Chris Kasper
CS 383-002
Homework #4

1. Theory

1.

Y	x_1	x_2	Count
+	T	T	3
+	T	F	4
+	F	T	4
+	F	F	1
-	T	T	0
-	T	F	1
-	F	T	3
-	F	F	5

a. Entropy of system:

$$H(Y) = H\left(\frac{12}{21}, \frac{9}{21}\right) = -\frac{12}{21} * \log_2(\frac{12}{21}) - \frac{9}{21} * \log_2\left(\frac{9}{21}\right) = 0.9852$$

b. Information Gains for x_1 and x_2:

For x_1:
T – (7/8, +), (1/8, -)
F – (5/13, +), (8/13, -)

$$H(x_1(T)) = H\left(\frac{7}{8}, \frac{1}{8}\right) = -\frac{7}{8} * \log_2(\frac{7}{8}) - \frac{1}{8} * \log_2\left(\frac{1}{8}\right) = 0.5436$$

$$H(x_1(F)) = H\left(\frac{5}{13}, \frac{8}{13}\right) = -\frac{5}{13} * \log_2(\frac{5}{13}) - \frac{8}{13} * \log_2\left(\frac{8}{13}\right) = 0.9612$$

$$E(x_1) = \frac{8}{21} * 0.5436 + \frac{13}{21} * 0.9612 = 0.8021$$

$$IG(x_1) = 0.9852 - 0.8021 = 0.1831$$

For x_2:
T – (7/10, +), (3/10, -)
F – (5/11, +), (6/11, -)

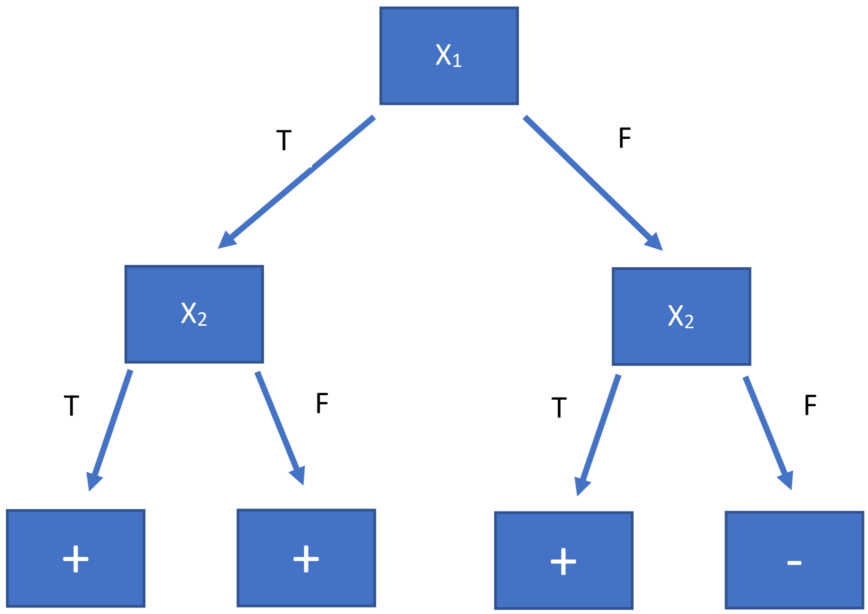
$$H(x_2(T)) = H\left(\frac{7}{10}, \frac{3}{10}\right) = -\frac{7}{10} * \log_2(\frac{7}{10}) - \frac{3}{10} * \log_2\left(\frac{3}{10}\right) = 0.8813$$

$$H(x_2(F)) = H\left(\frac{5}{11},\frac{6}{11}\right) = -\frac{5}{11} * \log_2(\frac{5}{11}) - \frac{6}{11} * \log_2\left(\frac{6}{11}\right) = 0.9940$$

$$E(x_2) = \frac{10}{21} * 0.8813 + \frac{11}{21} * 0.9940 = 0.9403$$

$$IG(x_2) = 0.9852 - 0.9403 = 0.0449$$

c. Decision tree using ID3 algorithm:



Note: x_2 nodes, when x_1 is T, could be collapsed to just "+"

2.

# of Chars	Average Word Length	Give an A
216	5.68	Yes
69	4.78	Yes
302	2.31	No
60	3.16	Yes
393	4.2	No

a. Class priors:

$$P(A = Y) = 3/5$$
$$P(A = N) = 2/5$$

b. Finding Gaussian parameters to do Gaussian Naïve Bayes classification on the decision to give an A or not:

$$data_{standardized} = \begin{bmatrix} 0.0551 & 1.2477 \\ -0.9572 & 0.5688 \\ 0.6473 & -1.2945 \\ -1.0192 & -0.6533 \\ 1.2740 & 0.1313 \end{bmatrix}, Y = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

$$mean = [208, 4.0260]$$
$$std = [145.2154, 1.3256]$$

Gaussian Parameters for A=Y models:

$$data(A = Y) = \begin{bmatrix} 0.0551 & 1.2477 \\ -0.9572 & 0.5688 \\ -1.0192 & -0.6533 \end{bmatrix}$$

$$\begin{array}{ccc} mean & std & \\ \multicolumn{2}{c}{\begin{bmatrix} -0.6404 & 0.6031 \\ 0.3877 & 0.9633 \end{bmatrix}} & \begin{array}{c} \#chars \\ avg\ word\ length \end{array} \end{array}$$

Gaussian Parameters for A=N models:

$$data(A = N) = \begin{bmatrix} 0.6473 & -1.2945 \\ 1.2740 & 0.1313 \end{bmatrix}$$

$$\begin{array}{ccc} mean & std & \\ \multicolumn{2}{c}{\begin{bmatrix} 0.9606 & 0.4431 \\ -0.5816 & 1.0082 \end{bmatrix}} & \begin{array}{c} \#chars \\ avg\ word\ length \end{array} \end{array}$$

c. Given an essay with 242 characters and an average word length of 4.56, determine whether or not it would get an A.

$$P(f_k = x_k \mid y = i) = \frac{1}{\sigma_i \sqrt{2\pi}} * e^{-\frac{(x_k - \mu_i)^2}{2\sigma_i^2}}$$

$$x = (242, 4.56) => x_{standardized} = (0.2341, 0.4028)$$

$$P(A = Yes \mid f = x) = P(A = Yes) * P(f_1 = x_1 \mid A = Yes) * \ P(f_2 = x_2 \mid A = Yes)$$

$$P(f_1 = x_1 \mid A = Yes) = \frac{1}{0.6031 * \sqrt{2\pi}} * e^{-\frac{(0.2341 - -0.6404)^2}{2*0.6031^2}} = 0.2312$$

$$P(f_2 = x_2 \mid A = Yes) = \frac{1}{0.9633 * \sqrt{2\pi}} * e^{-\frac{(0.4028-0.3877)^2}{2*0.9633^2}} = 0.4141$$

$$P(A = Yes \mid f = x) = \frac{3}{5} * \ 0.2312 * \ 0.4141 = 0.0574$$

$$P(A = No \mid f = x) = P(A = No) * P(f_1 = x_1 \mid A = No) * \ P(f_2 = x_2 \mid A = No)$$

$$P(f_1 = x_1 \mid A = No) = \frac{1}{0.4431 * \sqrt{2\pi}} * e^{-\frac{(0.2341-0.9606)^2}{2*0.4431^2}} = 0.2348$$

$$P(f_2 = x_2 \mid A = No) = \frac{1}{1.0082 * \sqrt{2\pi}} * e^{-\frac{(0.4028--0.5816)^2}{2*1.0083^2}} = 0.4141$$

$$P(A = No \mid f = x) = \frac{2}{5} * \ 0.2348 * \ 0.2457 = 0.0231$$

$$Since\ P(A = Yes \mid f = x) > P(A = No \mid f = x), this\ essay\ gets\ an\ A$$

3.

$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

$$\mathrm{g}(\mathrm{x}, \theta) = \frac{e^{x\theta} - e^{-x\theta}}{e^{x\theta} + e^{-x\theta}}$$

a. Augment log likelihood objective function to deal with the range of -1 <= tanh(z) <= 1

   For $P(y = 1|x, \theta)$: we need to account for the negative values that could be returned from $g(x, \theta)$ because the equation is bound from -1 to 1. When $g(x, \theta) \approx 1$, we want the probability to 1. When $g(x, \theta) \approx -1$, we want the probability to 0. When $g(x, \theta) \approx 0$, we want the probability to 0.5, hence why I add 1 to $g(x, \theta)$ and then divide by 2. The exponent accounts for the two binary outcomes, 1 and -1. When y=1, the exponent = 1, and when y= -1, the exponent = 0.

$$P(y = 1|x, \theta) = \left(\frac{g(x, \theta) + 1}{2}\right)^{\frac{1+y}{2}}$$

For $P(y=0|x,\theta)$: We just need to subtract 1 by $P(y=1|x,\theta)$. The exponent accounts for the two binary outcomes, 1 and -1. When y=1, the exponent = 0, and when y= -1, the exponent = 1.

$$P(y=0|x,\theta) = \left(1 - \frac{g(x,\theta)+1}{2}\right)^{\frac{1-y}{2}}$$

Therefore, the new log likelihood objective function is …

$$\ell(y|x,\theta) = \left(\frac{g(x,\theta)+1}{2}\right)^{\frac{1+y}{2}} \left(1 - \frac{g(x,\theta)+1}{2}\right)^{\frac{1-y}{2}}$$

b. Show that:

$$\frac{d}{d\theta_j}(\tanh(x,\theta)) = x_j(1-\tanh(x\theta)^2)$$

$$\frac{d}{d\theta_j}(\tanh(x,\theta)) = \frac{e^{x\theta}-e^{-x\theta}}{e^{x\theta}+e^{-x\theta}}$$

Using quotient rule…

$$\frac{d}{d\theta_j}(\tanh(x,\theta)) = \frac{\left(e^{x\theta}-e^{-x\theta}\right)'(e^{x\theta}+e^{-x\theta}) - \left(e^{x\theta}-e^{-x\theta}\right)\left(e^{x\theta}+e^{-x\theta}\right)'}{(e^{x\theta}+e^{-x\theta})^2}$$

$$\frac{d}{d\theta_j}(\tanh(x,\theta)) = \frac{\left(xe^{x\theta}+xe^{-x\theta}\right)(e^{x\theta}+e^{-x\theta}) - \left(e^{x\theta}-e^{-x\theta}\right)\left(xe^{x\theta}-xe^{-x\theta}\right)}{(e^{x\theta}+e^{-x\theta})^2}$$

$$\frac{d}{d\theta_j}(\tanh(x,\theta)) = \frac{\left(xe^{x\theta}+xe^{-x\theta}\right)(e^{x\theta}+e^{-x\theta}) - x\left(e^{x\theta}-e^{-x\theta}\right)^2}{(e^{x\theta}+e^{-x\theta})^2}$$

$$\frac{d}{d\theta_j}(\tanh(x,\theta)) = \frac{\left(xe^{x\theta}+xe^{-x\theta}\right)(e^{x\theta}+e^{-x\theta})}{(e^{x\theta}+e^{-x\theta})^2} - \frac{x\left(e^{x\theta}-e^{-x\theta}\right)^2}{(e^{x\theta}+e^{-x\theta})^2}$$

$$\frac{d}{d\theta_j}(\tanh(x,\theta)) = \frac{x_j\left(e^{x\theta}+e^{-x\theta}\right)(e^{x\theta}+e^{-x\theta})}{(e^{x\theta}+e^{-x\theta})^2} - x_j\tanh(x,\theta)^2$$

$$\frac{d}{d\theta_j}(\tanh(x,\theta)) = x_j - x_j * \tanh(x,\theta)^2$$

$$\frac{d}{d\theta_j}(\tanh(x,\theta)) = x_j(1-\tanh(x,\theta)^2)$$

c. Derivative of log likelihood function

$$\frac{d}{d\theta}(\ell(y|x,\theta)) = \left(\frac{g(x,\theta)+1}{2}\right)^{\frac{1+y}{2}}\left(1-\frac{g(x,\theta)+1}{2}\right)^{\frac{1-y}{2}}$$

$$\frac{d}{d\theta}\big(\ell(y|x,\theta)\big) = \left(\frac{g(x,\theta)+1}{2}\right)^{\frac{1+y}{2}}\left(1^{\frac{1-y}{2}}-\left(\frac{g(x,\theta)+1}{2}\right)^{\frac{1-y}{2}}\right)$$

$$\frac{d}{d\theta}\big(\ell(y|x,\theta)\big) = \left(\frac{g(x,\theta)+1}{2}\right)^{\frac{1+y}{2}} * 1^{\frac{1-y}{2}} - \left(\frac{g(x,\theta)+1}{2}\right)^{\frac{1+y}{2}} * \left(\frac{g(x,\theta)+1}{2}\right)^{\frac{1-y}{2}}$$

$$\frac{d}{d\theta}\big(\ell(y|x,\theta)\big) = \frac{g(x,\theta)+1}{2} - \left(\frac{g(x,\theta)+1}{2}\right)^{2}$$

$$\frac{d}{d\theta}\big(\ell(y|x,\theta)\big) = \frac{g(x,\theta)+1}{2} - \frac{g(x,\theta)^2+2g(x,\theta)+1}{4}$$

$$\frac{d}{d\theta}\big(\ell(y|x,\theta)\big) = \frac{g(x,\theta)+1}{2} - \frac{g(x,\theta)^2+2g(x,\theta)+1}{4}$$

$$\frac{d}{d\theta}\big(\ell(y|x,\theta)\big) = \frac{1}{2} * \frac{d}{d\theta}(g(x,\theta)+1) - \frac{1}{4} * \frac{d}{d\theta}(g(x,\theta)^2+2g(x,\theta)+1)$$

$$LHS: \frac{d}{d\theta}\big(\ell(y|x,\theta)\big) = \frac{1}{2} * \Big(x_j(1-\tanh(x\theta))\Big)$$

$$RHS: \frac{d}{d\theta}\big(\ell(y|x,\theta)\big)$$
$$= \frac{1}{4} * \big(2(\tanh(x\theta)) * x_j(1-\tanh(x\theta)) + 2(x_j(1-\tanh\,(x\theta)) + 1\big)$$

$$RHS: \frac{d}{d\theta}\big(\ell(y|x,\theta)\big) = \frac{1}{2} * \big(\tanh(x\theta) * x_j(1-\tanh(x\theta)) + (x_j(1-\tanh\,(x\theta)) + 1\big)$$

$$\frac{d}{d\theta}\big(\ell(y|x,\theta)\big) = \frac{1}{2} * \Big(x_j\big(1-\tanh\big(x\theta_j\big)\big)\Big)$$
$$-\frac{1}{2} * \big(\tanh\big(x\theta_j\big) * x_j(1-\tanh\big(x\theta_j\big)) + (x_j(1-\tanh\,(x\theta_j)) + 1\big)$$

2. Naïve Bayes Classifier

Precision:	68.16%
Recall:	95.75%
F-Measure:	79.63%
Accuracy:	81.23%