Assignment #: 1

Venkata Sai Chelagamsetty

September 14th, 2022

1. (a) Naive Bayes assumes independence between features i.e the presence of a particular feature in a class is unrelated to the presence of any other feature.Naive Bayes is suitable for solving multi-class prediction problems. If its assumption of the independence of features holds true, it can perform better than other models and requires much less training data.

   (b) KNN is better than logistic regression when boundary is non linear during classifying. KNN supports non-linear solutions where logistic regression supports only linear solutions.

   (c) The entropy for the set is as follows:

$$E(x) = \frac{1}{4}\log 4 + \frac{3}{4}\log\frac{4}{3}$$

   (d) The mean and the standard deviation for the classes is as follows:
   Class A: Mean = 3.4, Standard Deviation = 1.094
   Class B: Mean = 23.4, Standard Deviation = 0.8433
   Note that the standard deviation is calculated using n-1 in the denominator.
   And the priors are as follows:
   Class A: Prior = $\frac{20}{30}$ Class B: Prior = $\frac{10}{30}$

   (e) When $y = 0$:

Variable	**Values**	**Mean**	**Std. Dev**
$x^{(1)}$	4, -4	0	5.656
$x^{(2)}$	7, 5	6	1.414

   When $y = 1$:

Variable	**Values**	**Mean**	**Std. Dev**
$x^{(1)}$	2, 10	6	5.656
$x^{(2)}$	10, 4	7	4.242

   Say that the vector x can take values from the following set (x_1, x_2). Then,

$$P(x = x_1|y = 0) = \frac{1}{\sqrt{2\pi * 5.656}} e^{-\frac{(a_1-0)^2}{2(5.656)^2}}$$

$$P(x = x_2|y = 0) = \frac{1}{\sqrt{2\pi * 1.414}} e^{-\frac{(a_2-6)^2}{2(1.414)^2}}$$

$$P(x = x_1|y = 1) = \frac{1}{\sqrt{2\pi * 5.656}} e^{-\frac{(a_1-6)^2}{2(5.656)^2}}$$

$$P(x = x_2|y = 1) = \frac{1}{\sqrt{2\pi * 4.242}} e^{-\frac{(a_2-7)^2}{2(4.242)^2}}$$

Therefore,

$$P(y=0|x) = \frac{1}{\sqrt{2\pi}*5.656} e^{-\frac{(x_1-0)^2}{2(5.656)^2}} . \frac{1}{\sqrt{2\pi}*1.414} e^{-\frac{(x_2-6)^2}{2(1.414)^2}}$$

$$P(y=1|x) = \frac{1}{\sqrt{2\pi}*5.656} e^{-\frac{(x_1-6)^2}{2(5.656)^2}} . \frac{1}{\sqrt{2\pi}*4.242} e^{-\frac{(x_2-7)^2}{2(4.242)^2}}$$

2. (a) Let us see which one gives us the highest information gain in the first step. So there are three attributes. Each on of them has only 2 values. So the root of the tree will be split in two based upon the feature values. Let's take the attribute color to be the one that algorithm chooses. Then we will have the following split:

Entropy without any information:

$$E = \frac{7}{16}\log\frac{16}{7} + \frac{9}{16}\log\frac{16}{9} = 0.359$$

Now the entropy in the right leaf(Green):

$$E_r = \frac{2}{3}\log\frac{3}{2} + \frac{1}{3}\log\frac{3}{2} = 0.176$$

Now in the left(Yellow):

$$E_l = \frac{5}{13}\log\frac{13}{5} + \frac{8}{13}\log\frac{13}{8} = 0.289$$

Average of both:

$$E_t = \frac{3}{16}E_r + \frac{13}{16}E_l = 0.268$$

Now, Let's take the attribute size to be the one that algorithm chooses. Then we will have the following split:

Entropy without any information:

$$E = \frac{7}{16}\log\frac{16}{7} + \frac{9}{16}\log\frac{16}{9} = 0.359$$

Now the entropy in the right leaf(Small):

$$E_r = \frac{6}{8}\log\frac{8}{6} + \frac{2}{8}\log\frac{8}{2} = 0.244$$

Now in the left(Large):

$$E_l = \frac{3}{8}\log\frac{8}{3} + \frac{5}{8}\log\frac{8}{5} = 0.287$$

Average of both:

$$E_t = \frac{1}{2}E_r + \frac{1}{2}E_l = 0.265$$

Now, Let's take the attribute shaper to be the one that algorithm chooses. Then we will have the following split:

Entropy without any information:

$$E = \frac{7}{16}\log\frac{16}{7} + \frac{9}{16}\log\frac{16}{9} = 0.359$$

Now the entropy in the right leaf(Irregular):

$$E_r = \frac{1}{4}\log\frac{4}{1} + \frac{3}{4}\log\frac{4}{3} = 0.244$$

Now in the left(Regular):

$$E_l = \frac{6}{12}\log\frac{12}{6} + \frac{6}{12}\log\frac{12}{6} = 0.301$$

Average of both:

$$E_t = \frac{4}{16}E_r + \frac{12}{16}E_l = 0.286$$

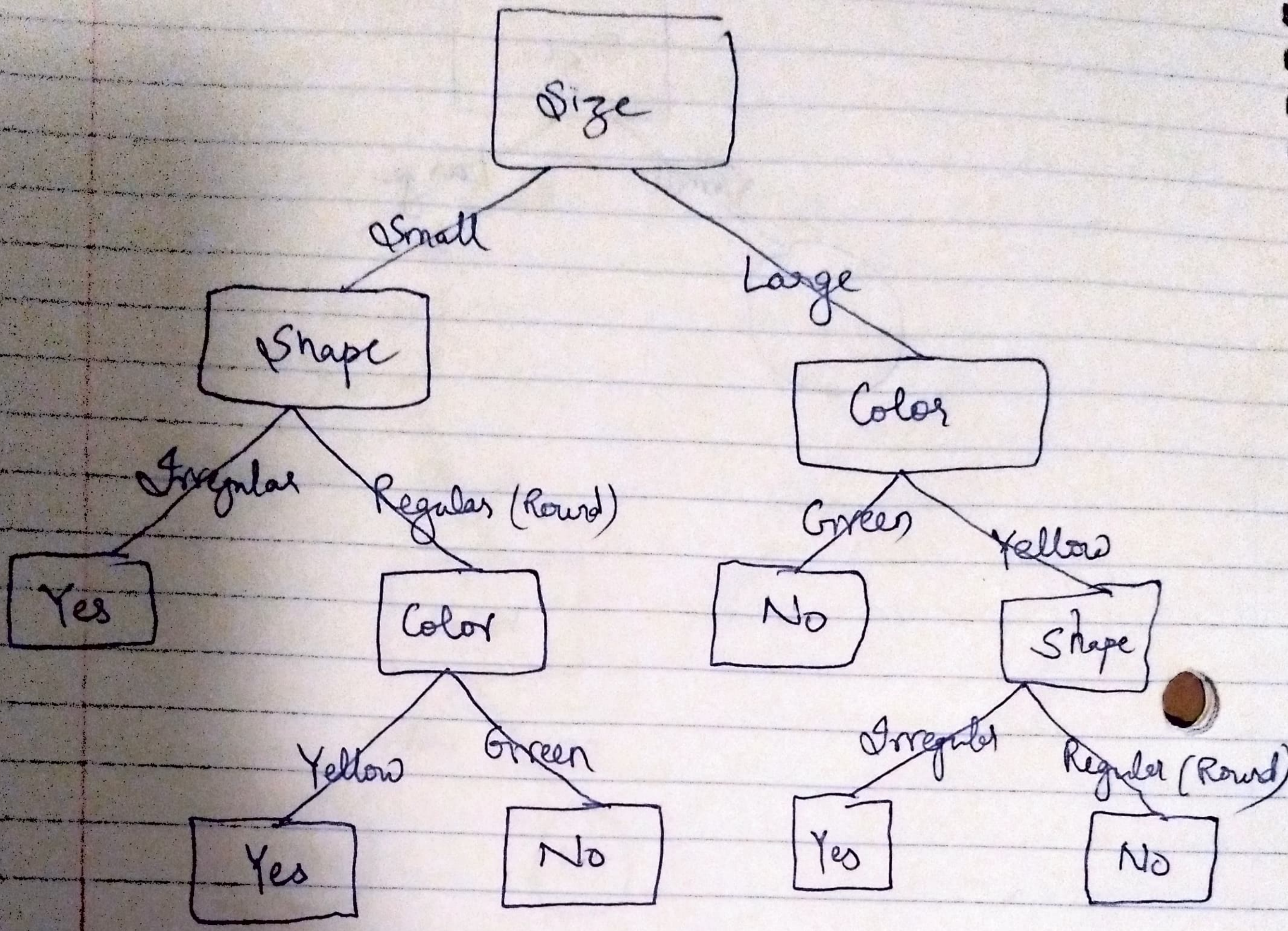
Therefore we will be choosing the size attribute. (As the inital entropy is same for all, we will choose the function with minimum avearge entropy after division)

(b) Having done the above calculations, in a similar way we proceed with other attributes and arrive at the following decision tree.

(c) Yes, some problems do arise. If each value is treated as a discrete value, we will have a very large tree. This results in overfitting of the data. We aren't able to set any thresholds. If we go to the testing phase, only if the incoming value is out of one of the training samples, we can predict correctly. Without generalization, we aren't able to predict the correct outputs leading to high test error.

(d) Code

(e) No standardization done.


Size
Small
Large
Shape
Color
Irregular
Regular (Round)
Green
Yellow
Yes
Color
No
Shape
Yellow
Green
Irregular
Regular (Round)
Yes
No
Yes
No