22417.202210 Homework 1 - Decision Trees

Pankaj Kumar Jatav

TOTAL POINTS

31.95 / 36

QUESTION 1

1 Math 1 1 / 1

✓ **- 0 pts** *Correct*

- 1 pts Incorrect

QUESTION 2

2 Math 2 2 / 2

✓ **- 0 pts** *Correct*

- 0.2 pts Why are f(X,Y) and f(X)f(Y) equal?

- 0.8 pts Explanation of why is E[X,Y] equal to E[X]E[Y] missing.

- 0.1 pts Slight error in notation

- 2 pts Correlation is defined based on covariance -- not the other way around.

- 2 pts "Show" means "show mathematically", not "explain"

QUESTION 3

Math 3 4 pts

3.1 Result 1.55 2 / 2

✓ **- 0 pts** *Correct*

- 1 pts Correct derivative, wrong solution for $$\mu_{ML}$$.

- 0.5 pts In your solution, why would we remove the "constant" $$\sigma^2$$?

- 1 pts Incorrect partial derivative

- 0.1 pts What's up with the d/du everywhere?

- 0 pts We don't "differentiate both sides".

- 0.2 pts Slightly incorrect partial derivative

- 0.3 pts Notational issues

- 1 pts Step 3 not explained.

- 2 pts I cannot follow what's happening here.

3.2 Result 1.56 2 / 2

✓ **- 0 pts** *Correct*

- 0.4 pts We need to take the derivative with respect to $$\sigma^2$$, not with respect to $$\sigma$$.

- 2 pts I don't understand what's happening here

- 1 pts Notational Issues. What does each "then" correspond to?

QUESTION 4

4 Math 4 4 / 4

✓ **- 0 pts** *Correct*

QUESTION 5

DT 7 pts

5.1 DT - Q.5 1 / 1

✓ **- 0 pts** *Correct*

- 1 pts Incorrect (the answer is 1)

- 1 pts Blank

5.2 DT - Q.6 1 / 1

✓ **- 0 pts** *Correct*

- 0.5 pts Wrong rounding

- 1 pts The correct answer is 0.0488

- 0 pts Click here to replace this description.

5.3 DT - Q.7 **1 / 1**

✓ **- 0 pts** *Correct*

- 1 pts Missing weight in summation

5.4 DT - Q.8 **0.95 / 1**

- 0 pts Correct

- 1 pts Wrong application of the definition of information gain

- 1 pts Missing weighted sum

- 1 pts Incorrect

✓ **- 0.05 pts** *Error in rounding*

5.5 DT - Q.9 **1 / 1**

✓ **- 0 pts** *Correct*

- 1 pts Correct answer is C

5.6 DT - Q.10 **1 / 1**

✓ **- 0 pts** *Correct*

- 1 pts Incorrect answer (correct answer is A)

5.7 DT - Q.11 **1 / 1**

✓ **- 0 pts** *Correct (3)*

- 0 pts Correct (the actual depth is 3)

QUESTION 6

Programming 18 pts

6.1 Decision Tree Plot **2 / 2**

✓ **- 0 pts** *Correct*

- 2 pts Incorrect

6.2 Q.13 **1 / 2**

- 0 pts Correct

✓ **- 1 pts** *Unclear answer. Does not mention improvement in accuracy.*

- 1 pts Does not address the "how do you know" part. Should mention improvement in terms of information gain or train/dev accuracy.

6.3 Q.14 **1 / 1**

✓ **- 0 pts** *Correct*

6.4 Q.15 **2 / 2**

✓ **- 0 pts** *Correct*

6.5 Q.16 Dev Acc **1 / 1**

✓ **- 0 pts** *Correct*

- 0.2 pts Slightly off answer (correct is 0.62)

- 0.5 pts Off answer (correct is 0.62)

6.6 Q.17 - Three Plots **3 / 3**

✓ **- 0 pts** *Correct*

- 1 pts Partially correct.

- 3 pts Incorrect

6.7 Q.18 - Discussion **0 / 2**

- 0 pts Correct

- 1 pts Partially correct

✓ **- 2 pts** *Incorrect*

6.8 Q.19 - Test Error 1 **0.5 / 1**

- 0 pts Correct answers are 0.3875 or 0.385.

- 1 pts Incorrect answer (Correct answers are 0.3875 or 0.385)

✓ **- 0.5 pts** *Slightly incorrect answer (Correct answers are 0.3875 or 0.385)*

6.9 Q.20 - Test Error 2 **1 / 1**

✓ **- 0 pts** *Correct*

- 0.5 pts Slightly different from 0.3675

- 1 pts Incorrect answer

6.10 Q.21 - Test Error 3 **0.5 / 1**

- 0 pts Correct

✓ **- 0.5 pts** *Slightly different from correct answer*

0.355

- 1 pts Incorrect answer. Correct ans: 0.355

6.11 Q.22 - Last **2 / 2**

✓ **- 0 pts** *Correct*

- 1 pts Partially correct. Using training data will overfit; using testing data is biased. Correct answer: dev data.

- 2 pts Incorrect. Using training data will overfit; using testing data is biased. Correct answer: dev data.

QUESTION 7

7 Collaboration Questions **0 / 0**

✓ **- 0 pts** *Correct*

HOMEWORK 1
DECISION TREES AND OVERFITTING[1]

CS 688 MACHINE LEARNING (SPRING 2022)
https://nlp.cs.gmu.edu/course/cs688-spring22/

OUT: Feb 3, 2022
DUE: Feb 9, 2022

Name: Pankaj Kumar Jatav

GID: G01338769

[1]Compiled on Friday 11th February, 2022 at 04:04

1 Written Questions [36 pts]

1.1 A bit of math

1. (1 point) The textbook (Bishop) defines the variance of $f(x)$ (definition 1.38) as:

$$\text{var}[f] = \mathbb{E}\Big[(f(x) - \mathbb{E}[f(x)])^2\Big]$$

Use that definition to show that $\text{var}\,[f(x)]$ satisfies that:

$$\text{var}[f] = \mathbb{E}\big[f(x)^2\big] - \mathbb{E}[f(x)]^2.$$

Work

$$\text{var}[f] = \mathbb{E}\Big[(f(x) - \mathbb{E}[f(x)])^2\Big]$$

$$= \mathbb{E}\Big[(f(x)^2 - 2f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2\Big]$$

$$= \mathbb{E}\Big[f(x)^2 + [-2\mathbb{E}[f(x)]] + \mathbb{E}[f(x)]^2\Big]$$

(Use theorem, $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ where X and Y are random variable)

$$= \mathbb{E}[f(x)^2] + \mathbb{E}[-2f(x)\mathbb{E}[f(x)]] + \mathbb{E}[f(x)]^2$$

(Use theorem, $\mathbb{E}[aX] = a\mathbb{E}[X]$ where X is random variable and c is constant,
In this case $f(x)$ is random variable and $-2\mathbb{E}[f(x)]$ is a constant)

$$= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2$$

$$= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]^2 + \mathbb{E}[f(x)]^2$$

$$= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

- Hence Proved

2. (2 points) Show that if two discrete random variables x and y are independent, then their covariance is 0.

Work

For two random variables x and y, the covariance is defined by (Bishop 1.41)

$$cov[x,y] = \mathbb{E}_(x,y)[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}]$$

$$= \mathbb{E}_(x,y)[xy - y\mathbb{E}[x] + x\mathbb{E}[y] + \mathbb{E}[x]\mathbb{E}[y]]$$

(Use theorems, $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ and $\mathbb{E}[aX] = a\mathbb{E}[X]$)

$$= \mathbb{E}_(x,y)[xy] - \mathbb{E}[y]\mathbb{E}[x] + \mathbb{E}[x]\mathbb{E}[y] + \mathbb{E}[x]\mathbb{E}[y]$$

$$= \mathbb{E}_(x,y)[xy] - \mathbb{E}[y]\mathbb{E}[x] - \mathbb{E}[x]\mathbb{E}[y] + \mathbb{E}[x]\mathbb{E}[y]$$

$$= \mathbb{E}_(x,y)[xy] - \mathbb{E}[y]\mathbb{E}[x]$$

Now we have show that $\mathbb{E}[xy] = \mathbb{E}[y]\mathbb{E}[x]$, which result cov(x,y) is zero.

$$p(x,y) = p((y|x)p(x)$$

As we know that x and y are independent variable, therefore $p(y|x) = p(y)$

$$p(x,y) = p(x)p(y) - - - - - - - - - - -(1)$$

Expection of f(x)

$$\mathbb{E}[f] = \sum_x p(x)f(x)$$

So,

$$\mathbb{E}_(x,y)[xy] = \sum_x \sum_y p(xy)f_{x,y}(x,y)$$

As we know that x and y are the random variable, using eq 1

$$\mathbb{E}_(x,y)[xy] = \sum_x \sum_y p(x)p(y)f_x(x)f_y(y)$$

$$\mathbb{E}_(x,y)[xy] = \sum_x p(x)f_x(x) \sum_y p(y)f_y(y)$$

$$\mathbb{E}_(x,y)[xy] = \mathbb{E}[y]\mathbb{E}[x]$$

Now we know that $\mathbb{E}[xy] = \mathbb{E}[y]\mathbb{E}[x]$, which means cov(x,y) is zero.

- Hence Proved

3. (4 points) The log likelihood function of a Gaussian distribution is given by (Bishop 1.54):

$$\ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2}\Sigma_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln 2\pi.$$

By setting its derivatives with respect to μ and σ^2 to zero, verify that the maximum likelihood estimates $\mu_{\text{ML}}, \sigma^2_{\text{ML}}$ are (Bishop result 1.55):

$$\mu_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n.$$

and (Bishop result 1.56):

$$\sigma^2_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\text{ML}})^2.$$

Work for result 1.55

$$\ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2}\Sigma_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln 2\pi$$

Taking the $\frac{d}{d\mu}$ for both side

$$\frac{d}{d\mu}\ln p(\mathbf{x}|\mu, \sigma^2) = \frac{d}{d\mu}(-\frac{1}{2\sigma^2}\Sigma_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln 2\pi)$$

$$\frac{d}{d\mu}\ln p(\mathbf{x}|\mu, \sigma^2) = \frac{d}{d\mu}(-\frac{1}{2\sigma^2}\Sigma_{n=1}^{N}(x_n - \mu)^2) - \frac{d}{d\mu}(\frac{N}{2}\ln\sigma^2) - \frac{d}{d\mu}(\frac{N}{2}\ln 2\pi)$$

Last two term in right side will be zero as they are constant with respect to μ

$$\frac{d}{d\mu}\ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2}(2\Sigma_{n=1}^{N}(x_n - \mu))) * (-1) - 0 - 0$$

Setting $\frac{d}{d\mu}\ln p(\mathbf{x}|\mu, \sigma^2)$ to zero to maximum

$$0 = \frac{1}{\sigma^2}\Sigma_{n=1}^{N}(x_n - \mu_{ML})$$

$$0 = \Sigma_{n=1}^{N}(x_n - \mu_{ML})$$

$$0 = \Sigma_{n=1}^{N}(x_n) - \Sigma_{n=1}^{N}(\mu_{ML})$$
$$\Sigma_{n=1}^{N}(x_n) = \Sigma_{n=1}^{N}(\mu_{ML})$$
$$\Sigma_{n=1}^{N}(x_n) = N(\mu_{ML})$$
$$\mu_{ML} = \frac{1}{N}\Sigma_{n=1}^{N}x_n$$

- Hence Proved

Work for result 1.56

$$\ln p(\mathbf{x}|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\Sigma_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln 2\pi$$

Taking the $\frac{d}{d\sigma^2}$ for both side

$$\frac{d}{d\sigma^2}\ln p(\mathbf{x}|\mu,\sigma^2) = \frac{d}{d\sigma^2}(-\frac{1}{2\sigma^2}\Sigma_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln 2\pi)$$

$$\frac{d}{d\sigma^2}\ln p(\mathbf{x}|\mu,\sigma^2) = \frac{d}{d\sigma^2}(-\frac{1}{2\sigma^2}\Sigma_{n=1}^{N}(x_n-\mu)^2) - \frac{d}{d\sigma^2}(\frac{N}{2}\ln\sigma^2) - \frac{d}{d\sigma^2}(\frac{N}{2}\ln 2\pi)$$

Last term in right side will be zero as they are constant with respect to σ^2

$$\frac{d}{d\sigma^2}\ln p(\mathbf{x}|\mu,\sigma^2) = (-\frac{1}{2}\frac{-1}{\sigma^4}\Sigma_{n=1}^{N}(x_n-\mu)^2) - \frac{N}{2}\frac{1}{\sigma^2} - 0$$

Setting $\frac{d}{d\sigma^2}\ln p(\mathbf{x}|\mu,\sigma^2)$ to zero to maximum

$$0 = \frac{1}{2\sigma_{ML}^4}\Sigma_{n=1}^{N}(x_n-\mu_{ML})^2 - \frac{N}{2}\frac{1}{\sigma_{ML}^2}$$

Taking $\frac{1}{2\sigma_{ML}^4}$ common

$$0 = \frac{1}{2\sigma_{ML}^4}(\Sigma_{n=1}^{N}(x_n-\mu_{ML})^2 - N\sigma_{ML}^2)$$

$$N\sigma_{ML}^2 = \Sigma_{n=1}^{N}(x_n-\mu_{ML})^2$$

$$\sigma_{ML}^2 = \frac{1}{N}\Sigma_{n=1}^{N}(x_n-\mu_{ML})^2$$

- Hence Proved

4. (4 points) Suppose that the variance of a Gaussian is estimated using the result (Bishop 1.56; see above question) but with the maximum likelihood estimate μ_{ML} replaced with the true value μ of the mean. Show that this estimator has the property that its expectation is given by the true variance σ^2.

Work

$$\mu_{ML} = \frac{1}{N}\Sigma_{n=1}^{N} x_n$$

$$\mathbb{E}[\mu_{ML}] = \mathbb{E}[\frac{1}{N}\Sigma_{n=1}^{N} x_n]$$

$$\mathbb{E}[\mu_{ML}] = \frac{1}{N}\Sigma_{n=1}^{N} \mathbb{E}[x_n]$$

$$\mathbb{E}[\mu_{ML}] = \frac{1}{N} N\mathbb{E}[x]$$

$$\mathbb{E}[\mu_{ML}] = \mathbb{E}[x]$$

Now we have to show that below equation is equal to true variance

$$\sigma_{ML}^2 = \frac{1}{N}\Sigma_{n=1}^{N} (x_n - \mu_{ML})^2$$

$$\sigma_{ML}^2 = \frac{1}{N}\Sigma_{n=1}^{N} (x_n^2 - 2x_n\mu_{ML} + \mu_{ML}^2)$$

$$\sigma_{ML}^2 = \frac{1}{N}(\Sigma_{n=1}^{N}(x_n^2) - \Sigma_{n=1}^{N}(2x_n\mu_{ML}) + \Sigma_{n=1}^{N}(\mu_{ML}^2))$$

$$\sigma_{ML}^2 = \frac{1}{N}(\Sigma_{n=1}^{N}(x_n^2) - 2\Sigma_{n=1}^{N}x_n\Sigma_{n=1}^{N}\mu_{ML} + \Sigma_{n=1}^{N}(\mu_{ML}^2))$$

$$\sigma_{ML}^2 = \frac{1}{N}(\Sigma_{n=1}^{N}(x_n^2) - 2N\mu_{ML}^2 + N\mu_{ML}^2)$$

$$\sigma_{ML}^2 = \frac{1}{N}\Sigma_{n=1}^{N}(x_n^2) - \mu_{ML}^2$$

Taking the expectation both side

$$\mathbb{E}[\sigma_{ML}^2] = \mathbb{E}[\frac{1}{N}\Sigma_{n=1}^{N}(x_n^2) - \mu_{ML}^2]$$

$$\mathbb{E}[\sigma_{ML}^2] = \mathbb{E}[\frac{1}{N}\Sigma_{n=1}^{N}(x_n^2)] - \mathbb{E}[\mu_{ML}^2]$$

$$\mathbb{E}[\sigma_{ML}^2] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$$

Above equation is equal to the true variance (Bishop 1.40)
- Hence Proved

1.2 Warm-Up to decision trees

Let's think a little bit more about decision trees. The following dataset D consists of 8 examples, each with 3 attributes, (A, B, C), and a label, Y.

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

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 that the dataset contains duplicate rows; treat each of these as their own example, do not remove duplicate rows.

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.

5. (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.[2] (Please include one number rounded to the fourth decimal place, e.g. 0.1234)

$H(Y)$	Work
1	The data is uniformly distributed $H(Y) = -0.5log_2(0.5) - 0.5log_2(0.5) H(Y) = -0.5(-1) - 0.5(-1) = 1$

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

6. (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
0.0488	

7. (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
0.3111	

8. (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
0.5486	

9. (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

10. (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

11. (1 point) If the same algorithm continues until the tree perfectly classifies the data, what would the depth of the tree be?

    Depth

    3

1.3 Empirical Questions

In this programming exercise, we'll explore using decision trees to make classification decisions on one simple binary classification task: sentiment analysis (is this review a positive or negative evaluation of a product?).

We'll use for prediction are simply the presence/absence of words in the text. If you look in `data/sentiment.tr`, you'll see training data for the sentiment prediction task. The first column is zero or one (one = positive, zero = negative). The rest is a list of all the words that appear in this product review. These are *binary* features: any word listed has value "=1" and any word not listed has value "=0" (implicitly... it would be painful to list all non-occurring words!).

Before we begin...

- Make sure you have installed SciKit-Learn correctly:

```
% python
>>> from sklearn.tree import DecisionTreeClassifier

```

If that doesn't work, something is wrong with your installation.

Part 1: Understanding what Decision Trees are doing Train a decision tree of (maximum) depth 2 on the sentiment data.

First, we need to load the data:

```
>>> from data import *
>>> X,Y,dictionary = loadTextDataBinary("dataset/sentiment.tr")
>>> X.shape
(1400, 3473)
>>> Y.shape
(1400,)
```

We have successfully loaded 1400 examples of sentiment training data. The vocabulary sie is 3473 words; we can look at the first ten words (arbitrarily sorted):

```
>>> dictionary[:10]
['hanging', 'woody', 'originality', 'bringing', 'wooden', 'woods', 'stereotypical',
    'shows', 'replaced', 'china']
```

Now, we can train a depth one decision tree (aka "decision stump") on this data:

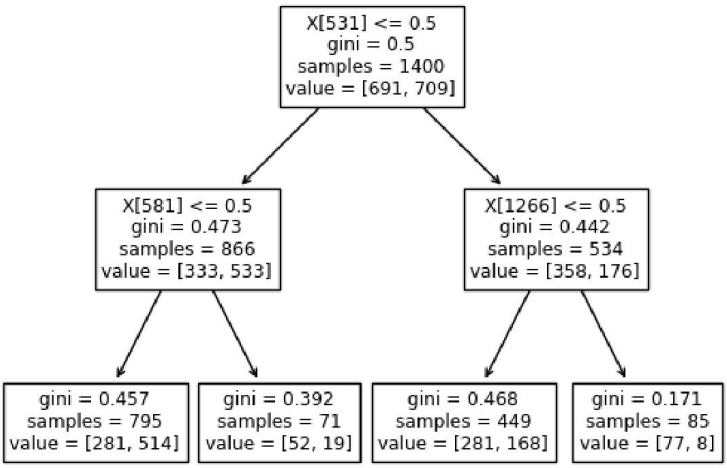
```
>>> from sklearn.tree import DecisionTreeClassifier
>>> dt = DecisionTreeClassifier(max_depth=1)
>>> dt.fit(X, Y)
DecisionTreeClassifier(compute_importances=None, criterion='gini',
            max_depth=1, max_features=None, min_density=None,
            min_samples_leaf=1, min_samples_split=2, random_state=None,
            splitter='best')
>>> showTree(dt, dictionary)
bad?
-N-> class 1   (333 for class 0, 533 for class 1)
-Y-> class 0   (358 for class 0, 176 for class 1)
```

This shows that if you only have one question you can ask about the review it's that you should ask if the review contains the word "bad" or not. If it does not ("N") then it's probably a positive review (by a vote of 533 to 333); if it does ("Y") then it's probable a negative review (by a vote of 358 to 176).

Your first task is to build a depth two decision tree.

12. (2 points) Draw it on a piece of paper or use scikit-learn's built in functions to draw it.

Decision Tree




```
bad?
-N-> worst?
|    -N-> class 1   (281 for class 0, 514 for class 1)
|    -Y-> class 0   (52 for class 0, 19 for class 1)
-Y-> stupid?
|    -N-> class 0   (281 for class 0, 168 for class 1)
|    -Y-> class 0   (77 for class 0, 8 for class 1)
```

13. (2 points) Convince yourself whether or not it is useful to go from depth one to depth two on this data. How do you know?

Work

Model 1(Max depth 1) perform well on 80-20 split. Also Below are the results:

$$MaxDepth1 0.6607142857142857$$

$$MaxDepth2 0.6428571428571429$$

After review the results, So we can say that it is not useful to go from depth one to depth two.

14. (1 point) It's important to recognize that decision trees are essentially learning *conjunctions* of features. In particular, you can convert a decision tree to a sequence of if-then-else statements, of the form:

```
if    A and  B and  C and  D then return POSITIVE
elif  A and  B and  C and !D then return NEGATIVE
elif  ...
```

This is called a "*decision list.*" Write down the decision list corresponding to the tree that you learned of depth 2.

Work

```
      if !bad and !worse then return 0
      elif  !bad and worse then return 1
      elif  bad and !worse then return 0
      elif  bad and worse then return 0

```

15. (2 points) Build a depth three decision tree and "explain" it. In other words, if your boss asked you to tell her, intuitively, what your tree is doing, how would you explain it? Write a few sentences.

Work

When we build the DT with the depth 3, we have 3 layer for predicting for the any unseen data or in other word we are performing max 8 binary classification.
In the first level, we are checking is this sentiment is bad or not. If it is bad then we are checking if the sentiment is worse or stupid, which is the second depth, And now in the third depth we are checking for wonderfully and bob under the worse and stupit and bob under stupid. And this can be understand easily using below diagram.

```
-N-> worst?
|    -N-> many?
|    |    -N-> class 1  (204 for class 0, 274 for class 1)
|    |    -Y-> class 1  (77 for class 0, 240 for class 1)
|    -Y-> present?
|    |    -N-> class 0  (52 for class 0, 13 for class 1)
|    |    -Y-> class 1  (0 for class 0, 6 for class 1)
-Y-> stupid?
|    -N-> wonderfully?
|    |    -N-> class 0  (280 for class 0, 153 for class 1)
|    |    -Y-> class 1  (1 for class 0, 15 for class 1)
|    -Y-> bob?
|    |    -N-> class 0  (76 for class 0, 4 for class 1)
|    |    -Y-> class 1  (1 for class 0, 4 for class 1)
None
```

16. (1 point) It's not enough to just think about training data; we need to see how well these trees generalize to new data. First, let's look at training accuracy for different trees:

    Depth 1:

```
>>> np.mean(dt.predict(X) == Y)
0.63642857142857145
```

    (Brief explanation: dt.predict(X) returns one prediction for each training example. We check to see if each of these is equal to Y or not. We want the average number that are equal, which is what the mean is doing. This gives us our training accuracy.)

    Depth 2:

```
>>> np.mean(dt.predict(X) == Y)
0.66000000000000003
```

    So the depth two tree does indeed fit the training data better. What about development data?

```
>>> Xde,Yde,_ = loadTextDataBinary("data/sentiment.de", dictionary)
```

(Note: when we load the development data, we have to give it the dictionary we built on the training data so that words are mapped to integers in the same way!)

What accuracy do you obtain on the dev data with the depth-2 tree?

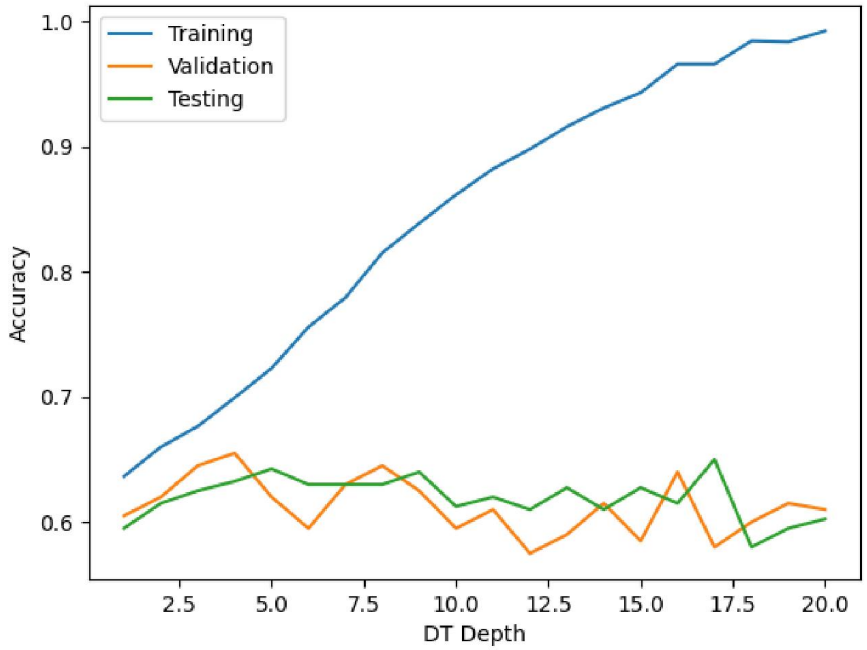
Dev Accuracy

0.62

Here, you should see that the accuracy has dropped a bit.

17. (3 points) For all possible depths from depth 1 to depth 20, compute training error, development error and test error for the corresponding decision tree (hint: use a for loop :P). Plot these three curves (yes, by hand if you must; we recommend the `matplotlib` or `seaborn` Python libraries).

Plots



18. (2 points) What trend do you observe for the training error rates? Why should this happen? Write a few sentences:

Work

As we increase the depth of the decision tree, the training data accuracy increase but the model does not perform well on dev and test data, Because model is over-fit with large DT depth.

19. (1 point) If you were to choose the depth hyperparameter based on TRAINING data, what TEST error would you get?

Test error

0.395

20. (1 point) If you were to choose depth based on the DEV data, what TEST error would you get?

Test error

0.3675

21. (1 point) Finally, if you were to choose the depth based on the TEST data, what TEST error would you get.

Test error

0.3525

22. (2 points) Precisely one of these three above is "correct" – which one and why?

Work

Second one(20) with test error is 0.3675 is correct because at that DT depth four the model is not too easy or too complex, So it is not affected by underfit or overfit.

Collaboration Questions Please answer the following:

1. Did you receive any help whatsoever from anyone in solving this assignment?
   No.
   - If you answered 'yes', give full details: _____
   - (e.g. "Jane Doe explained to me what is asked in Question 3.4")
2. Did you give any help whatsoever to anyone in solving this assignment?
   No.
   - If you answered 'yes', give full details: _____
   - (e.g. "I pointed Joe Smith to section 2.3 since he didn't know how to proceed with Question 2")
3. Did you find or come across code that implements any part of this assignment ?
   No. (See below policy on "found code")
   - If you answered 'yes', give full details: _____
   - (book & page, URL & location within the page, etc.).