Part 0

0.1

2.4

Nathaniel Li

3.3, 3.4

Foris Kuang

0.2

None

0.3

I have read and understood these policies.

Part 1

1.1

MIN(p‘, n’)

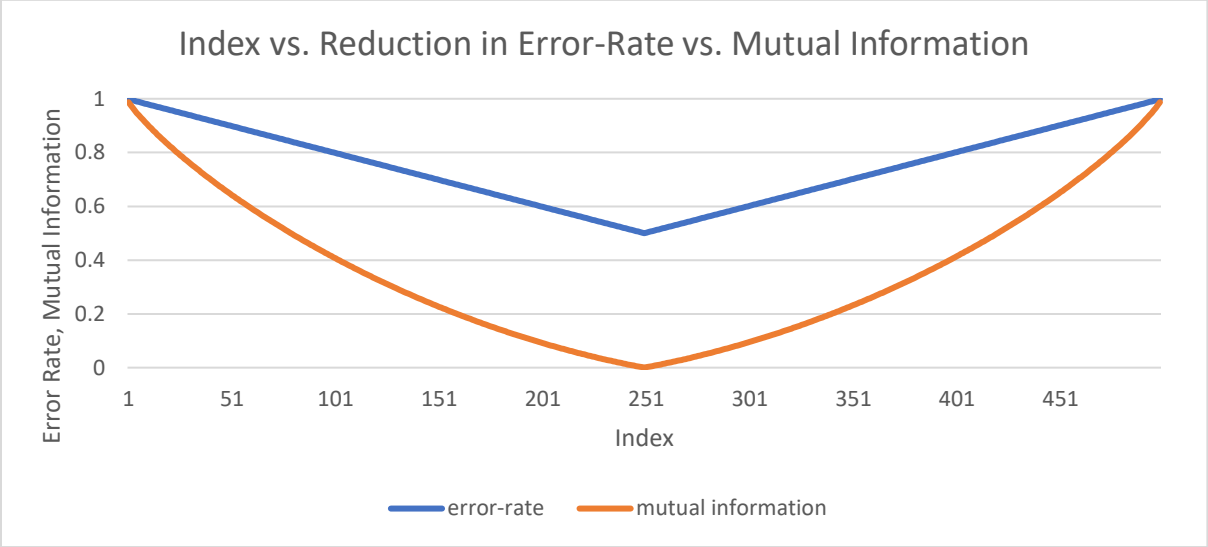
Once we reach the maximum depth and get our subset D’, we have to choose the MAX(p’, n’) as the correct leaf. Thus, the MIN of the two will be our error as we chose the MAX.

1.2

MIN(n_0 ‘, p_0 ‘) + MIN(n_1 ‘, p_1 ‘)

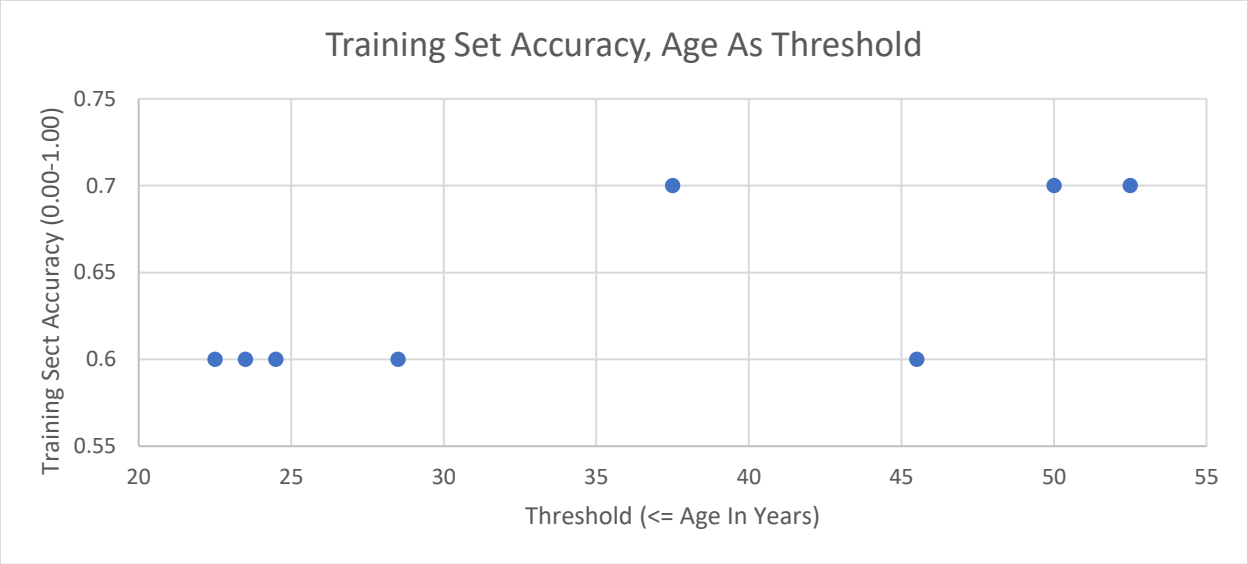
We split on feature phi. If we look at the 0 branch, we see we have n_0:p_0 results. We will take the MAX of those two to be correct, thus leaving the MIN to be our mistakes. Similarly for the 1 branch, we have n_1:p_1 results, and the MIN of those two are our mistakes. Thus the total sum of mistakes by splitting on feature phi is the sum of the mistakes in each branch.

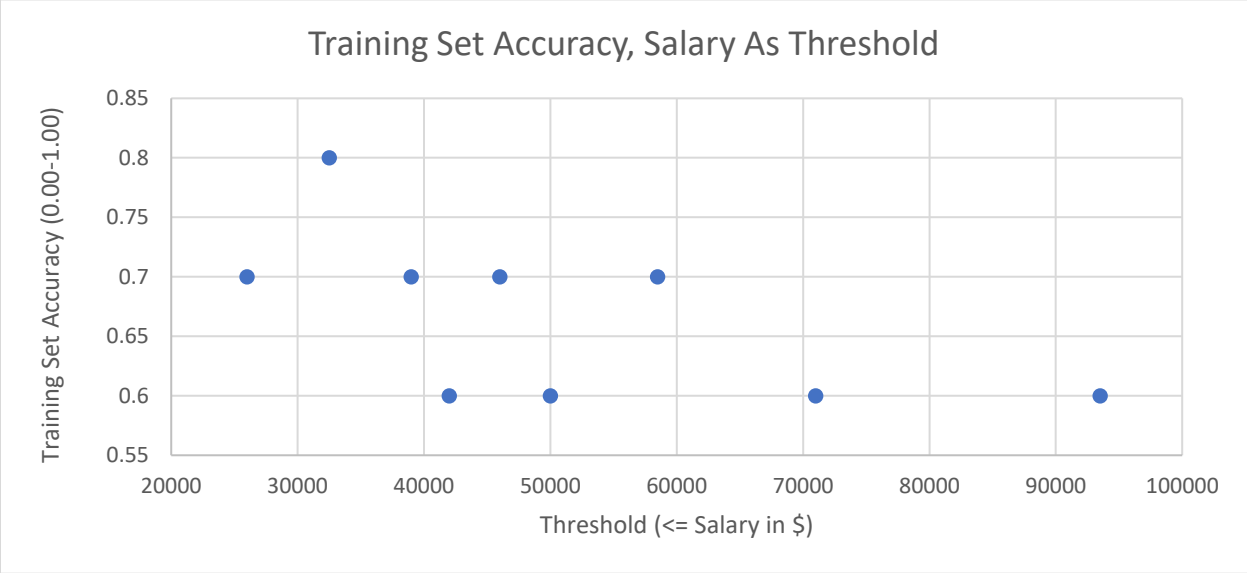
1.3

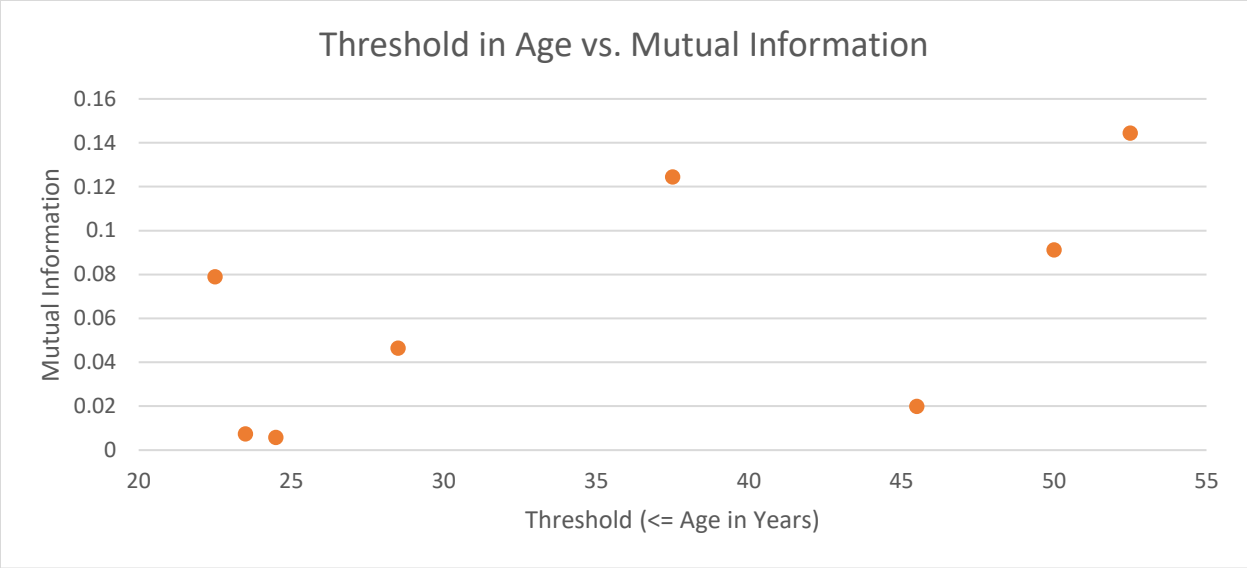


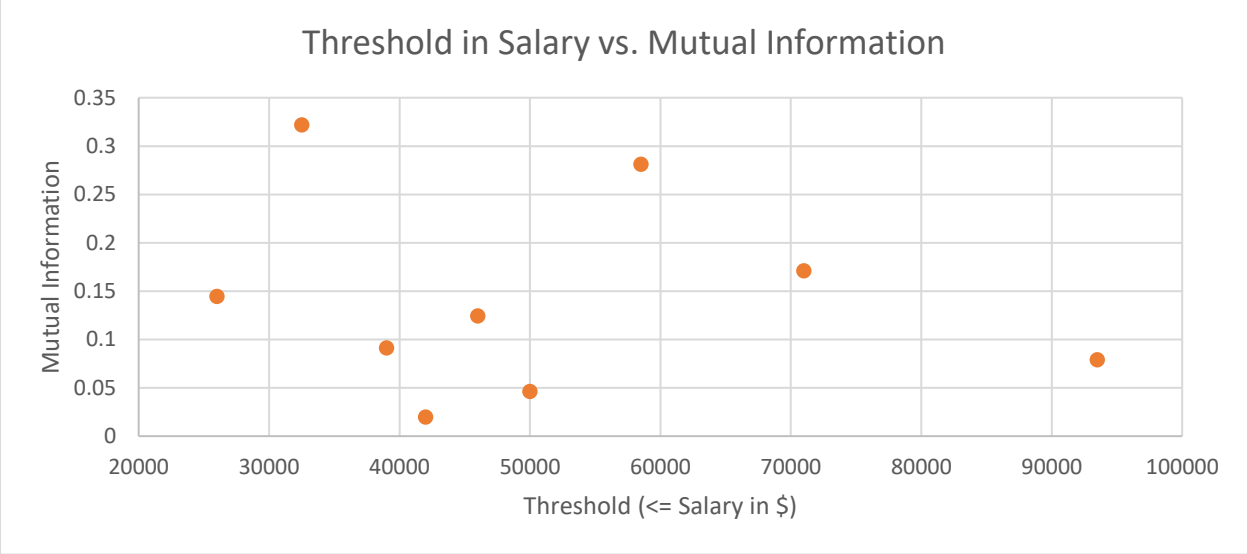
Part 2

2.1


Training Set Accuracy, Age As Threshold
Training Sect Accuracy (0.00-1.00)
0.75
0.7
0.65
0.6
0.55
20
25
30
35
40
45
50
55
Threshold (<= Age In Years)


Training Set Accuracy, Salary As Threshold
Training Set Accuracy (0.00-1.00)
0.85
0.8
0.75
0.7
0.65
0.6
0.55
20000
30000
40000
50000
60000
70000
80000
90000
100000
Threshold (<= Salary in $)

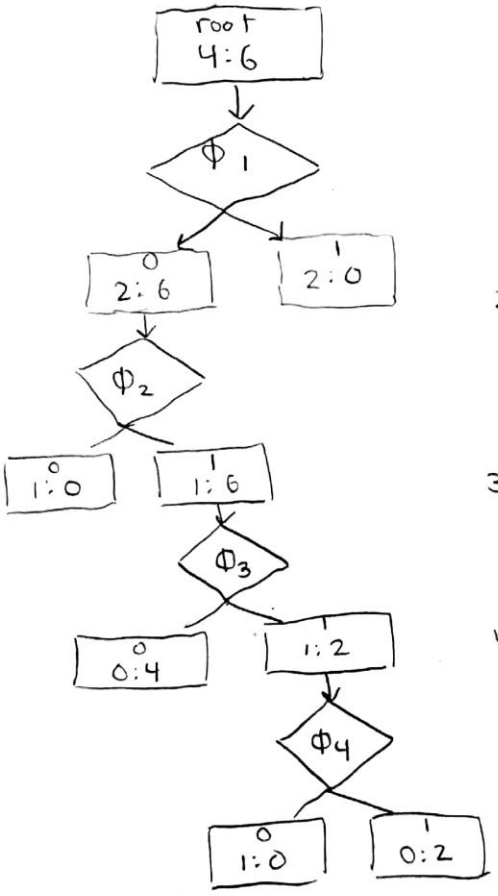

Threshold in Age vs. Mutual Information
Mutual Information
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
20
25
30
35
40
45
50
55
Threshold (<= Age in Years)


Threshold in Salary vs. Mutual Information
Mutual Information
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
20000
30000
40000
50000
60000
70000
80000
90000
100000
Threshold (<= Salary in $)

2.2

2.2

Greedy



1) Choose Φ_1 (Salary $\leq$ 32,500)
0.8 highest Training set accuracy
8/10
80%.

2) Choose Φ_2 (Age $\leq$ 52.5)
0.7

3) Choose Φ_3 (Salary $\leq$ 46,000)
0.7

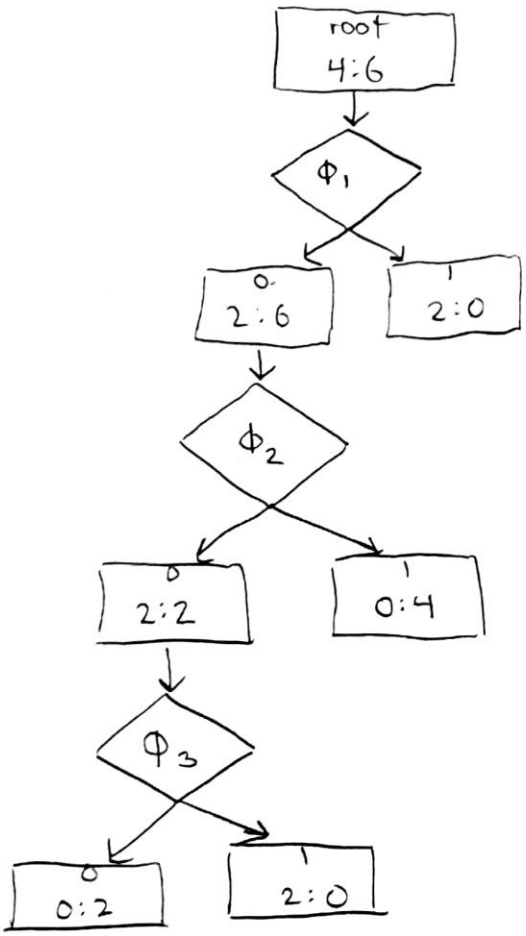
4) Choose Φ_4 (Age $\leq$ 37.5)
0.7

Training set
Error rate has been
reduced to 0%

o incorrectly classified

2.2

Mutual Information



1) Choose Φ_1 (Highest mutual information score = 0.3219, Salary $\leq$ 32,500)

2) Choose Φ_2 (Highest mutual information score = 0.3066, Age $\leq$ 37.5

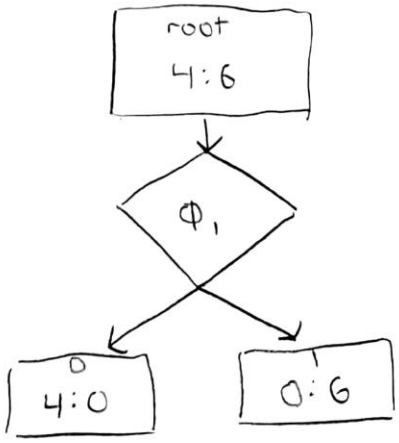
3) Choose Φ_3 (Highest mutual information score = 0.7118, Salary $\leq$ 58,500)

Training set error rate has been reduced to 0%.

0 incorrect classifications

2.3

2.3



$\alpha = -1100, \beta = 1$

Φ_1 = sign of
$-1100(\text{Age}) + \text{Income} - 1$

0% error rate

$\alpha x_{AGE} + \beta x_{INCOME} - 1$

$= -1100 x_{AGE} + x_{INCOME} - 1$

2.4

Advantage: smaller/shallower tree because less splits must be made as we can combine information.

Disadvantage: Harder to understand because splits are not made orthogonally and may be combined in non-intuitive ways.

Part 3

3.1

2^m inputs for a m length binary input, where m > 0. Consider m = 1. There are 2^1 inputs, as a binary string of length 1 is either 0 or 1. Now consider m = 2. There are 2^2 inputs, as we take each possible string of $2^{2-1} = 2^1$ and push a 0 or a 1 to the front of the string, thus doubling our inputs. Every time the length of the binary string increases by 1, the possible strings double, as every binary string of 2^m can be represented by $2 * 2^{m-1}$ inputs by adding a 0 or 1 to the front of the 2^{m-1} strings.

Specifically for m = 5, 2^5 = **32** possible binary strings.

3.2

The function *f* maps a binary string to either 1 or 0. So from 2^m inputs, there are $2 * 2^m = 2^{m+1}$ possible mappings.

Specifically for m = 5, 2^6 = **64** possible mappings.

3.3

Bob must observe at least $(2^m/2) + 1$ training set pairs where m is the length of the binary strings the function *f* acts upon. Because he observes just above 50% of all possible values, he can guarantee a loss with a constant less than 50% because he has observed the behavior of over 50% of possible values.

Specifically for m = 5, he must observe $(2^5 / 2) + 1$ = **17** distinct training set examples.

3.4

Bob could have split his data into data that he trains his function f hat on and a test set to test how good his function is. He can demonstrate to Janet that his function, which is trained on the training data, classifies the test set in such a way that is non-trivially better than chance. If Bob did choose convince Janet, it is an example of the No Free Lunch Theorem in action: although Bob's function works well in regard to Alice's function, it may not work well for all functions that map a binary string to some binary label.

Part 4

4.5

I believed that as K increased, the distortion would strictly decrease. In the plot, as K increases, the distortion is strictly decreasing, but the distortion decreases by a lower magnitude each time K increases. This confirmed my expectation. This is to be expected, as we increase the number of groups, more alike observations will be grouped into the group that is most appropriate for them.

I believed that as K increased, the mistake rate would decrease. The plot had sharp decreases in mistake rate from K = 1 to K = 2, and K = 2 to K = 3. After that, the mistake rate generally decreased, but there were some cases where the mistake rate increased as K increased. This is surprising, but can be explained by the fact that there might be some noise in the random initialization.