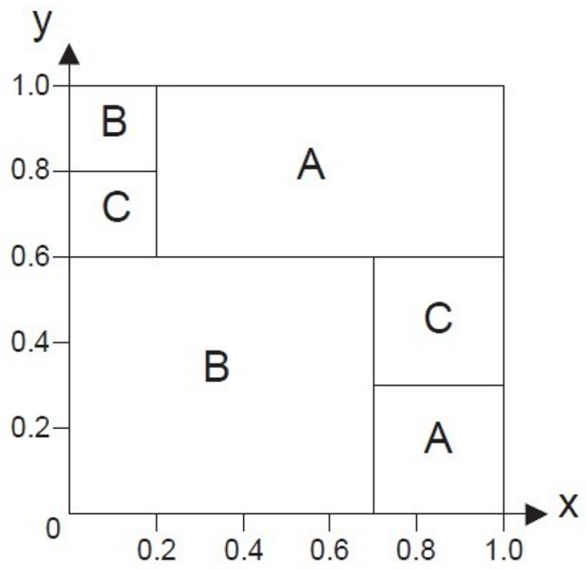

Figure 1: 2D Region

(b) Compare the entropy when the data is split at $x \leq 0.2$, $x \leq 0.7$, and $y \leq 0.6$.

(c) Based on your answer in part (b), which attribute split condition should be used as the root of the decision tree.

(d) Draw the full decision tree for the data set. Constraints: Use only the decision boundaries $x = 0.7$, $x = 0.2$, $y = 0.6$, $y = 0.3$, $y = 0.8$ to contruct your tree. Each level of the tree will maximize the information gain. Use the Tree class to save the tree within a dictionary of dictionaries. (Show an example of how this is done) (TODO).

Question 3:
Consider the training examples shown in Table (Table 2) below for a binary classification problem.

(a) Compute the Gini index for the overall collection of training examples.

(b) Compute the Gini index for the Customer ID attribute.

Customer ID	Gender	Car Type	Class
1	M	Family	C0
2	M	Sports	C0
3	M	Sports	C0
4	M	Sports	C0
5	M	Sports	C0
6	M	Sports	C0
7	F	Sports	C0
8	F	Sports	C0
9	F	Sports	C0
10	F	Luxury	C0
11	M	Family	C1
12	M	Family	C1
13	M	Family	C1
14	M	Luxury	C1
15	F	Luxury	C1
16	F	Luxury	C1
17	F	Luxury	C1
18	F	Luxury	C1
19	F	Luxury	C1
20	F	Luxury	C1

Figure 2:

(c) Compute the Gini index for the Gender attribute.

(d) Compute the Gini index for the Car Type attribute using multiway split.

(e) Which of the three attributes has the lowest Gini index? How did you come to this conclusion?

(f) Which of the three attributes will you use for splitting at the root node? Explain your choice.

Question 4:

Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

Example: Age in years. Answer: Discrete, quantitative, ratio

(a) Time in terms of AM or PM.

(b) Brightness as measured by a light meter.

(c) Brightness as measured by peoples judgments.

(d) Angles as measured in degrees between 0 and 360.

(e) Bronze, Silver, and Gold medals as awarded at the Olympics.

(f) Height above sea level.

(g) Number of patients in a hospital.

(h) ISBN numbers for books. (Look up the format on the Web.)

(i) Ability to pass light in terms of the following values: opaque, translucent, transparent.

(j) Military rank.

(k) Distance from the center of campus.

(l) Density of a substance in grams per cubic centimeter.

(m) Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.)

Question 5:
You are given a classification dataset with 100 instances, which has been partitioned into two subsets, dataset A with 50 instances and dataset B with 50 instances. Dataset A is used for training and dataset B is used for testing. You are supposed to compare two classification models: Model 1, which is an unpruned decision tree, and Model 2, which is a pruned version of the decision tree. The accuracy of the two classification models on datasets A and B are shown in the table below.

Classification Accuracy	Dataset A	Dataset B
Model 1	0.98	0.72
Model 2	0.82	0.8

(a) Based on the accuracies shown in the table above, which classification model would you expect to have better performance on unseen instances? Support your answer with a brief explanation.

(b) Now, you tested Model 1 and Model 2 on the entire dataset (A + B) and found that the classification accuracy of Model 1 on dataset (A + B) is 0.85, whereas the classification accuracy of Model 2 on the dataset (A + B) is 0.81. Based on this new information and your observations from the table above, which classification model would you finally choose for classification? Provide a brief explanation.

(c) Both Minimum Description Length (MDL) and the pessimistic error estimate are techniques used for incorporating model complexity into the loss function. State one similarity and one difference between them in the context of decision trees.

Question 6:
Consider the two-dimensional data shown in Figure **??**. The data consists of two classes: A and B.

(a) Draw a 2-level decision tree for the data (see Figure 3.10(b)). Use gini index as the splitting criterion. Assume the classier uses a binary split, i.e., the splitting criterion at each internal node must be specied either as $x \leq c$ or $y \leq c$, where c is some constant. In other words, do not specify the splitting criteria as $0.5 \leq x \leq 1.0$ or $x + y \leq 1$.

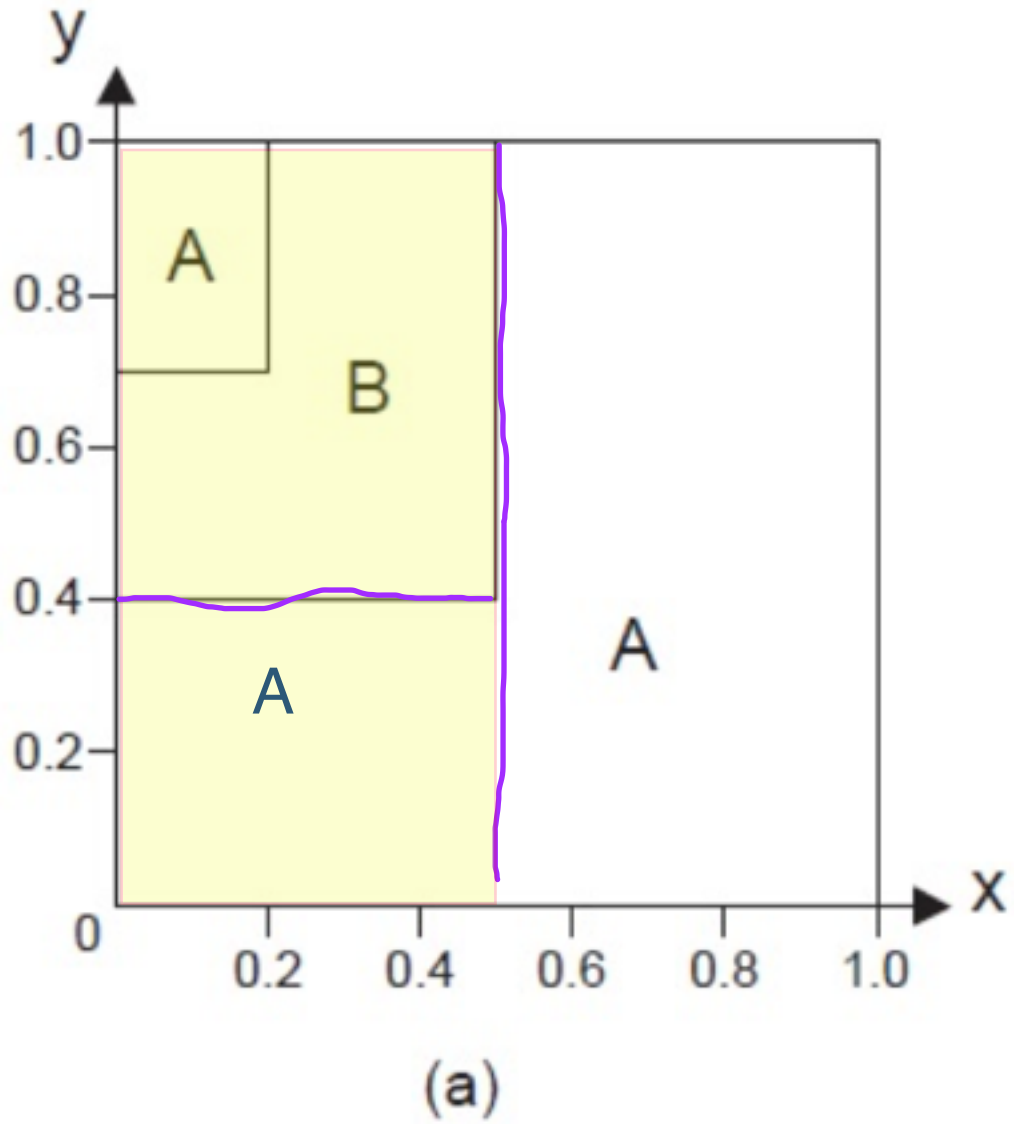

Figure 3: 2D Region

(b) Compute the expected error rate of your decision tree when it is applied to a test set randomly sampled from the same 2-d space.

Question 7:

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 dataID (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.

(a) Compute the information gain if we use ID as the splitting attribute.

(b) Repeat part (a) using Handedness as the splitting attribute.

(c) Based on your answers in parts (a) and (b), which attribute will be chosen according to information gain?

(d) Repeat part (a) using gain ratio (instead of information gain).

(e) Repeat part (b) using gain ratio (instead of information gain).

(f) Based on your answers in parts (d) and (e), which attribute will be chosen according to gain ratio?