CS 529 Assignment 2

Mohit Mayank – mayank@purdue.edu

October 13, 2022

PUID: 033744160

1

1.
   - assumption: features of the class are independent of each other
   - Naive-Bayes is useful when this assumption of independence and equality holds true where this model could outperform other models.

2. KNN could be better than Logistic Regression when we sufficient domain knowledge about the problem at hand since KNN depends on distance measure, as this would support derivation of an appropriate measure. Also, KNN would perform better for models with less number of parameters and dataset size.

3. $$Entropy = -\frac{150}{200} \cdot \log \frac{150}{200} - \frac{50}{200} \cdot \log \frac{50}{200}$$

4.

Mean of A	3.4
Mean of B	23.4
Std. Dev. of A	1.067
Std. Dev. of B	0.8
Prior of A	$20/30 = 0.67$
Prior of B	$10/30 = 0.33$

5. Given:

$$P(y = 0) = \frac{1}{2}$$
$$P(y = 1) = \frac{1}{2}$$

From Gaussian distribution:

$$P(x^{(t)}|y=y_i) = \frac{1}{\sigma_t\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-\mu_t)^2}{\sigma_t^2}}$$

For $x^{(1)}$:

$$\begin{aligned}\mu(x^{(1)}|y=1) &= 6\\ \sigma(x^{(1)}|y=1) &= 5.65\\ \mu(x^{(1)}|y=0) &= 0\\ \sigma(x^{(1)}|y=0) &= 5.65\end{aligned}$$

$$P(x^{(1)}|y=1) = \frac{1}{5.65\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-6)^2}{5.65^2}}$$
$$P(x^{(1)}|y=0) = \frac{1}{5.65\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x)^2}{5.65^2}}$$

For $x^{(1)}$:

$$\begin{aligned}\mu(x^{2}|y=1) &= 7\\ \sigma(x^{2}|y=1) &= 4.24\\ \mu(x^{2}|y=0) &= 6\\ \sigma(x^{2}|y=0) &= 1.41\end{aligned}$$

$$P(x^{(2)}|y=1) = \frac{1}{4.24\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-7)^2}{4.24^2}}$$
$$P(x^{(2)}|y=0) = \frac{1}{1.41\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-6)^2}{1.41^2}}$$

Using Naive Bayes for $y=0$:

$$P(y=0|x) = P(x^{(1)}|y=0) * P(x^{(2)}|y=0)$$
$$P(y=0|x) = \frac{1}{5.65\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x)^2}{5.65^2}} * \frac{1}{1.41\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-6)^2}{1.41^2}}$$

Similarly for $y = 1$:

$$P(y = 1|x) = P(x^{(1)}|y = 1) * P(x^{(2)}|y = 1)$$

$$P(y = 1|x) = \frac{1}{5.65\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-6)^2}{5.65^2}} * \frac{1}{4.24\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-7)^2}{4.24^2}}$$

From above:

$$P(y = 0|x).P(y = 1|x) = \frac{1}{5.65\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x)^2}{5.65^2}} * \frac{1}{1.41\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-6)^2}{1.41^2}}$$
$$* \frac{1}{5.65\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-6)^2}{5.65^2}} * \frac{1}{4.24\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-7)^2}{4.24^2}}$$

2

1. Information gain calculation:

$$
\begin{aligned}
\text{InfoGain}(S, A) &= H(S) - \sum_{v \in V} \left(\frac{|S_v|}{|S|} \right) H(S_v) \\
H(S) &= \frac{-9}{16} \ln \left(\frac{9}{16} \right) - \frac{7}{16} \ln \left(\frac{7}{16} \right) \\
&= 0.685314 \\
\text{InfoGain(S,"Color")} &= H(S) - \left(\frac{13}{16} \left(\frac{-8}{13} \ln \left(\frac{8}{13} \right) - \frac{5}{13} \ln \left(\frac{5}{13} \right) \right) + \frac{3}{16} \left(\frac{-1}{3} \ln \left(\frac{1}{3} \right) - \frac{2}{3} \ln \left(\frac{2}{3} \right) \right) \right) \\
&= 0.02461 \\
\text{InfoGain(S,"Size")} &= H(S) - \left(\frac{8}{16} \left(\frac{-6}{8} \ln \left(\frac{6}{8} \right) - \frac{2}{8} \ln \left(\frac{2}{8} \right) \right) + \frac{8}{16} \left(\frac{-3}{8} \ln \left(\frac{3}{8} \right) - \frac{5}{8} \ln \left(\frac{5}{8} \right) \right) \right) \\
&= 0.07336 \\
\text{InfoGain(S,"Shape")} &= H(S) - \left(\frac{12}{16} \left(\frac{-6}{12} \ln \left(\frac{6}{12} \right) - \frac{6}{12} \ln \left(\frac{6}{12} \right) \right) + \frac{4}{16} \left(\frac{-3}{4} \ln \left(\frac{3}{4} \right) - \frac{1}{4} \ln \left(\frac{1}{4} \right) \right) \right) \\
&= 0.02461
\end{aligned}
$$

   Since **size** has the highest information gain, we choose it as the root of the tree.

2. Below is the Decision Tree diagram:

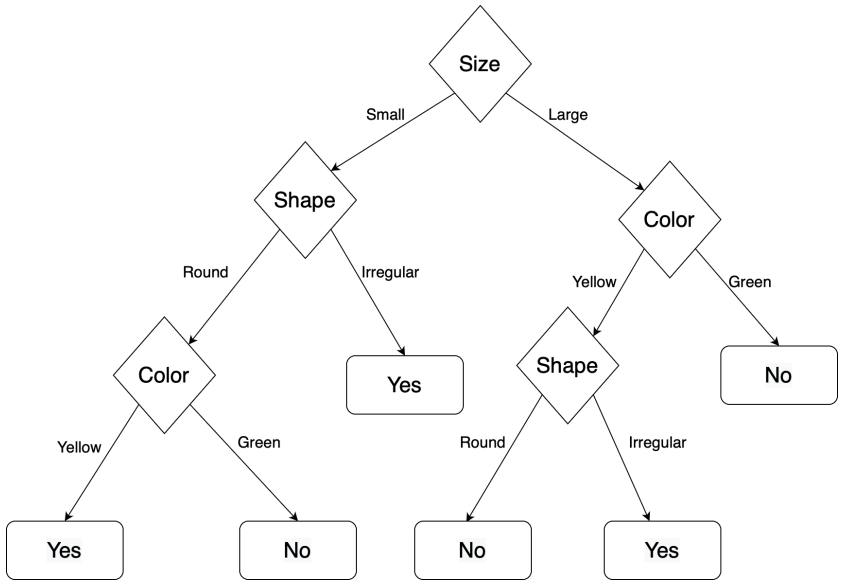

Figure 1: Decision Tree

3. If we use numerical features as separate categories, we would be creating too many categories (this would be even more huge in case of real valued features). With such features, it is highly likely that test data would contain unique values (eg: train data might contain values: 3.1, 3.112, 3.3, 4.55 and test data might contain some value like 3.35), in such cases the decision tree would fail to predict.

   We need to create bins (or ranges) to solve this problem if we really need to use Decision Trees for such a problem.