CS434 Assignment 2 PART 2 Solution

1. We have two identical bags. Bag A contains 4 red marbles and 6 black marbles and bag B contains 5 red marbles and 5 black marbles. Now we random chose a bag and drew a marble from the chosen bag and it turns out to be black. What is the probability that the chosen bag is bag A?

Solution:

Let A (B) stands for the event that bag A (B) is chosen. Let R (Black) stands for the event that red (black) marble is drawn. Let's first write down some probabilities that we know.

$$P(A) = P(B) = 0.5$$
$$P(R \mid A) = 0.4; P(Black \mid A) = 0.6$$
$$P(R \mid B) = 0.5; P(Black \mid B) = 0.5$$

We need to compute P(A|Black) and we will use Bayes rule to compute this.

$$P(A \mid Black)$$
$$= \frac{P(Black \mid A)P(A)}{P(Black)}$$
$$= \frac{0.6 * 0.5}{P(Black \mid A)P(A) + P(Black \mid B)P(B)}$$
$$= \frac{0.3}{0.6 * 0.5 + 0.5 * 0.5} = \frac{0.3}{0.3 + 0.25} = \frac{0.3}{0.55} = 0.545$$

2. Suppose we have class variable Y and three attributes X_1, X_2, X_3 and we wish to calculate $p(Y \mid X_1; X_2; X_3)$, and we have no conditional independence information.

   (a) Which of the following sets of probabilities are sufficient for calculation?
   - i. $p(Y)$; $p(X_1 \mid Y)$; $p(X_2 \mid Y)$; $p(X_3 \mid Y)$
   - ii. $p(X_1; X_2; X_3)$; $p(Y)$; $p(X_1; X_2; X_3 \mid Y)$
   - iii. $p(X_1; X_2; X_3)$; $p(Y \mid X_1)$; $p(Y \mid X_2)$; $p(Y \mid X_3)$

   (b) Now suppose we know that the variables X_1, X_2, X_3 are conditionally independent given the class variable Y. Which of the above 3 sets are sufficient now?

Solution:

(a) $p(Y \mid X_1, X_2, X_3) = \frac{p(X_1, X_2, X_3 \mid Y)p(Y)}{p(X_1, X_2, X_3)}$

Clearly (ii) provides enough information to compute this but not the others.

(b) $p(Y \mid X_1, X_2, X_3) = \frac{p(X_1, X_2, X_3 \mid Y)p(Y)}{p(X_1, X_2, X_3)} = \frac{p(X_1 \mid Y)p(X_2 \mid Y)p(X_3 \mid Y)p(Y)}{p(X_1, X_2, X_3)}$

Now both (i) and (ii) now suffice. In particular, some of you might have question regarding the availability of p(X1, X2, X3). Note that this can be calculated using the following formula:

$$p(X_1, X_2, X_3) = p(X_1, X_2, X_3|Y = 1)P(Y = 1) + p(X_1, X_2, X_3|Y = 1)P(Y = 0)$$

(Note that here we assume that $p(X_1, X_2, X_3|Y)$ is referring to the distribution of X_1, X_2, X_3 given Y.)

Further to show that (iii) is not enough, we can see that to get $P(X_1|Y)$ from $P(Y|X_1)$ via Bayes rule $P(X_1|Y) = \frac{P(Y|X_1)P(X_1)}{P(Y)}$, we will have to have $P(X_1)$, and P(Y), which cannot be computed using what's provided in (iii).

3. (Naïve Bayes Classifier) We will use the following training set to build a Naïve Bayes classifier. A, B, and C are three binary attributes and Y is the target class label.

A	B	C	Y
0	1	1	0
1	1	1	0
0	0	0	0
1	1	0	1
0	1	0	1
1	0	1	1

   a. Based on the training data, calculate the prior distribution for Y: P(Y)
   P(Y=1)=3/6=0.5; P(Y=0)=1-0.5=0.5

   b. Based on the training data, calculate the distributions P(A|Y), P(B|Y) and P(C|Y).

	P(A|Y=1)
A=1	2/3
A=0	1/3

	P(A|Y=0)
A=1	1/3
A=0	2/3

	P(B|Y=1)
B=1	2/3
B=0	1/3

	P(B|Y=0)
B=1	2/3
B=0	1/3

	P(C|Y=1)
C=1	1/3
C=0	2/3

	P(C|Y=0)
C=1	2/3
C=0	1/3

c. What prediction will the naïve bayes classifier make for a new example (A=1, B=0, C=0)?

$P(Y=1|A=1,B=0, C=0) = P(A=1,B=0,C=0|Y=1)P(Y=1)/$
$[P(A=1,B=0,C=0|Y=1)P(Y=1)+ P(A=1,B=0,C=0|Y=0)P(Y=0)]$

$P(A=1,B=0,C=0|Y=1) = P(A=1|Y=1)\ P(B=0|Y=1)\ P(C=0|Y=1) = 2/3*1/3*2/3=4/9$

$P(A=1,B=0,C=0|Y=0) = P(A=1|Y=0)\ P(B=0|Y=0)\ P(C=0|Y=0) = 1/3*1/3*1/3=1/9$

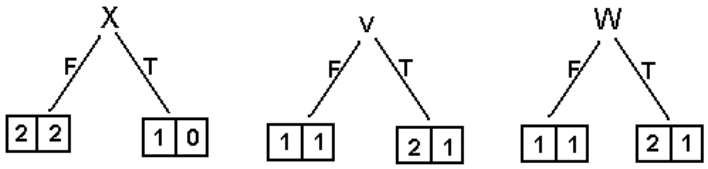
$P(Y=1|A=1,B=0, C=0) = 4/9*0.5/[4/9*0.5+1/9*0.5] = 0.8$, $P(Y=0|A=1,B=0, C=0) = 0.2$

$P(Y=1|A=1,B=0, C=0) > P(Y=0|A=1,B=0, C=0)$

Thus we predict Y=1 for (A=1, B=0, C=0)

3. Decision tree and pruning.

(a). Compute the information gain of features X, V and W respectively,



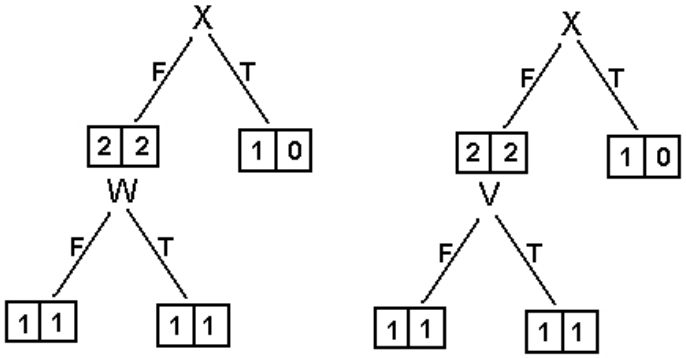
For feature X:
$E_{root} = -0.4 \log 0.4 - 0.6 \log 0.6 = 0.971$

$E_{Left} = -0.5 \log 0.5 - 0.5 \log 0.5 = 1$
$E_{Right} = 0$
Information gain $= 0.971 - 4/5 = 0.171$

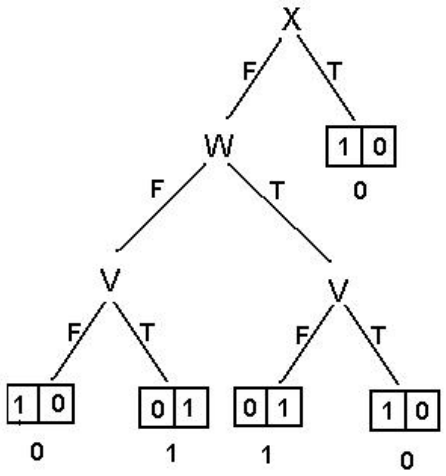
For feature V:
$E_{root} = 0.971$
$E_{Left} = -0.5 \log 0.5 - 0.5 \log 0.5 = 1$
$E_{Right} = -\frac{2}{3} \log \frac{2}{3} - \frac{1}{3} \log \frac{1}{3} = 0.918$
Information gain $= 0.971 - \frac{2}{5} \times 1 - \frac{3}{5} \times 0.918 = 0.02$

The computation for feature W is exactly the same as feature V and its information gain is 0.02.

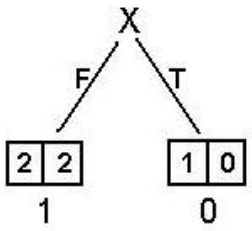
(b). Write down the full decision tree.
From (a), we see that the root node test should be on X. Its right child contains only one class-0 instance, thus no further testing required and is turned into a leave node predicting class 0. Its left child contains 2 class-0 instances and 2 class-1 instances. Therefore it needs further testing. We compute the information gain of W and V using the remaining four instances.



From the above picture, we can see that the information gain is zero for both V and W. So we simply randomly select one, let's say W, as the test for this node. Because the result children are all impure, we need to further test on V to complete the tree. The final decision tree is depicted as follows. Note that it is also correct if we switch the position of V and W in the decision tree.



– [c.] Prunning based on information gain
For strategy i, node X will not be pruned because its information gain is $0.171 > \epsilon$. However, The second level test W will be pruned and turned into a leaf node because its information gain is zero. The resulting tree is as follows. Note that because for this new leaf node, the two classes are equally represented, we can simply randomly select a label for the leaf node, both 0 and 1 will be correct. The resulting training error will be 40%.



For strategy ii, both of the third level tests have information gain 1, thus are intact. The tree remains the same and the training error is 0.

– [d.] Comparing strategy i and ii.
Strategy i is computationally cheaper - because we can stop early and avoid building the full tree. But as we can see it may prune too much. Strategy ii is more computationally expensive because we have to build a full tree first, which can become exponentially large. Another less obvious problem with stratgety ii (you are not required to make this observation in your solution to get the full mark) is that at the low level (close to the bottom) of the tree the number of examples in the subtree can get quite small so information gain might not be an appropriate criterion for pruning.