Kshitij Jain 16528151

CPSC 340
Assignment 1

Q1.1

1. 14
2. 0
3. 2
   6
   2
4. 5
5. 6
   5
   7
6. 19
7. 11 10 10
   10 14 10
   10 11 14

Q1.2

1. True
2. True
3. False
4. False
5. False
6. True
7. False
8. True
9. True
10. False

Q2.1

1. 5 $
2. P(B) = .55
3. P(B) = .392

Q2.2

1. .010096
2. False positives
3. .0097
4. No
5. Carry out multiple tests

Q3.1

1. $6x - 2$
2. $1 - 2x$
3. $1 - \frac{e^{-x}}{1+e^{-x}}$

Q3.2

1. $2x + e^{x+2x_2}, 2e^{x+2x_2}$
2. $\frac{e^{x_1}}{e^{x_1}+e^{x_2}+e^{x_3}}, \frac{e^{x_2}}{e^{x_1}+e^{x_2}+e^{x_3}}, \frac{e^{x_3}}{e^{x_1}+e^{x_2}+e^{x_3}}$
3. a
4. $2x_1, 2x_2$
5. $< x_1, x_2, x_3, \ldots . x_d >$

Q3.3

1. $\frac{14}{3}$
2. .25
3. 0,1
4. .5
5. 1
6. (0,0)

Q4.1

1. 6
2. 6

Q4.2

1. O(n)
2. O(log n)
3. O(1)
4. O(n)
5. O(n^2)

Q4.3

1. O(N)
2. O(N)
3. O(N)
4. $O(N^2)$

Q5.1

1. Minimum: 0.35200000000000004
2. Maximum: 4.862
3. Mean: 1.3246250000000002
4. Median: 1.1589999999999998
5. Mode: 0.77
6. 5% quantile: 0.46495000000000003
7. 25% quantile: 0.718
8. 50% quantile: 1.1589999999999998
9. 75% quantile: 1.81325
10. 95% quantile: 2.624049999999999
11. Highest mean: 1.5669615384615392
12. Lowest mean: 1.0632115384615384
13. Highest variance: 0.7834400188609465
14. Lowest variance: 0.3158458206360948

Q5.2

1. D: The option is for a single value histogram which plot D is for frequency vs percentage
2. C: The option is for multi value histogram which is represented by plot C
3. B: The plot represents the various points with its range along with outliers which box plot does
4. A: The plot is like a line graph with multiple points with percentages vs weeks
5. F: There is high cohesion present in the graph as represented from the linear alignment of the points
6. E: There is low cohesion present in the graph as represented from the random alignment of the points

Q6.1

1. I think categorical features would be beneficial for equality-based splitting rule

Q6.2

The updated error is: .265

Q6.3

The updated error is .328

Q6.4

Linked in the readme file

Q6.5

The DecisionStumpErrorRate took 13.786 seconds, DecisionStumpInfoGain took 17.865822 seconds and scikit-learn's decision tree took 0.010649 seconds. Surprisingly infogain took a bit more time compared to the stump error rate because of the various thresholds being compared. The scikit-learn is the most efficient. Also looking at the graph

uptill a point the 3 mechanisms produce the same time but after that point, there is extreme change in the efficiency of the models

Q6.6
Another experiment that could be used to check if the model produces the same curve is by testing both on the unknown dataset, so any overfitting validation is removed.

Q6.7
The cost of the fitting of decision tree of depth m would be $O(ndlogn * m)$ assuming that the cost of transferring from one node to another is constant