Security Analytics

Assignment 2

Student Name: Chandrika Mukherjee
Student ID: 32808289
Email: cmukherj@purdue.edu

Problem 1

- **1. Principal Assumption in Naive Bayes' Model:** Input features are independent, do not affect each other. If the features are $X = (x_1, x_2, x_3, ..., x_n)$, then due to the assumptions, we can write as ,
  $P(y/x_1, x_2, ..., x_n) \propto P(y)P(x_1/y)P(x_2/y)...P(x_n/y)$

  Assumption is useful : When in real-world, features don't depend on each other, then Naive Bayes works better than other classifier. It will converge quicker than other models.

- **2. k-NN do better than Logistic Regression:** When the ML model is non-linear, as Logistic regression support only linear model

- **3. 150 examples in + class, 50 examples in - class**
  So, total examples $= (150 + 50) = 200$
  Entropy of the class $= -[(150/200)log_2(150/200) + (50/200)log_2(50/200)]$ which is equivalent to 0.244

- **4. Number of Requests from Domain A = {2, 3, 4, 3, 4, 3, 3, 4, 5, 3, 2, 3, 4, 3, 2, 2, 3, 4, 5, 6} and Number of Requests from Domain B= {22, 23, 24, 23, 24, 23, 23, 24, 25, 23}**
  total requests from A $[Sum(A)]$ = ((4*2)+(8*3)+(5*4)+(2*5)+(1*6)) = 68
  total requests from B $[Sum(B)]$ = ((1*22)+(5*23)+(3*24)+(1*25)) = 234

  Mean for A $[\mu_A] = Sum(A)/$(total request instances from A) $= 68/20 = 3.4$
  Mean for B $[\mu_B] = Sum(B)/$(total request instances from B) $= 234/10 = 23.4$

  Variance for A $[Var(A)] =$
  $((4 * (2 - 3.4)^2) + (8 * (3 - 3.4)^2) + (5 * (4 - 3.4)^2) + (2 * (5 - 3.4)^2) + (1 * (6 - 3.4)^2))/$ (total request instances from A) $= (7.84 + 1.28 + 1.8 + 5.12 + 6.76)/20 = 1.14$

  Variance for B $[Var(B)] =$
  $((1 * (22 - 23.4)^2) + (5 * (23 - 23.4)^2) + (3 * (24 - 23.4)^2) + (1 * (25 - 23.4)^2))/$ (total request instances from B) $= (1.96 + 0.8 + 1.08 + 2.56)/10 = 0.64$

  Standard Deviation for A $[SD_A] = \sqrt{Var(A)} = \sqrt{1.14} = 1.067$
  Standard Deviation for B $[SD_B] = \sqrt{Var(B)} = \sqrt{0.64} = 0.8$

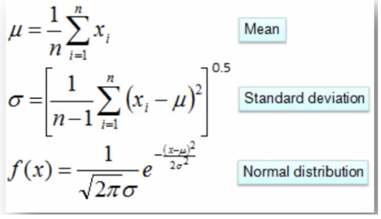
Prior for A $[P(A)/P(A+B)] = 68/(68+234) = 68/302 = 0.2251$
Prior for B $[P(B)/P(A+B)] = 234/(68+234) = 234/302 = 0.7748$

- **5.** Given,

$x^{(1)}$	$x^{(2)}$	y
4	7	0
-4	5	0
2	10	1
10	4	1

$P(y=0) = 2/4 = 0.5$
$P(y=1) = 2/4 = 0.5$

Using the formulas as following,



$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$ Mean

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2\right]^{0.5}$$ Standard deviation

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$ Normal distribution

Table for $x^{(1)}$,

y	$x^{(1)}$	Mean	SD
0	4, -4	$\frac{(4-4)}{2} = 0$	$\frac{\sqrt{(4-0)^2+(-4-0)^2}}{(2-1)} = \sqrt{32}$
1	2, 10	$\frac{(10+2)}{2} = 6$	$\frac{\sqrt{(2-6)^2+(10-6)^2}}{(2-1)} = \sqrt{32}$

Table for $x^{(2)}$,

y	$x^{(2)}$	Mean	SD
0	7, 5	$\frac{(7+5)}{2} = 6$	$\frac{\sqrt{(7-6)^2+(5-6)^2}}{(2-1)} = \sqrt{2}$
1	10, 4	$\frac{(10+4)}{2} = 7$	$\frac{\sqrt{(10-7)^2+(4-7)^2}}{(2-1)} = \sqrt{18}$

According to Naive Bayes Assumption, we can write as below,

$$\begin{aligned}
P(y=0|x) &= P(y=0) * [P(x^{(1)}|y=0) * P(x^{(2)}|y=0)] \\
&= 0.5 * [(\frac{1}{\sqrt{2\pi} * \sqrt{32}} * e^{\frac{-(x-0)^2}{(2*32)}}) * (\frac{1}{\sqrt{2\pi} * \sqrt{32}} * e^{\frac{-(x-6)^2}{(2*32)}})] \\
&= 0.5 * [\frac{1}{64\pi} * e^{\frac{-(x-0)^2}{(64)}} * e^{\frac{-(x-6)^2}{(64)}}] \\
&= 0.5 * [\frac{1}{64\pi} * e^{-\frac{(x^2 + (x-6)^2)}{64}}]
\end{aligned} \tag{1}$$

$$\begin{aligned}
P(y=1|x) &= P(y=1) * [P(x^{(1)}|y=1) * P(x^{(2)}|y=1)] \\
&= 0.5 * [(\frac{1}{\sqrt{2\pi} * \sqrt{2}} * e^{\frac{-(x-6)^2}{(2*2)}}) * (\frac{1}{\sqrt{2\pi} * \sqrt{18}} * e^{\frac{-(x-7)^2}{(2*18)}})] \\
&= 0.5 * [\frac{1}{12\pi} * e^{\frac{-(x-6)^2}{(4)}} * e^{\frac{-(x-7)^2}{(36)}}] \\
&= 0.5 * [\frac{1}{12\pi} * e^{-\frac{(9*(x-6)^2 + (x-7)^2)}{36}}]
\end{aligned} \tag{2}$$

Problem 2

1. Decision on which attribute to consider as Root:

Total instances = 16 and out of these, 9 are “Yes” and 7 are “No”

$$\begin{aligned}
H(S) &= -[\frac{9}{16} log_2(\frac{9}{16}) + \frac{7}{16} log_2(\frac{7}{16})] \\
&= 0.98847
\end{aligned} \tag{3}$$

Now, will be calculating Information gain using Color, Shape and Size.

Information Gain Using Color Attribute:
Total Yellow Instances = 13, Green Instance = 3

Out of 13 Yellow instances, for 8, we get “Yes” output and for 5, we get “No” output.

$$\begin{aligned}
H(S|Yellow) &= -[\frac{8}{13} log_2(\frac{8}{13}) + \frac{5}{13} log_2(\frac{5}{13})] \\
&= 0.96070
\end{aligned} \tag{4}$$

Out of 3 Green instances, for 1, we get “Yes” output and for 2, we get “No” output.

$$\begin{aligned}
H(S|Yellow) &= -[\frac{1}{3} log_2(\frac{1}{3}) + \frac{2}{3} log_2(\frac{2}{3})] \\
&= 0.91784
\end{aligned} \tag{5}$$

Therefore, Information gain using Color Attribute is as below,

$$\begin{aligned}InfoGain(S, Color) &= H(S) - [\frac{13}{16}H(S|Yellow) + \frac{3}{16}H(S|Green)] \\ &= 0.98847 - [\frac{13}{16} * 0.96070 + \frac{3}{16} * 0.91784] \\ &= 0.98847 - 0.952655 \\ &= 0.0358\end{aligned} \tag{6}$$

Information Gain Using Size Attribute:
Total Small Instances = 8, Large Instance = 8

Out of 8 Small instances, for 6, we get "Yes" output and for 2, we get "No" output.

$$\begin{aligned}H(S|Small) &= -[\frac{6}{8}log_2(\frac{6}{8}) + \frac{2}{8}log_2(\frac{2}{8})] \\ &= 0.811214\end{aligned} \tag{7}$$

Out of 8 Large instances, for 3, we get "Yes" output and for 5, we get "No" output.

$$\begin{aligned}H(S|Large) &= -[\frac{3}{8}log_2(\frac{3}{8}) + \frac{5}{8}log_2(\frac{5}{8})] \\ &= 0.95405\end{aligned} \tag{8}$$

Therefore, Information gain using Size Attribute is as below,

$$\begin{aligned}InfoGain(S, Size) &= H(S) - [\frac{8}{16}H(S|Small) + \frac{8}{16}H(S|Large)] \\ &= 0.98847 - [\frac{8}{16} * 0.811214 + \frac{8}{16} * 0.95405] \\ &= 0.98847 - 0.882632 \\ &= 0.105838\end{aligned} \tag{9}$$

Information Gain Using Shape Attribute:
Total Round Instances = 12, Irregular Instance = 4

Out of 12 Round instances, for 6, we get "Yes" output and for 6, we get "No" output.

$$\begin{aligned}H(S|Round) &= -[\frac{6}{12}log_2(\frac{6}{12}) + \frac{6}{12}log_2(\frac{6}{12})] \\ &= 1\end{aligned} \tag{10}$$

Out of 4 Irregular instances, for 3, we get "Yes" output and for 1, we get "No" output.

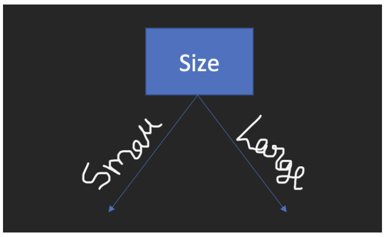
$$\begin{aligned}H(S|Irregular) &= -[\frac{3}{4}log_2(\frac{3}{4}) + \frac{1}{4}log_2(\frac{1}{4})] \\ &= 0.811214\end{aligned} \tag{11}$$

Therefore, Information gain using Shape Attribute is as below,

$$\begin{aligned} InfoGain(S,Shape) &= H(S) - [\frac{8}{16}H(S|Round) + \frac{8}{16}H(S|Irregular)] \\ &= 0.98847 - [\frac{12}{16} * 1 + \frac{4}{16} * 0.811214] \\ &= 0.98847 - 0.9528 \\ &= 0.03567 \end{aligned} \tag{12}$$

So, the information gain will be more if we choose **Size** as the root. (Information gain of Size is 0.105838 which is more than Information gain from Color (0.0358) and Information gain from Shape(0.03567))
Ans. We Choose Size as Root of Decision Tree.

2. Decision Tree Drawing and Corresponding Logic :
Taking the Size as Root (from the above derivation) as the Information gain is more with Root= Size



Splitting with Size attribute gives two results - Small and Large. There are total 8 Small instances. Out of 8, 6 gives “Yes” and 2 gives “No”.

$$\begin{aligned} H(Small) &= -[\frac{6}{8}log_2(\frac{6}{8}) + \frac{2}{8}log_2(\frac{2}{8})] \\ &= 0.811214 \end{aligned} \tag{13}$$

There are total 8 Large instances. Out of 8, 3 gives “Yes” and 5 gives “No”

$$\begin{aligned} H(Large) &= -[\frac{3}{8}log_2(\frac{3}{8}) + \frac{5}{8}log_2(\frac{5}{8})] \\ &= 0.95405 \end{aligned} \tag{14}$$

Calculation of Information Gain with respect to Color on Small Size:
There are total 6 instances of Small Size and Yellow Color. Out of these 6, 5 give “Yes” and 1 gives “No”

$$\begin{aligned} H(Small|Yellow) &= -[\frac{5}{6}log_2(\frac{5}{6}) + \frac{1}{6}log_2(\frac{1}{6})] \\ &= 0.6497 \end{aligned} \tag{15}$$

There are total 2 instances of Small Size and Green Color. Out of these 2, 1 gives “Yes” and 1 gives “No”

$$\begin{aligned} H(Small|Green) &= -[\frac{1}{2}log_2(\frac{1}{2}) + \frac{1}{2}log_2(\frac{1}{2})] \\ &= 1 \end{aligned} \tag{16}$$

Total Small size instances are 8, out of which 6 are Small and Yellow, other 2 are Small and Green

$$\begin{aligned} InfoGain(Small, Color) &= H(Small) - [\frac{6}{8}H(Small|Yellow) + \frac{2}{8}H(Small|Green)] \\ &= 0.811214 - [\frac{6}{8} * 0.6497 + \frac{2}{8} * 1] \\ &= 0.811214 - 0.7372 \\ &= 0.074014 \end{aligned} \tag{17}$$

Calculation of Information Gain with respect to Shape on Small Size:

There are total 6 instances of Small Size and Round Shape. Out of these 6, 4 give "Yes" and 2 give "No"

$$\begin{aligned} H(Small|Round) &= -[\frac{4}{6}log_2(\frac{4}{6}) + \frac{2}{6}log_2(\frac{2}{6})] \\ &= 0.91828 \end{aligned} \tag{18}$$

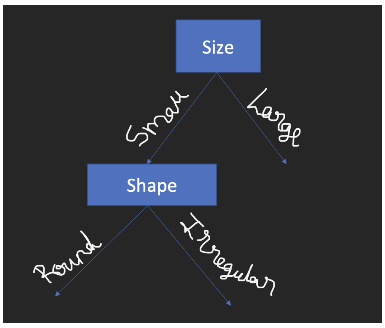
There are total 2 instances of Small Size and Irregular Shape. 2 of them give "Yes".

$$\begin{aligned} H(Small|Irregular) &= -[\frac{2}{2}log_2(\frac{2}{2})] \\ &= 0 \end{aligned} \tag{19}$$

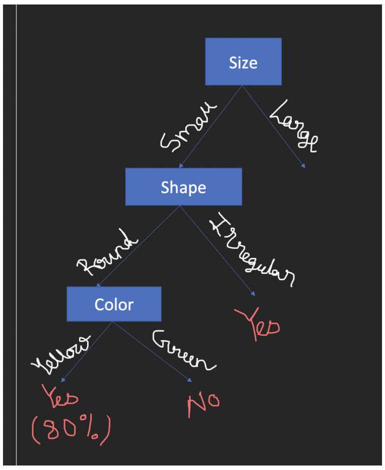
Total small size instances are 8, out of which 6 are Small and Round, other 2 are Small and Irregular

$$\begin{aligned} InfoGain(Small, Shape) &= H(Small) - [\frac{6}{8}H(Small|Round) + \frac{2}{8}H(Small|Irregular)] \\ &= 0.811214 - [\frac{6}{8} * 0.91828 + \frac{2}{8} * 0] \\ &= 0.811214 - 0.68871 \\ &= 0.122504 \end{aligned} \tag{20}$$

Therefore, after splitting using Size, in the Small branch, we should split using Shape as Information Gain by Shape is 0.122504 that is higher than Information Gain by Color (0.074014).



Now, After Splitting Small branch with Shape, Consequent Irregular branch will have only one outcome ("Yes"), Entropy is also Zero. But the Entropy of Round branch is more than zero (0.91828). Therefore, Need to split Round Branch with Color. After Splitting with Color, Green Branch has only one outcome "No" as in the training data, only one instance is present with {Size=Small, Shape= Round, Color= Green} which has output of "No". But there are total 5 instances of {Size=Small, Shape= Round, Color= Yellow}, out of these 5, 4 give "Yes" as output and 1 gives "No" as output. As we don't have any other feature, if we consider majority, Yellow branch will give "Yes" as output.



Calculation of Information Gain with respect to Color on Large Size:

There are total 7 instances of Large Size and Yellow Color. Out of these 7, 3 give "Yes" and 4 gives "No"

$$\begin{aligned} H(Large|Yellow) &= -[\frac{3}{7}log_2(\frac{3}{7}) + \frac{4}{7}log_2(\frac{4}{7})] \\ &= 0.98518 \end{aligned} \tag{21}$$

There are total 1 instance of Large Size and Green Color and it gives output as "No"

$$\begin{aligned} H(Large|Green) &= -[\frac{1}{1}log_2(\frac{1}{1})] \\ &= 0 \end{aligned} \tag{22}$$

Total Large size instances are 8, out of which 7 are Large and Yellow, other 1 is Large and Green

$$\begin{aligned} InfoGain(Large, Color) &= H(Large) - [\frac{7}{8}H(Large|Yellow) + \frac{1}{8}H(Large|Green)] \\ &= 0.95405 - [\frac{7}{8} * 0.98518 + \frac{1}{8} * 0] \\ &= 0.95405 - 0.86203 \\ &= 0.09202 \end{aligned} \tag{23}$$

Calculation of Information Gain with respect to Shape on Large Size:

There are total 6 instances of Large Size and Round Shape. Out of these 6, 2 give “Yes” and 4 give “No”

$$\begin{aligned} H(Large|Round) &= -[\frac{2}{6}log_2(\frac{2}{6}) + \frac{4}{6}log_2(\frac{4}{6})] \\ &= 0.91828 \end{aligned} \tag{24}$$

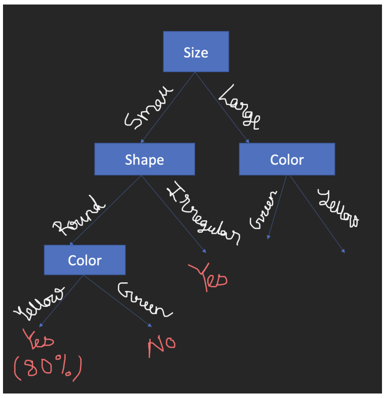
There are total 2 instances of Large Size and Irregular Shape. Out of 2, 1 gives “Yes” and 1 gives “No”.

$$\begin{aligned} H(Large|Irregular) &= -[\frac{1}{2}log_2(\frac{1}{2}) + \frac{1}{2}log_2(\frac{1}{2})] \\ &= 1 \end{aligned} \tag{25}$$

Total large size instances are 8, out of which 6 are Large and Round, other 2 are Large and Irregular

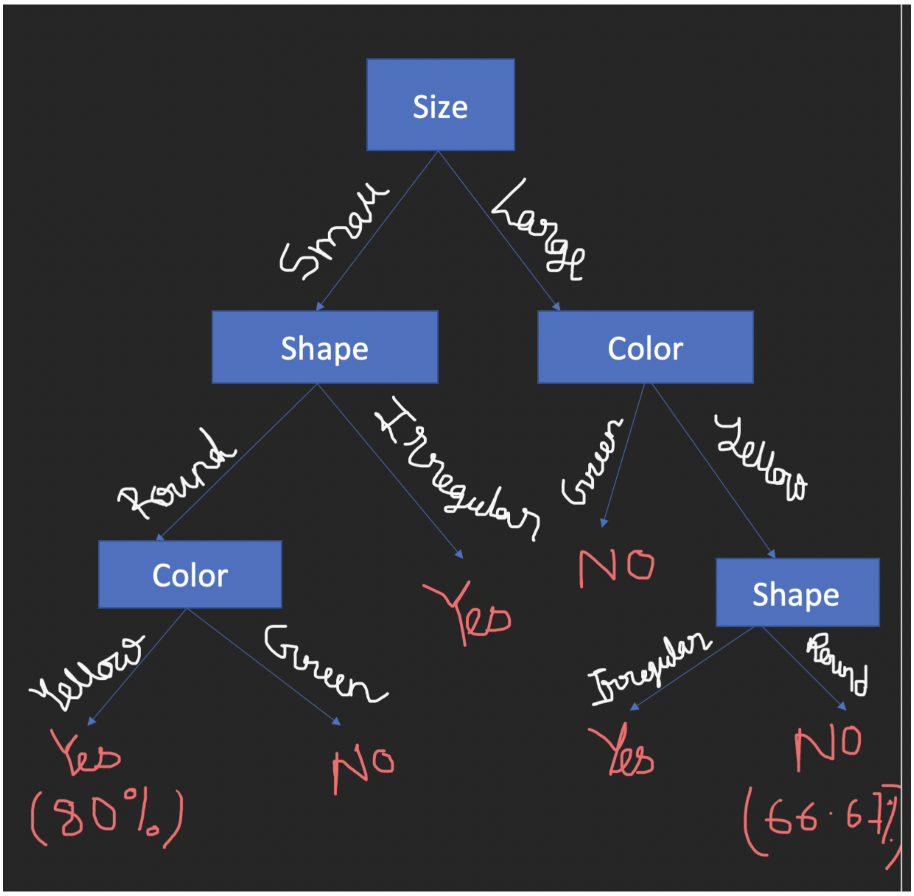
$$\begin{aligned} InfoGain(Large, Shape) &= H(Large) - [\frac{6}{8}H(Large|Round) + \frac{2}{8}H(Large|Irregular)] \\ &= 0.95405 - [\frac{6}{8} * 0.91828 + \frac{2}{8} * 1] \\ &= 0.95405 - 0.93871 \\ &= 0.01534 \end{aligned} \tag{26}$$

Therefore, after splitting using Size, in the Large branch, we should split using Color as Information Gain by Color is 0.09202 that is higher than Information Gain by Shape (0.01534).



Now, After Splitting Large branch with Color, Consequent Green branch will have only one outcome (“No”), Entropy is also Zero. But the Entropy of Yellow branch is more than zero (0.98518). Therefore, Need to split Yellow Branch with Shape. After Splitting with Shape, Irregular Branch has only one outcome “Yes” as in the training data, only one instance is present with {Size=Large, Color= Yellow, Shape= Irregular } which has output of “Yes”. But there are total 6 instances of {Size=Large, Color= Yellow, Shape= Round}, out of these 6, 2 give “Yes” as output and 4 give “No” as output. As we don’t have any other feature, if we consider majority, Round branch will give “No” as output.

This is the final Decision Tree.



3. If Features have Numerical Values and those are treated as Discrete, and we proceed as Categorical Decision Tree Algorithm :

When the decision tree is used to classify unseen data, problems will arise.
New unseen data may not be present in the sample training data set. So, the branch corresponding to the unseen value will not be present in the decision tree, therefore, from that point, decision process will stop.
For example, In the given data set of edible mushroom, if there was another feature "Height" which had numerical values. If we treated each value as discrete, then we could build the decision tree. Suppose,The heights present in the data were = {10,12,20,21,24,11,56,34,21,22,10,12,13,24,55,22}. Now, new unseen data has height = 28, but we don't have any branch corresponding to height 28, then we can't conclude on the output. Even if unseen data is present in training sample, treating each numerical value as discrete, make the decision tree more complex.