Solutions Template

1 Written Questions

Write your solutions to the HW2 Written Questions in this template. You may either enter your solutions into the .tex file directly, or print out the pdf version and write your solutions by hand in the boxes provided. When you have completed all of the written questions, you should upload your solutions to Gradescope in pdf format.

1.1 Warm-up

First, let's think a little bit about decision trees. The following dataset consists of 7 examples, each with 3 attributes, (A, B, C), and a label, Y.

A	B	C	Y
1	1	0	0
1	1	2	1
1	0	0	1
1	1	2	1
0	0	2	0
0	1	1	0
0	0	0	0

Use the data above to answer the following questions.

A few important notes:

- *All calculations should be done without rounding!* After you have finished all of your calculations, write your rounded solutions in the boxes below.
- Note that, throughout this homework, we will use the convention that the leaves of the trees do not count as nodes, and as such are not included in calculations of depth and number of splits. (For example, a tree which classifies the data based on the value of a single attribute will have depth 1, and contain 1 split.)

1. What is the entropy of Y in bits, $H(Y)$? In this and subsequent questions, when we request the units in *bits*, this simply means that you need to use log base 2 in your calculations.[1] (Please include one number rounded to the fourth decimal place, e.g. 0.1234)

2. What is the mutual information[2] of Y and A in bits, $I(Y; A)$? (Please include one number rounded to the fourth decimal place, e.g. 0.1234)

[1] If instead you used log base e, the units would be *nats*; log base 10 gives *bats*.

[2] In the context of decision trees, and therefore in this assignment, the terms "information gain" and "mutual information" are synonymous (e.g. they have the same meaning as one another). However, in information theory and machine learning in general, "information gain" is synonymous with "Kullback-Leibler (K-L) divergence", often referred to as relative entropy, which is not the same as mutual information. For more information, go here.

3. What is the mutual information of Y and B in bits, $I(Y;B)$? (Please include one number rounded to the fourth decimal place, e.g. 0.1234)

4. What is the mutual information of Y and C in bits, $I(Y;C)$? (Please include one number rounded to the fourth decimal place, e.g. 0.1234)

5. Consider the dataset given above. Which attribute (A, B, or C) would a decision tree algorithm pick first to branch on, if its splitting criterion is mutual information?

6. Consider the dataset given above. Which is the second attribute you would pick to branch on, if its splitting criterion is mutual information? (*Hint:* Notice that this question correctly presupposes that there is *exactly one* second attribute.)

7. If the same algorithm continues until the tree perfectly classifies the data, what would the depth of the tree be?

8. Draw your completed Decision Tree. Label the non-leaf nodes with which attribute the tree will split on (e.g. B), the edges with the value of the attribute (e.g. 1 or 0), and the leaf nodes with the classification decision (e.g. $Y = 0$).

1.2 Empirical Questions

The following questions should be completed as you work through the programming portion of this assignment.

9. Train and test your decision tree on the politician dataset and the education dataset with four different values of max-depth, $\{0, 1, 2, 4\}$. Report your findings in the table below. A Decision Tree with max-depth 0 is simply a *majority vote classifier*; a Decision Tree with max-depth 1 is called a *decision stump*. If desired, you could even check that your answers for these two simple cases are correct using your favorite spreadsheet application (e.g. Excel, Google Sheets).

Dataset	Max-Depth	Train Error	Test Error
politician	0		
politician	1		
politician	2		
politician	4		
education	0		
education	1		
education	2		
education	4		

10. For the politicians dataset, create a plot showing error on the y-axis against depth of the tree on the x-axis. Plot *both* training error and testing error, clearly labeling which is which. That is, for each possible value of max-depth ($0, 1, 2, \ldots$, up to the number of attributes in the dataset), you should train a decision tree and report train/test error of the model's predictions.

11. Suppose your research advisor asks you to run some model selection experiments and then report your results. You select the Decision Tree model's max-depth to be the one with lowest test error in metrics.txt and then report that model's test error as the performance of our classifier on held out data. Is this a good experimental setup? If so, what is the name for this type of model selection? If not, why not?

12. In this assignment, we used max-depth as our stopping criterion, and as a mechanism to prevent overfitting. Alternatively, we could stop splitting a node whenever the mutual information for the best attribute is lower than a threshold value. This threshold would be another hyperparameter. How would this threshold value affect the number of nodes and depth of the learned trees? In a practical setting, how would you choose this parameter?

13. Print the decision tree which is produced by your algorithm for the politician data with max depth 3. Instructions on how to print the tree could be found in the HW2 writeup.

14. After you have completed all other components of this assignment, report your answers to the collaboration policy questions detailed in the Academic Integrity Policies found here. (a) Did you receive any help whatsoever from anyone in solving this assignment? Is so, include full details. (b) Did you give any help whatsoever to anyone in solving this assignment? Is so, include full details? (c) Did you find or come across code that implements any part of this assignment ? If so, include full details.