Vrund Patel

1 Theory

1.

(a)

$$H(Y) = H(P(y=0), P(y=1)) = -P(y=0) * \log_2 P(y=0) + -P(y=1) * \log_2 P(y=1)$$

$$= \left(-\frac{9}{21}\right) * \log_2 \frac{9}{21} + \left(-\frac{12}{21}\right) * \log_2 \frac{12}{21}$$

$$= .98522$$

(b)

$$IG(A) = H\left(\frac{9}{21}, \frac{12}{21}\right) - \mathbb{E}(H(A))$$

Feature $\boldsymbol{x_1} = \{\mathbf{0}, \mathbf{1}\} = \{\mathbf{F}, \mathbf{T}\}$

$$IG(A) = H\left(\frac{12}{21}, \frac{9}{21}\right) - \left(\frac{p_0 + n_0}{p + n} H\left(\frac{p_0}{p_0 + n_0}, \frac{n_0}{p_0 + n_0}\right) + \frac{p_1 + n_1}{p + n} H\left(\frac{p_1}{p_1 + n_1}, \frac{n_1}{p_1 + n_1}\right)\right)$$

$$IG(A) = H\left(\frac{12}{21}, \frac{9}{21}\right) - \left(\frac{5 + 8}{12 + 9} H\left(\frac{5}{5 + 8}, \frac{8}{5 + 8}\right) + \frac{7 + 1}{12 + 9} H\left(\frac{7}{7 + 1}, \frac{1}{7 + 1}\right)\right)$$

$$IG(A) = 0.98522 - \left(\frac{13}{21} H\left(\frac{5}{5 + 8}, \frac{8}{5 + 8}\right) + \frac{8}{21} H\left(\frac{7}{7 + 1}, \frac{1}{7 + 1}\right)\right)$$

$$IG(A) = 0.9852 - \left(\frac{13}{21}\left(-\frac{5}{13} * \log_2\left(\frac{5}{13}\right) + \left(-\frac{8}{13} * \log_2\left(\frac{8}{13}\right)\right)\right) + \frac{8}{21} H\left(\frac{7}{7 + 1}, \frac{1}{7 + 1}\right)\right)$$

$$IG(A) = 0.9852 - \left(0.5951 + \frac{8}{21}\left(-\frac{7}{8} * \log_2\left(\frac{7}{8}\right) + \left(-\frac{1}{8} * \log_2\left(\frac{1}{8}\right)\right)\right)\right)$$

$$\boldsymbol{IG(A) = 0.183}$$

Feature $\boldsymbol{x_2} = \{\mathbf{0}, \mathbf{1}\} = \{\mathbf{F}, \mathbf{T}\}$

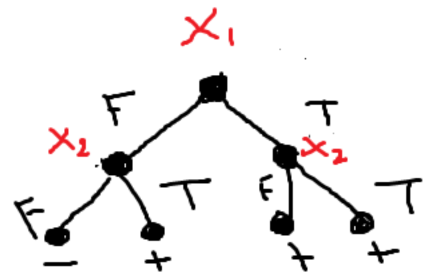
$$IG(A) = H\left(\frac{12}{21}, \frac{9}{21}\right) - \left(\frac{p_0 + n_0}{p + n} H\left(\frac{p_0}{p_0 + n_0}, \frac{n_0}{p_0 + n_0}\right) + \frac{p_1 + n_1}{p + n} H\left(\frac{p_1}{p_1 + n_1}, \frac{n_1}{p_1 + n_1}\right)\right)$$

$$IG(A) = 0.9852 - \left(\frac{5 + 6}{12 + 9} H\left(\frac{5}{5 + 6}, \frac{6}{5 + 6}\right) + \frac{7 + 3}{12 + 9} H\left(\frac{7}{7 + 3}, \frac{3}{7 + 3}\right)\right)$$

$$IG(A) = 0.9852 - \left(\frac{11}{21}\left(-\frac{5}{11} * \log_2\left(\frac{5}{11}\right) + \left(-\frac{6}{11} * \log_2\left(\frac{6}{11}\right)\right)\right) + \frac{10}{21}\left(-\frac{7}{10} * \log_2\left(\frac{7}{10}\right) + \left(-\frac{3}{10} * \log_2\left(\frac{3}{10}\right)\right)\right)\right)$$

$$\boldsymbol{IG(A) = 0.0449}$$

(c)



2.

(a)

$$P(A = Yes) = \frac{3}{5} = 0.6$$

$$P(A = No) = \frac{2}{5} = 0.4$$

(b)

$$\textbf{Chars mean} = \frac{216+69+302+60+393}{5} = \frac{1040}{5} = 208$$

$$\textbf{Chars standard deviation} = \sqrt{\frac{(216-208)^2+(69-208)^2+(302-208)^2+(60-208)^2+(393-208)^2}{5-1}} =$$

$$= \sqrt{\frac{84350}{4}} = 145.21$$

$$\textbf{Word Length mean} = \frac{5.68+4.78+2.31+3.16+4.2}{5} = 4.03$$

$$\textbf{Word Length standard deviation} = \sqrt{\frac{(5.68-4.03)^2+(4.78-4.03)^2+(2.31-4.03)^2+(3.16-4.03)^2+(4.2-4.03)^2}{5-1}}$$

$$= 1.33$$

# of Chars Standardized	Average Word Length Standardized	Give an A

$\frac{216-208}{145.21}=0.055$	$\frac{5.68-4.03}{1.33}=1.248$	Yes
$\frac{69-208}{145.21}=-0.957$	$\frac{4.78-4.03}{1.33}=0.564$	Yes
$\frac{302-208}{145.21}=0.647$	$\frac{2.31-4.03}{1.33}=-1.293$	No
$\frac{60-208}{145.21}=-1.019$	$\frac{3.16-4.03}{1.33}=-0.654$	Yes
$\frac{393-208}{145.21}=1.274$	$\frac{4.2-4.03}{1.33}=0.128$	No

(c) –

$$\text{C} = \text{Characters Standardized} = \frac{242-208}{145.21} = .234$$

$$\text{L} = \text{Word Length standardized} = \frac{4.56-4.03}{1.33} = .398$$

$$\mathbf{P(A = yes) = 0.6}$$

$$\text{P(A = yes | C = .234, L = .398)}$$

$$\text{P(C = .234) = 0.34}$$
$$\text{P(L = .398) = 0.34}$$

$$\text{P(A = Yes | C = .234, L = .398)} = \frac{P(A)P(C,L|A)}{P(C,L)}$$

$$P(A = Yes|C = 0.2341, L = 0.4028\ ) = \frac{0.6 * P(C|A)P(L|A)}{P(C)P(L)}$$

$$P(A = Yes|C = 0.2341, L = 0.4028\ ) = \frac{0.6 * P(C|A)P(L|A)}{0.34 * 0.34}$$

2 Logistic Regression Spam Classification

(a) Precision
0.8671454219030521

(b) Recall
0.8370883882149047

(c) F-measure
0.8518518518518519

(d) Accuracy
0.8904109589041096

3 Naive Bayes Classifier

(a) Precision
0.6465116279069767

(b) Recall
0.9636048526863085

(c) F-measure
0.7738343771746694

(d) Accuracy
0.7879973907371167

4 Decision Trees

I was not able to compute the stats as I had not completed finished implementing the DTL algorithm.