HOMEWORK 2: DECISION TREES

10-601 Introduction to Machine Learning (Spring 2021)
http://www.cs.cmu.edu/~mgormley/courses/10601/
DUE: Monday, February 22, 2021 11:59 PM

Summary It's time to build your first end-to-end learning system! In this assignment, you will build a Decision Tree classifier and apply it to several binary classification problems. This assignment consists of several parts: In Written component, you will work through some Information Theory basics in order to "learn" a Decision Tree on paper. Then in Programming component, you will implement Decision Tree learning, prediction, and evaluation. Using that implementation, you the empirical questions found at the end of the Written component.

START HERE: Instructions

- **Collaboration Policy**: Please read the collaboration policy here: http://www.cs.cmu.edu/~mgormley/courses/10601/about.html
- **Late Submission Policy:** See the late submission policy here: http://www.cs.cmu.edu/~mgormley/courses/10601/about.html
- **Submitting your work:** You will use Gradescope to submit answers to all questions and code. Please follow instructions at the end of this PDF to correctly submit all your code to Gradescope.
  - **Written:** For written problems such as short answer, multiple choice, derivations, proofs, or plots, we will be using Gradescope (https://gradescope.com/). Please use the provided template. Submissions must be written in LaTeX. Regrade requests can be made, however this gives the staff the opportunity to regrade your entire paper, meaning if additional mistakes are found then points will be deducted. Each derivation/proof should be completed in the boxes provided. For short answer questions you **should not** include your work in your solution. If you include your work in your solutions, your assignment may not be graded correctly by our AI assisted grader.
  - **Programming:** You will submit your code for programming questions on the homework to Gradescope (https://gradescope.com). After uploading your code, our grading scripts will autograde your assignment by running your program on a virtual machine (VM). When you are developing, check that the version number of the programming language environment (e.g. Python 3.6.9, OpenJDK 11.0.5, g++ 7.4.0) and versions of permitted libraries (e.g. numpy 1.17.0 and scipy 1.4.1) match those used on Gradescope. You have unlimited Gradescope programming submissions. However, we recommend debugging your implementation on your local machine (or the linux servers) and making sure your code is running correctly first before submitting you code to Gradescope.
- **Materials:** The data that you will need in order to complete this assignment is posted along with the writeup and template on Piazza.

Written Questions (30 points)

1. Warm-Up

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

A	B	C	Y
0	2	0	0
0	1	0	1
0	0	1	0
0	1	0	1
1	2	0	1
1	1	1	0
1	2	1	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.)

Note: Showing your work in these questions is optional, but it is recommended to help us understand where any misconceptions may occur. Only your numerical answer in the left box will be graded.

(a) (1 point) 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)

$H(Y)$	Work
0.9852	$H(Y) = -\left(\frac{4}{7}\log_2\left(\frac{4}{7}\right) + \frac{3}{7}\log_2\left(\frac{3}{7}\right)\right)$ $= 0.9852$

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

(b) (1 point) What is the mutual information of Y and A in bits, $I(Y;A)$? (Please include one number rounded to the fourth decimal place, e.g. 0.1234)

$I(Y;A)$	Work
0.0202	$I(Y;A) = H(Y) - H(Y\|A=0) - H(Y\|A=1)$ $= 0.9852 - \frac{4}{7}[-(\frac{1}{2}\log_2(\frac{1}{2})\cdot 2] - \frac{3}{7}[(\frac{2}{3}\log_2(\frac{2}{3}) + \frac{1}{3}\log_2(\frac{1}{3})]$ $= 0.0202$

(c) (1 point) 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)

$I(Y;B)$	Work
0.1981	$I(Y;B) = H(Y) - H(Y\|B=0) - H(Y\|B=1) - H(Y\|B=2)$ $= 0.9852 - 0 - 2\cdot\frac{3}{7}[-(\frac{2}{3}\log_2(\frac{2}{3}) + \frac{1}{3}\log_2(\frac{1}{3})]$ $= 0.1981$

(d) (1 point) 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)

$I(Y;C)$	Work
0.5216	$I(Y;C) = H(Y) - H(Y\|C=0) - H(Y\|C=1)$ $= 0.9852 - \frac{4}{7}[-(\frac{1}{4}\log_2(\frac{1}{4}) + \frac{3}{4}\log_2(\frac{3}{4})] - 0$ $= 0.5216$

(e) (1 point) 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?

Select one:

- ○ A
- ○ B
- ● C

(f) (1 point) Consider the dataset given above. After making the first split, which attribute would pick to branch on next, if the splitting criterion is mutual information? (*Hint:* Notice that this question correctly presupposes that there is *exactly one* second attribute.)

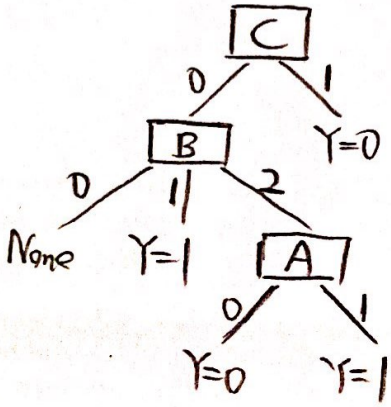
Select one:

- ○ A
- ● B
- ○ C

(g) (1 point) If the same algorithm continues until the tree perfectly classifies the data, what would the depth of the tree be?

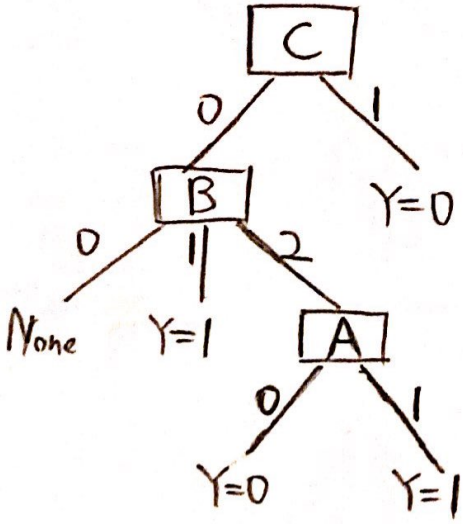
Depth

3



(h) (4 points) 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. $Y = 0$). You should include an image file below using the provided, commented out code in LaTeX, switching out *DecTree.png* to your file name as needed. The image may be hand-drawn.

Decision Tree



2. Empirical Questions

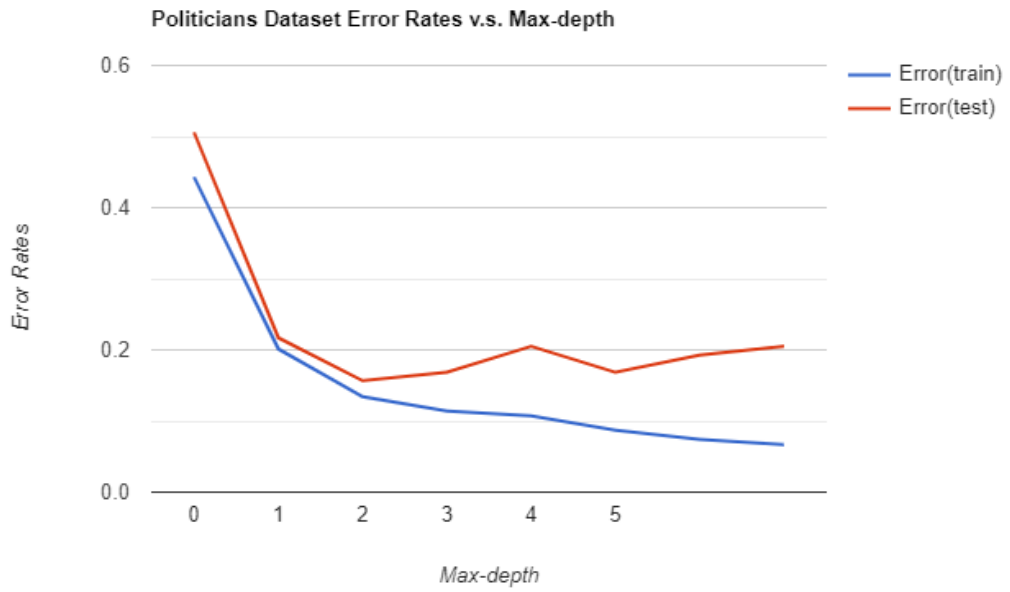
The following questions should be completed as you work through the programming portion of this assignment.

(a) (3 points) 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 HW2 solutions template provided. 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). (Please round each number to the fourth decimal place, e.g. 0.1234)

Dataset	Max-Depth	Train Error	Test Error
politician	0	0.4430	0.5060
politician	1	0.2013	0.2169
politician	2	0.1342	0.1566
politician	4	0.1074	0.2048
education	0	0.325	0.31
education	1	0.195	0.23
education	2	0.195	0.23
education	4	0.13	0.155

(b) (3 points) For the politicians dataset, create a computer-generated 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. You should include an image file below using the provided, commented out code in LaTeX, switching out *politician.png* to your file name as needed.

Plot



(c) (3 points) 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 test data. Is this a good experimental setup? If so, why? If not, why not?

Answer

No, because for the decision tree we are using batch learning, which means the algorithm that's been developed should learn from all the examples at once instead of gradually learning over time. Thus, the hyperparameter optimization, which is selecting the max-depth value in this case, should be determined by looking at the lowest training error. We wouldn't know the performance of the current setup described in the question until a new test data is given.

(d) (3 points) 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. Theoretically, how would increasing this threshold value affect the number of nodes and depth of the learned trees?

Answer

Theoretically, there would be less nodes and a shallower learned tree. Everytime the data is split, the subsets created should have lower entropies as the tree grows deeper until the subsets are perfectly classified, which indicates 0 entropy. If there exists a threshold for mutual information, the tree stops growing earlier when the entropy of a subset is low enough.

(e) (3 points) Continuing from the previous question, how would you set-up model training to choose the threshold value?

Answer

In class, we learned a method called cross-validation. We could split the training data into N subsets and choose the threshold value to be 0 to 1 with an interval of 0.1. Next, we obtain the N-fold cross validation error for each threshold, and pick the threshold that produces the lowest cross-validation error.

(f) (4 points) Print (do not handwrite!) 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 section 3.4.

Output

```
% YOUR ANSWER
% Text here will be compiled verbatim.
% So do not add unnecessary indents


[83 democrat/66 republican]
| Superfund_right_to_sue = y: [28 democrat/64 republican]
| | Aid_to_nicaraguan_contras = n: [13 democrat/58 republican]
| | | Export_south_africa = y: [13 democrat/38 republican]
| | | Export_south_africa = n: [0 democrat/20 republican]
| | Aid_to_nicaraguan_contras = y: [15 democrat/6 republican]
| | | Mx_missile = n: [12 democrat/0 republican]
| | | Mx_missile = y: [3 democrat/6 republican]
| Superfund_right_to_sue = n: [55 democrat/2 republican]
| | Export_south_africa = y: [55 democrat/1 republican]
| | | Immigration = y: [9 democrat/1 republican]
| | | Immigration = n: [46 democrat/0 republican]
| | Export_south_africa = n: [0 democrat/1 republican]
```

(g) After you have completed all other components of this assignment, report your answers to these questions regarding the collaboration policy. Details of the policy can be found here. 1. Did you receive any help whatsoever from anyone in solving this assignment? Is so, include full details. 2. Did you give any help whatsoever to anyone in solving this assignment? Is so, include full details. 3. Did you find or come across code that implements any part of this assignment ? If so, include full details.

Answer

3. I copied code from HW1 that I wrote myself.