CS 383 - Machine Learning

Assignment 2 - Classification

1 Theory

1. Consider the following set of training examples for an unknown target function: $(x_1, x_2) \rightarrow y$:

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) What is the sample entropy, H(Y) from this training data (using log base 2)?

$$\rightarrow Count(+) = 3 + 4 + 4 + 1 = 12, \quad Count(-) = 0 + 1 + 3 + 5 = 9$$

$$\therefore P(+) = \frac{12}{21}, \quad P(-) = \frac{9}{21}$$

$$H(Y) = -\sum_{i}^{n} P(v_i) \log_2(P(v_i))$$

$$\rightarrow H(+,-) = -\frac{12}{21}\log_2\left(\frac{12}{21}\right) - \frac{9}{21}\log_2\left(\frac{9}{21}\right) = 0.98522813602$$

(b) What are the information gains for branching on variables x_1 and x_2?

$$IG(A) = H\Big(\frac{p}{p+n}, \frac{n}{p+n}\Big) - \boldsymbol{E}(H(A))$$

$$TP_{x_1} = 3 + 4 = 7, \quad TN_{x_1} = 0 + 1 = 1$$

$$FP_{x_1} = 4 + 1 = 5, \quad FN_{x_1} = 3 + 5 = 8$$

$$\frac{TP + FP}{Total} = \frac{8}{21}, \quad \frac{TN + FN}{Total} = \frac{13}{21}$$

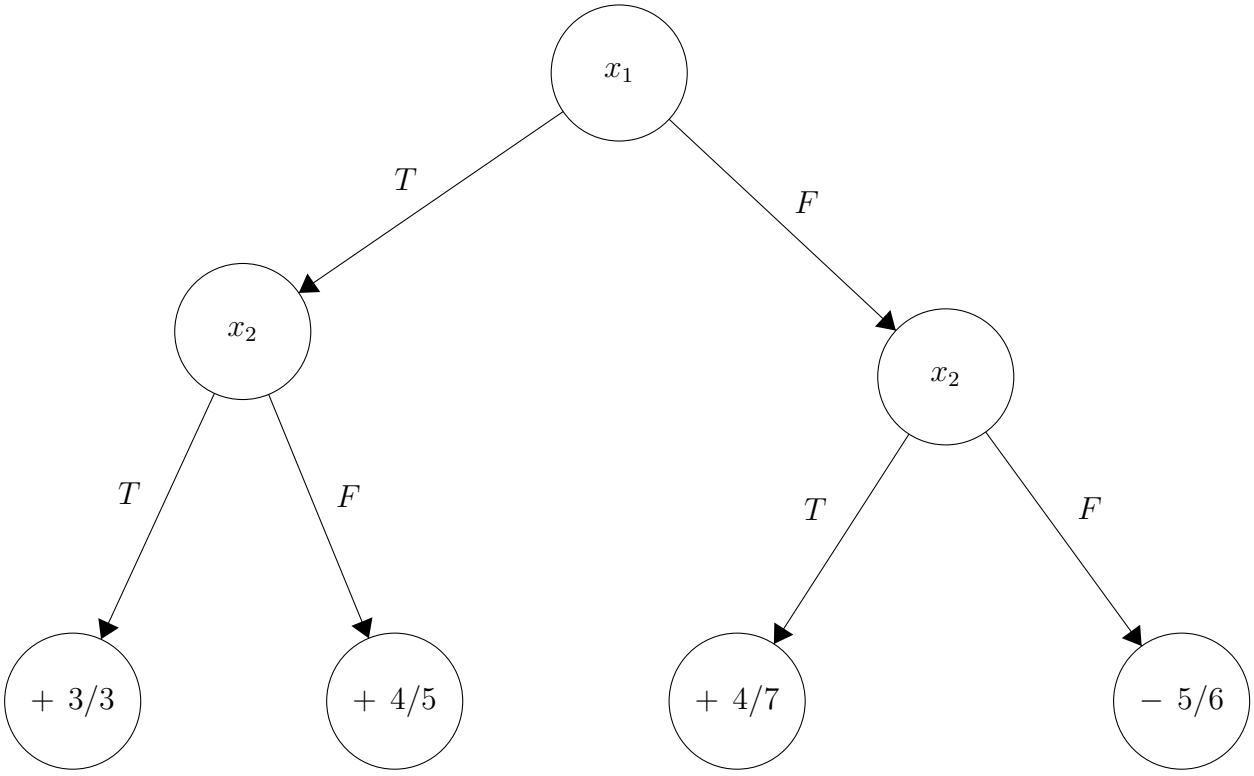
$$H\Big(\frac{7}{8}, \frac{1}{8}\Big) = -\frac{7}{8}\log_2\left(\frac{7}{8}\right) - \frac{1}{8}\log_2\left(\frac{1}{8}\right) = 0.5435644432$$

$$H\Big(\frac{5}{13}, \frac{8}{13}\Big) = -\frac{5}{13}\log_2\left(\frac{5}{13}\right) - \frac{8}{13}\log_2\left(\frac{8}{13}\right) = 0.96123660471$$

$$\boldsymbol{E}(H(x_1)) = \frac{8}{21} \times 0.5435644432 + \frac{13}{21} \times 0.96123660471 = 0.801212340032$$
$$IG(x_1) = H(+,-) - \boldsymbol{E}(H(x_1)) = 0.98522813602 - 0.801212340032 = 0.1831047357$$

$$TP_{x_2} = 3 + 4 = 7, \quad TN_{x_2} = 0 + 3 = 3$$
$$FP_{x_2} = 4 + 1 = 5, \quad FN_{x_2} = 1 + 5 = 6$$
$$\frac{TP + FP}{Total} = \frac{10}{21}, \quad \frac{TN + FN}{Total} = \frac{11}{21}$$
$$H\Big(\frac{7}{10}, \frac{3}{10}\Big) = -\frac{7}{10}\log_2\Big(\frac{7}{10}\Big) - \frac{3}{10}\log_2\Big(\frac{3}{10}\Big) = 0.88129089923$$
$$H\Big(\frac{5}{11}, \frac{6}{11}\Big) = -\frac{5}{11}\log_2\Big(\frac{5}{11}\Big) - \frac{6}{11}\log_2\Big(\frac{6}{11}\Big) = 0.99403021147$$
$$\boldsymbol{E}(H(x_2)) = \frac{10}{21} \times 0.88129089923 + \frac{11}{21} \times 0.99403021147 = 0.94034482468$$
$$IG(x_2) = H(+,-) - \boldsymbol{E}(H(x_2)) = 0.88129089923 - 0.99403021147 = 0.04488331134$$

(c) Draw the decision tree that would be learned by the ID3 algorithm without pruning from this training data.



2. We decided that maybe we can use the number of characters and the average word length an essay to determine if the student should get an A in a class or not. Below are five samples of this data:

# of Chars	Average Word Length	Given an A
216	5.68	Yes
69	4.78	Yes
302	2.31	No
60	3.16	Yes
393	4.20	No

(a) What are the class priors, P(A = Yes), P(A = No)?

$$P(A = Yes) = \frac{3}{5} \quad P(A = No) = \frac{2}{5}$$

(b) Find the parameters of the Gaussians necessary to do Gaussian Naive Bayes classification on this decision to give an A or not. Standardize the features first over all the data together so that there is no unfair bias towards the features of different scales.

$$X = \begin{bmatrix} 216 & 5.68 \\ 69 & 4.78 \\ 302 & 2.31 \\ 60 & 3.16 \\ 393 & 4.20 \end{bmatrix} \Rightarrow \mu = \begin{bmatrix} 208 & 4.026 \end{bmatrix}, \quad \sigma \begin{bmatrix} 145.2154 & 1.3256 \end{bmatrix}$$

$$\Rightarrow X_{std} = \frac{X - \mu}{\sigma} = \begin{bmatrix} 0.0551 & 1.2477 \\ -0.9572 & 0.5688 \\ 0.6473 & -1.2945 \\ -1.0192 & -0.6533 \\ 1.2740 & 0.1313 \end{bmatrix}$$

$$A = True \Rightarrow \begin{bmatrix} 0.0551 & 1.2477 \\ -0.9572 & 0.5688 \\ -1.0192 & -0.6533 \end{bmatrix}$$

$$\mu(A = True, axis = 0) = \begin{bmatrix} -0.6404 & 0.3877 \end{bmatrix}$$
$$\sigma(A = True, axis = 0) = \begin{bmatrix} 0.4925 & 0.7865 \end{bmatrix}$$

$$A = False \Rightarrow \begin{bmatrix} 0.6473 & -1.2945 \\ 1.2740 & 0.1313 \end{bmatrix}$$

$$\mu(A = False, axis = 0) = \begin{bmatrix} 0.9607 & -0.5816 \end{bmatrix}$$
$$\sigma(A = False, axis = 0) = \begin{bmatrix} 0.31335 & 0.7129 \end{bmatrix}$$

(c) Using your response from the prior question, determine if an essay with 242 characters and an average word length of 4.56 should get an A or not.

$$newX = \begin{bmatrix} 242 & 4.56 \end{bmatrix}$$

$$newX_{std} = \frac{newX - \mu}{\sigma} = \begin{bmatrix} 0.2341 & 0.4028 \end{bmatrix}$$

$$Posterior(A) \approx P(A) \times P(\#\ ofChars \mid A) \times P(AvgWordLength \mid A)$$

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

$$P(\#\ ofChars \mid A) := \frac{e^{\frac{-(0.2341-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} = 0.2312$$

$$P(AvgWordLength \mid A) := \frac{e^{\frac{-(0.4028-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} = 0.5071$$

$$Posterior(A) = 0.07034$$

$$P(A) := \frac{2}{5}$$

$$P(\#\ ofChars \mid NotA) := \frac{e^{\frac{-(0.2341-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} = 0.00088675$$

$$P(AvgWordLength \mid NotA) := \frac{e^{\frac{-(0.4028-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} = 0.2157$$

$$Posterior(NotA) = 0.000076508$$

Posterior(A) > Posterior(Not A) therefore the essay should get an A.

2 Logistic Regression Spam Classification

$$TP = 422,\ TN = 922,\ FP = 35,\ FN = 154$$

$$Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN}, F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, Accuracy = \frac{TP + TN}{total}$$

$$Precision = 82.0175\%,\ Recall = 92.2697\%\ , F_1 = 86.8421\%,\ Accuracy = 88.8084\%$$

3 Naive Bayes Classifier

$$Precision = 62.4722\%,\ Recall = 97.9058\%\ , F_1 = 76.2746\%,\ Accuracy = 77.0244\%$$

4 Decision Trees

$$Precision = 90.0169\%,\ Recall = 87.5000\%\ , F_1 = 88.7406\%,\ Accuracy = 91.1126\%$$