CS 613 - Machine Learning

Assignment 2 - Classification
Xiangang Lai

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) (2pts)?

The total count is 21, while +Y is 12 and -Y has 9.
Therefore, $P(Y = +) = 12/21 = 4/7$**,** $P(Y = -) = 3/7$**,**
The entropy is $H(4/7, 3/7) = -4/7 * log_2(3/7) - 3/7 * log_2(4/7) = 0.9852$

(b) What are the information gains for branching on variables x_1 and x_2 (2pts)?
Variable x1:

$$T\colon p1 = 7, n1 = 1$$
$$F\colon p2 = 5, n2 = 8$$
$$All\colon p = 12, n = 9$$

$$\begin{aligned} E(H(x1)) &= (7+1)/(12+9) - 7/8 log_2(1/8) - 1/8 log_2(7/8) \\ &+ (5+8)/(12+9) - 5/13 log_2(8/13) - 8/13 log_2(5/13) \\ &= 0.8021 \end{aligned}$$

$$IG(x1) = H(4/7, 3/7) - E(H(x1)) = 0.1831$$

Variable x2:

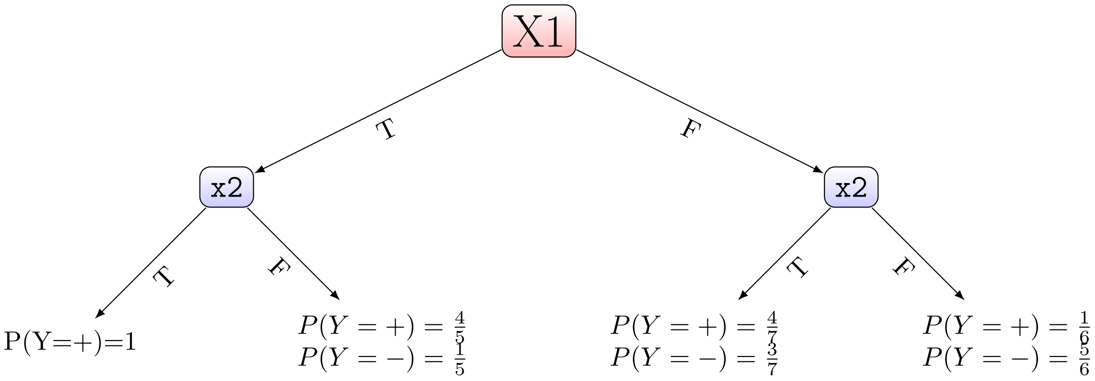
$$
\begin{aligned}
T &: p1 = 7, n1 = 3 \\
T &: p2 = 5, n2 = 6 \\
All &: p = 12, n = 9
\end{aligned}
$$

$$
\begin{aligned}
E(H(x2)) &= (7+3)/(12+9) - 7/10 log_2(3/10) - 3/10 log_2(7/10) \\
&+ (5+6)/(12+9) - 5/11 log_2(6/11) - 6/11 log_2(5/11) \\
&= 0.9403
\end{aligned}
$$

$$IG(x2) = H(4/7, 3/7) - E(H(x4)) = 0.0449$$

So we should prioritize variable x1

(c) Draw the deicion tree that would be learned by the ID3 algorithm without pruning from this training data (3pts)?



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, $P(A = Yes), P(A = No)$? (2pt)

$P(A = Yes) = 0.6, P(A = No) = 0.4$

(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 (2pts).

A: $mean = [-0.640428130.38774181], std = [0.492456430.78656431]$
Not A: $mean = [0.9606422 - 0.58161272], std = [0.313327740.71287162]$
(code in jupyter notebook.as well as (c))

(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 (3pts).

Calculate its proportional probability of A and Not A based on the Naive Bayes classification, get $P(A = Yes|x) = 0.3200, P(A = No|x) = 0.0469$**, obviously** $P(A = Yes|x) > P(A = No|x)$**, therefore the student should get an A**

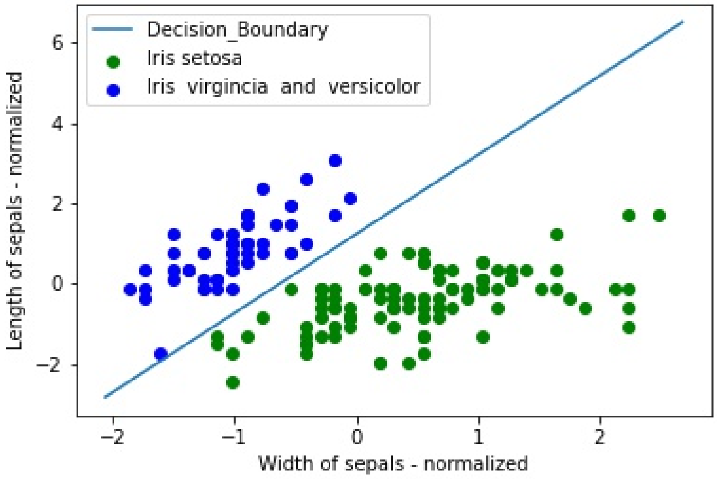
3. Consider the following questions pertaining to a k-Nearest Neighbors algorithm (1pt):

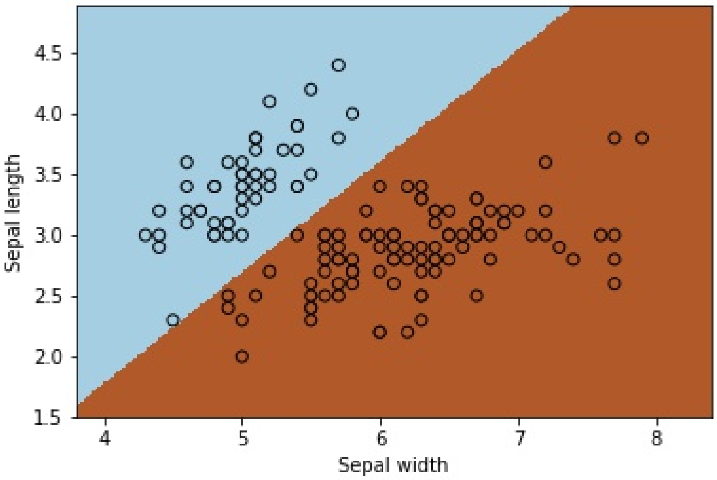
(a) How could you use a *validation set* to determine the user-defined parameter k?

Try a few k values such as (3,5,7,9) and compare its accuracy with the validation set, choose the k value with highest accuracy.

2 Logistic Regression

The final model theta values is: [7.39825871 11.73857941 -5.98106822]





These two figures agree well.

3 Logistic Regression Spam Classification

The precision is: 0.943502824858757
The recall is: 0.8119935170178282
The F_measure is: 0.8728222996515679
The accuracy is: 0.9048239895697523

4 Naive Bayes Classifier

The precision is: 0.7120291616038882
The recall is: 0.9512987012987013
The F_measure is: 0.8144544822793607
The accuracy is: 0.8259452411994785

5 Decision Trees

The precision is: 0.86084142394822
The recall is: 0.9094017094017094
The F$_{m}easureis$: 0.884455527847049
$Theaccuracyis$: 0.909387222946545