Aslan Oztreves

CS 383

Matthew Burlick

Homework #4

Theory Problems

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

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) (5pts)?

(b) What are the information gains for branching on variables x_1 and x_2 (5pts)?

(c) Draw the decision tree that would be learned by the ID3 algorithm without pruning from this training data. All leaf nodes should have a single class choice at them. If necessary use the mean class or, in the case of a tie, choose one at random.(10pts)?

a)

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

b)

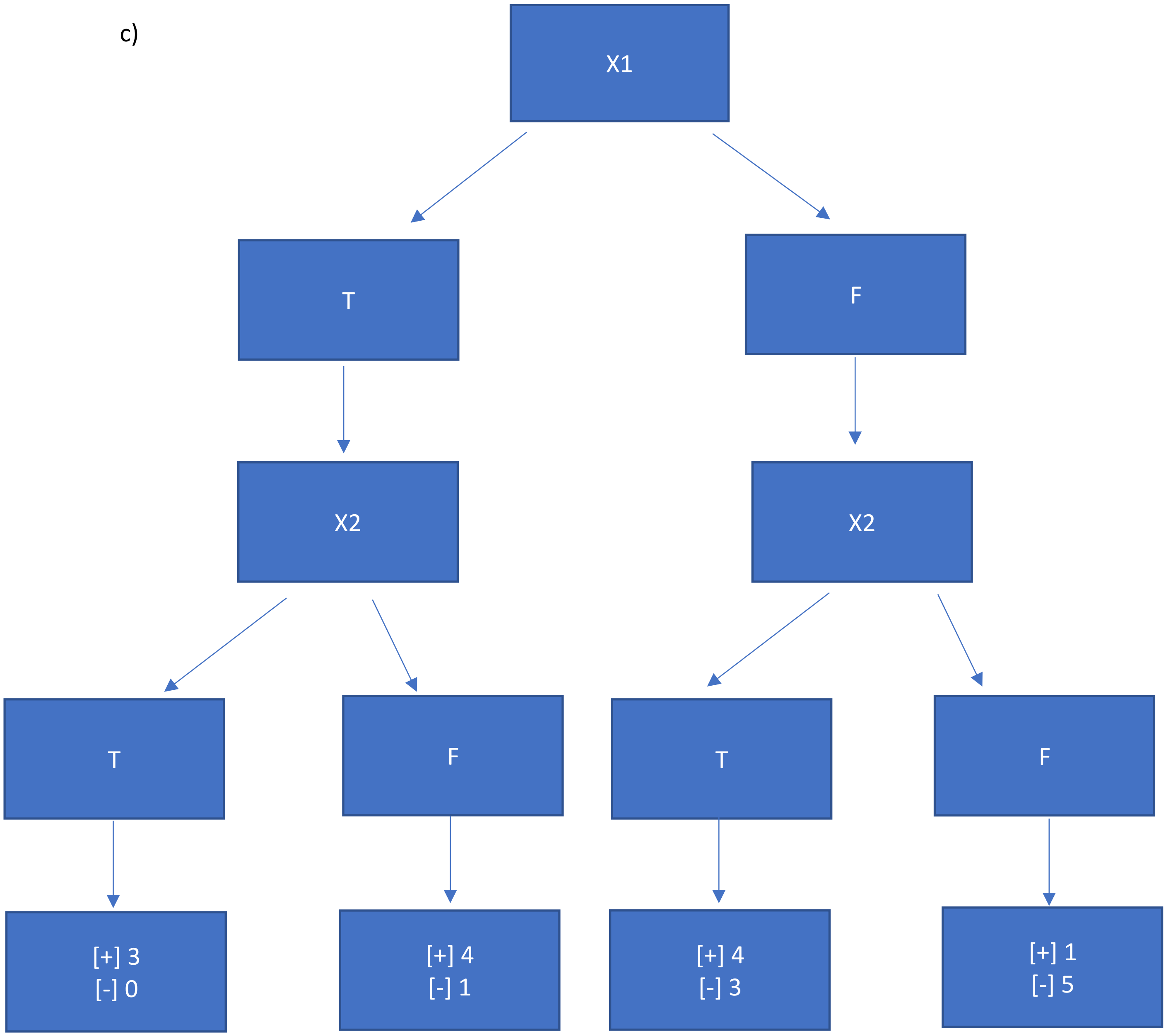
$$remainder(x_1) = \frac{8}{21} H\left(\frac{7}{8}, \frac{1}{8}\right) + \frac{13}{21} H\left(\frac{5}{13}, \frac{8}{13}\right) = 0.8021$$

$$remainder(x_2) = \frac{10}{21} H\left(\frac{7}{10}, \frac{3}{10}\right) + \frac{11}{21} H\left(\frac{5}{11}, \frac{6}{11}\right) = 0.9403$$

$$IG(x_1) = \mathrm{H}\left(\frac{12}{21}, \frac{9}{21}\right) - remainder(x_1) = 0.1831$$

$$IG(x_2) = \mathrm{H}\left(\frac{12}{21}, \frac{9}{21}\right) - remainder(x_2) = 0.0449$$

c)



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	Give an A
216	5.68	Yes
69	4.78	Yes
302	2.31	No
60	3.16	Yes
393	4.2	No

(a) What are the class priors, $P(A = Yes), P(A = No)$? (5pts)

(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 (5pts).

(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. Show the math to support your decision (10pts).

a)

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

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

b)

Standardized feature, $\mu_1 = 208$, $\mu_2 = 4.026$, $\sigma_1 = 145.2154$, $\sigma_2 = 1.3256$

$$X = \begin{bmatrix} 0.0551 & 1.2477 \\ -0.9572 & 0.5688 \\ 0.6473 & -1.2945 \\ -1.0192 & -0.6533 \\ 1.274 & 0.1313 \end{bmatrix}$$

For P(Yes):

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

$$\mu_1 = -0.6404, \mu_2 = 0.3877, \sigma_1 = 0.6031, \sigma_2 = 0.9633$$

For P(No):

$$X = \begin{bmatrix} 0.6473 & -1.2945 \\ 1.2740 & 0.1313 \end{bmatrix}$$

$$\mu_1 = 0.9607, \mu_2 = -0.5816, \sigma_1 = 0.4431, \sigma_2 = 1.0082$$

c) From the previous part we get 0.2341 and 0.4028 when standardized. We then calculate for

$$P(getA|R) \propto P(getA) \times P(\#char|getA) \times P(avgWord|getA) =$$

$$= \frac{3}{5} \times 0.2312 \times 0.4141 = 0.0574$$

$$P(notA|R) \propto P(notA) \times P(\#char|notA) \times P(avgWord|notA) =$$

$$= \frac{2}{5} \times 0.2347 \times 0.2457 = 0.0232$$

$0.0574 > 0.0232$ so gets an A.

3. Another common activation function for use in logistic regression or artificial neural networks is the hyperbolic tangent function, $tanh$, which is defined as:

$$tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \quad (1)$$

(a) Since the hyperbolic tangent function outputs values in the range of $-1 <= tanh(z) <= 1$ we will have to augment our log likelihood objective function to deal with this range. If we opt to use this function for logistic regression (as opposed to $\frac{1}{1+e^{-z}}$), what will this object function be? Show your work. (5pts)

(b) In order to compute the gradient of your previous answer with respect to θ_j, we'll need to compute the gradient of the hyperbolic tangent function itself. Use the exponential definition of the hyperbolic tangent function provided at the top of this problem to show that
$\frac{\partial}{\partial\theta_j}(tanh(x\theta)) = x_j(1 - tanh(x\theta)^2)$. (5pts)

(c) Using the fact that $\frac{\partial}{\partial\theta_j}(tanh(x\theta)) = x_j(1 - tanh(x\theta)^2)$, what is the gradient of your log likelihood function in part (a) with respect to θ_j? Show your work. (5pts)

a)

$$= \ln\left(\frac{e^z - e^{-z}}{e^z + e^{-z}}\right) = \frac{e^{2z} - 1}{e^{2z} + 1}$$

$$e^{2z} = u$$

$$t = \frac{u-1}{u+1}, so\ u = -1$$

$$e^{2z} = -1, z = NaN$$

b)

Naïve Bayes Classifier

Data Results:

precision = 0.6874

recall = 0.9591
fmeasure = 0.8009
accuracy = 0.8174