Name- Sayem Lincoln HW7

PID - A54207835

1) From Bayes Naive Classifier, we evaluate $P(a_i|v_j)$ and it is given by ⇒

$P(a_i|v_j) = \frac{n_c + mp}{n+m}$. Here, m = equivalent sample size.

n_c = number of examples for which $v = v_j$, $va = a_i$

P = prior estimate for $P(a_i|v_j)$

n = number of training examples for which $v = v_j$

Here, first we have to calculate ⇒

$P(Suv|Yes)$, $P(Red|Yes)$, $P(Domestic|yes)$, $P(Suv|No)$, $P(Red|No)$

$P(Domestic|No)$

Evaluating these values ⇒

No ⇒

SUV ⇒ $P = 0.5$, $n_c = 3$, $m = 3$, $n = 5$

Red ⇒ $P = 0.5$, $n_c = 2$, $m = 3$, $n = 5$

Domestic ⇒ $P = 0.5$, $n_c = 3$, $m = 3$, $n = 5$

yes ⇒

SUV ⇒ $P = 0.5$, $n_c = 1$, $m = 3$, $n = 5$

Red ⇒ $P = 0.5$, $n_c = 3$, $m = 3$, $n = 5$

Domestic ⇒ $P = 0.5$, $n_c = 2$, $m = 3$, $n = 5$

Here using Bayes Naives classifier to calculate the respective probabilities

⇒ $P(Suv|Yes) = \frac{n_c+mp}{n+m} = \frac{1 + 3\times0.5}{5+3} = 0.31$

$P(Red|Yes) = \frac{n_c+mp}{n+m} = \frac{3+3\times0.5}{5+3} = 0.56$

$P(Domestic|Yes) = \frac{n_c+mp}{n+m} = \frac{2+3\times0.5}{3+5} = 0.44$

$P(SUV|NO) = \frac{n_c+mp}{n+m} = \frac{3+3\times0.5}{5+3} = 0.56$

$P(Red|No) = \frac{n_c+mp}{n+m} = \frac{2+3\times0.5}{5+3} = 0.44$

$P(Domestic|NO) = \frac{n_c+mp}{n+m} = \frac{3+3\times0.5}{5+3} = 0.56$

Then we have ⇒ $P(yes) = 0.5$ $P(NO) = 0.5$

Now for the stolen (Yes) = $P(yes) \times P(Red|yes) \times P(suv|Yes) \times P(Domestic|yes)$

$= 0.5 \times 0.56 \times 0.31 \times 0.44$

$= 0.038$

Now for stolen (NO) = $P(NO) \times P(Red|NO) \times P(SUV|NO) \times P(Domestic|NO)$

$= 0.5 \times 0.44 \times 0.56 \times 0.56$

$= 0.069$

Since $0.069 > 0.038$, therefore the Red Domestic SUV gets classified as "No".

2) a) Naives Bayes is called navie because it makes the naive assumption that features a sense of zero correlation with each other, implying independency. By doing so, the joint distilbution can be found easily just by multiplying the probability of each feature whilst in the real world they may not be independent and you have to find the correct joint distribution, hence it is ~~exact~~ naive due to this simplification process.

b) The numbers of parameters that are to be estimated are of infinite numbers, if each possible datapoint has a different possibility then you have an exponential number of possibilities, Since there an exponential number of different data points. So, the parameters are infinite, and by Naives Bayes method the parameters can be estimated and then the independeny of these parameters correct joint distribution be figured out ~~from~~ for the data.

3) $P(\text{physical} \mid \text{atom} \mid \text{carbon} \mid \text{life} \mid \text{earth})$

$= \dfrac{P(\text{atom}, \text{carbon}, \text{life}, \text{earth})\ p(\text{physics})}{P(\text{atom}, \text{carbon}, \text{life}, \text{earth})}$

$= \dfrac{P(\text{atom}|\text{physics})\ P(\text{carbon}|\text{physics})\ p(\text{life}|\text{physics})\ P(\text{earth}|\text{physics})\ P(\text{physics})}{\sum p(\text{atom}|c)\ p(\text{carbon}|c)\ p(\text{earth}|c)\ p(\text{life}|c)}$

$= 0.00657$

$P(\text{Biology} \mid \text{atom}, \text{carbon}, \text{life}, \text{earth}) = 0.542$

$P(\text{chemistry} \mid \text{atom}, \text{carbon}, \text{life}, \text{earth}) = 0.452$

Biology is the most category $\Rightarrow$

$P(\text{physics} \mid \text{atom}, \text{carbon}, \text{proton})$

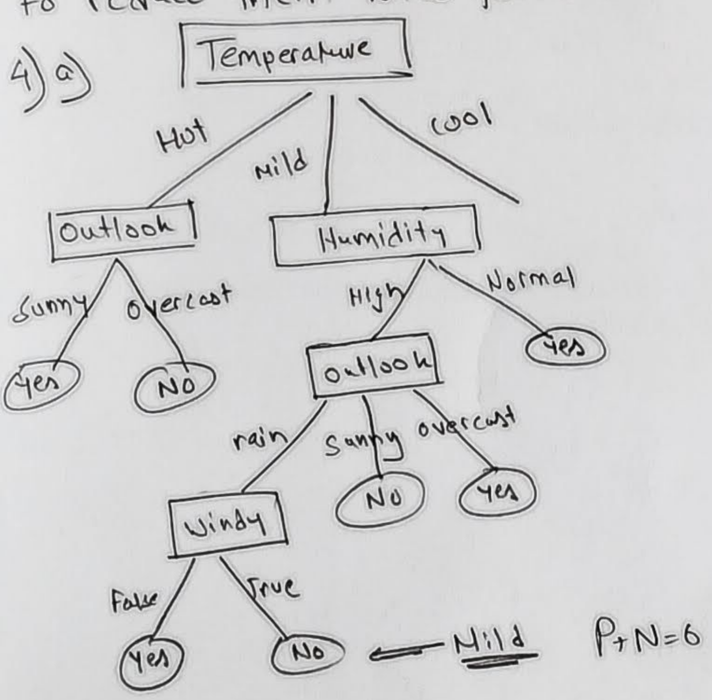
$= \dfrac{P(\text{atom}, \text{carbon}, \text{proton} \mid \text{physics})\ P(\text{physics})}{P(\text{atom}, \text{carbon}, \text{proton})} = 0.0654$

$P(\text{Biology} \mid \text{atom}, \text{carbon}, \text{proton}) = 0.00082$

$P(\text{chemistry} \mid \text{atom}, \text{carbon}, \text{proton}) = 0.94$

Chemistry is the most likely useful, when the words are first stemmed to reduce their base form.

4) a)



Temperature = hot

Outlook	humidity	windy	Play Golf
Sunny	high	False	N
Sunny	high	true	N
Overcast	high	true	Y
Overcast	Normal	true	Y

Temperature = Mild

Outlook	humidity	Windy	Play Golf
rain	high	False	Y
sunny	high	false	N
rain	normal	false	Y
sunny	normal	True	Y
overcast	high	True	Y
rain	high	True	N

$I(4,2) = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}$

$= 0.9114$

Rain (outlook) = $I(4,2) - E(O)$

Outlook	Play	No plays
rain	2	1
sunny	1	1
overcast	1	0

$E(O) = \sum_{i=1}^{v} p_i + n_i \; I(p_i + n_i)$

$= \frac{3}{6} I(2,1) + \frac{2}{6} I(1,1) + \frac{1}{6} I(1,0) = 0.77$

Gain (outlook) = 0.1414

Similarly ⇒ Gain (Humidity) = 0.2414 (maximum)

Gain (windy) = 0.0014

Humidity = high

outlook	windy	Play
rain	false	Y
sunny	false	N
overcast	True	Y
rain	True	N

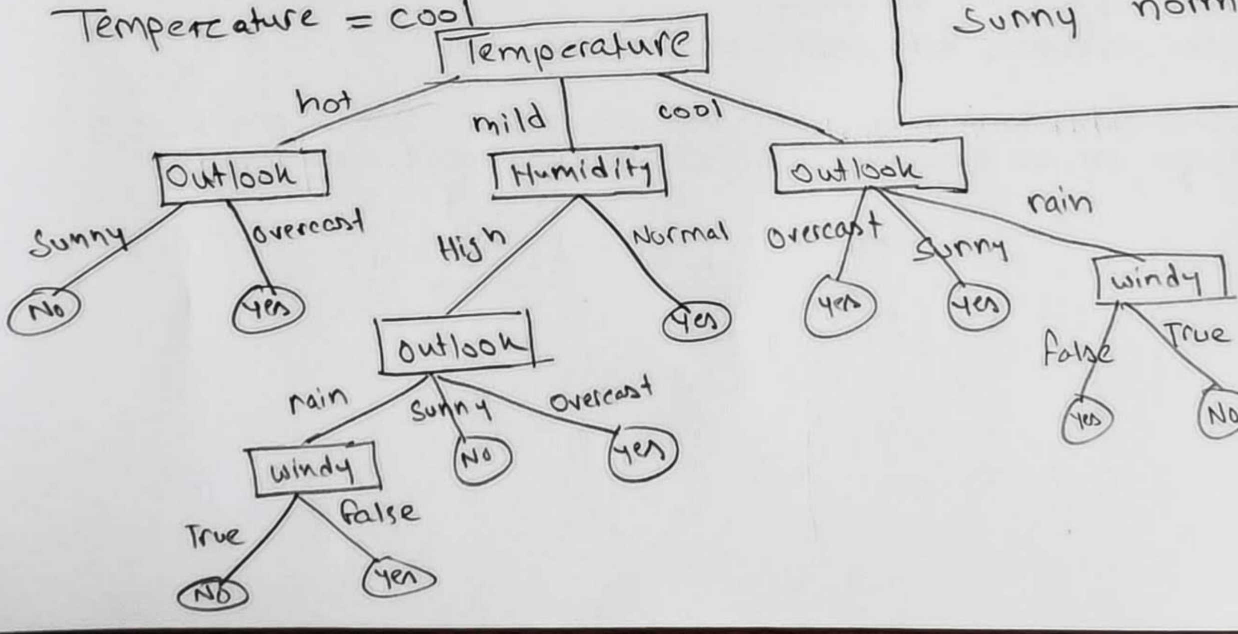
Outlook	humidity	windy	Play
rain	normal	false	Y
rain	normal	True	N
overcast	normal	True	Y
sunny	normal	false	Y

$I(2,2) = -\frac{2}{4}\log_2\frac{2}{4} \log_2\frac{2}{4} = 1$

Gain (outlook) = 0.5 ⟵ maximum

Gain (windy) = 0

Temperature = cool



b) Choose root node

$I(9,5) = 0.9401$

$E(\text{outlook}) = 0.69$

- Gain (outlook) = $0.9401 - 0.69 = 0.2501$
- Gain (temperature) = $0.9401 - 0.878 = 0.0621$
- Gain (humidity) = $0.9401 - 0.7809 = 0.1592$
- Gain (windy) = $0.9401 - 0.09 = 0.0501$

where, $E(\text{outlook}) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} (3,2)$

$E(\text{temperature}) = \frac{4}{14} I(2,2) + \frac{6}{14} I(4,2) + \frac{4}{14} I(3,1)$

$E(\text{humidity}) = \frac{7}{14} I(3,4) + \frac{7}{14} I(6,1)$

$E(\text{windy}) = \frac{8}{14} I(6,2) + \frac{6}{14} I(3,3)$

Tree