Homework 1 ID3 Implementation

Zeynep Mutlu Hakguder

Department of Electrical and Computer Engineering

University of Nebraska-Lincoln, USA

zeynep.hakguder@huskers.unl.edu

Haluk Dogan

Department of Electrical and Computer Engineering

University of Nebraska-Lincoln, USA

haluk.dogan@huskers.unl.edu

September 25, 2016

Contents

1 Question 1

The training data consists of 6 examples each with a single integer-valued attribute and a binary label. The target function $\mathcal{C}$ is represented as a single interval using two points a and b.

1	-	6		11	
2		7	+	12	-
3		8	+	13	
4		9		14	
5	+	10		15	-

Table 1: Training Set

(a) A hypothesis consistent with the training set would be obtained if $a = 5$ and $b = 8$ so that an instance x is labeled positive if and only if $5 \leq x \leq 8$.

(b) The version space is the subset of hypothesis space $\mathcal{H}$ (which is chosen as the family of single intervals and is guaranteed to include target concept) that is consistent with the training set. For example, in the interval given in (a), all positive training examples are contained, if it was any smaller, it would exclude a positive training example and be inconsistent with the data. We could expand the interval to $2 \leq x \leq 11$, this is the largest interval that is consistent with the training set, if we were to expand it any further we would end up including a negative example in the interval that represents the target function. The version space is the set of all such hypotheses (represented by intervals) that are consistent with the training set. The interval's lower bound can be any of $2, 3, 4$ or 5, its upper bound can be any of $8, 9, 10$ or 11, the size of the version space is $4 \times 4 = 16$.

(c) To decrease the size of the version space we can specify the query $q_1 = 10$. If the label for 10 is:

- positive, the smallest interval consistent with our training data and q_1 would be $5 \leq x \leq 10$. We would eliminate 8 hypotheses that assign negative labels

to 9 and 10 from the version space.

- negative, the smallest interval consistent with our training data and q_1 would be $5 \leq x \leq 9$. We would eliminate 8 hypotheses that assign positive labels to 10 and 11 from the version space.

If we set $q_2 = 6$, this query wouldn't change the size of the version space regardless of the answer.

(d) If we have a training set with three examples $\{(x_1, -), (x_2, +), (x_3, -)\}$, where $0 < x_1 < x_2 < x_3$, in order to reduce the size of the version space, we can form queries in a manner similar to binary search. First, to find the upper bound of the interval we ask the label of the midpoint between x_2 and x_3, depending on the label we receive we form our next query; i.e if it is negatively labeled we ask the label for the midpoint between the last query and x_2, if it is positively labeled we ask the label for the midpoint between the last query and x_3. Similarly, to obtain the lower boundary we ask the label for the midpoint between x_1 and x_2, if it is negatively labeled we ask the label for the midpoint between the last query and x_2, if it's positively labeled we ask the label for the midpoint between the last query and x_1. Below is an example showing one of the worst case scenarios with training set $\{(1, -), (9, +), (17, -)\}$.

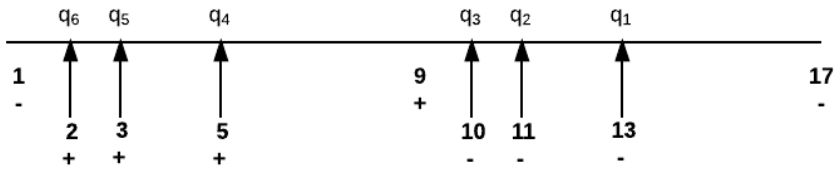

Figure 1: Query example to decrease version space size.

2 Question 2

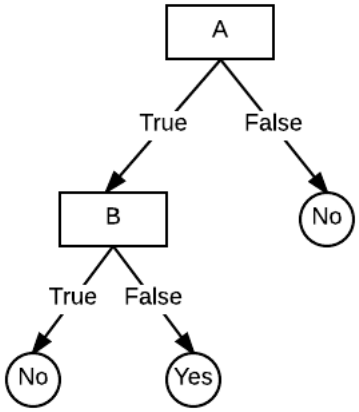

Figure 2: $A \wedge [\sim B]$

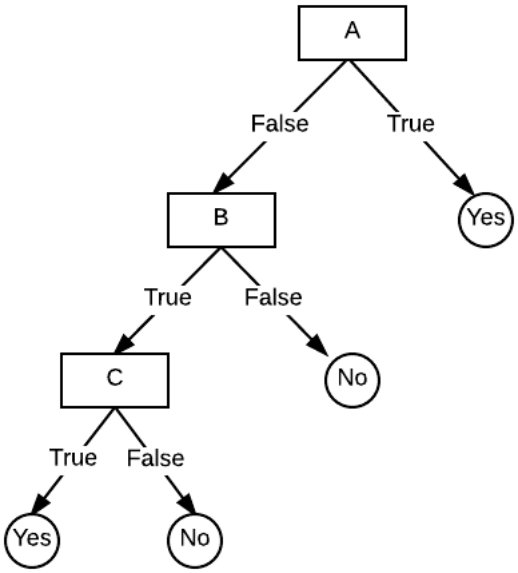

Figure 3: $A \vee [B \wedge C]$

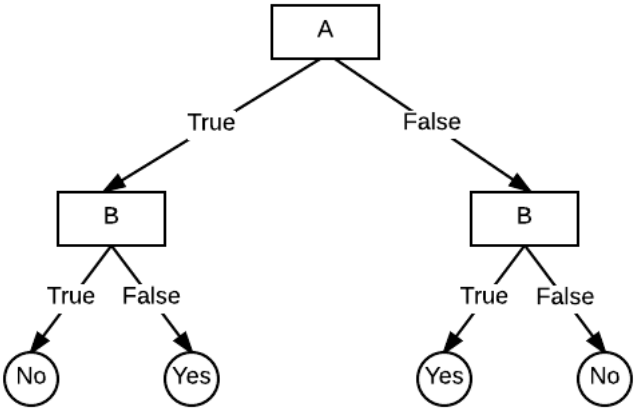

Figure 4: $A \oplus B$

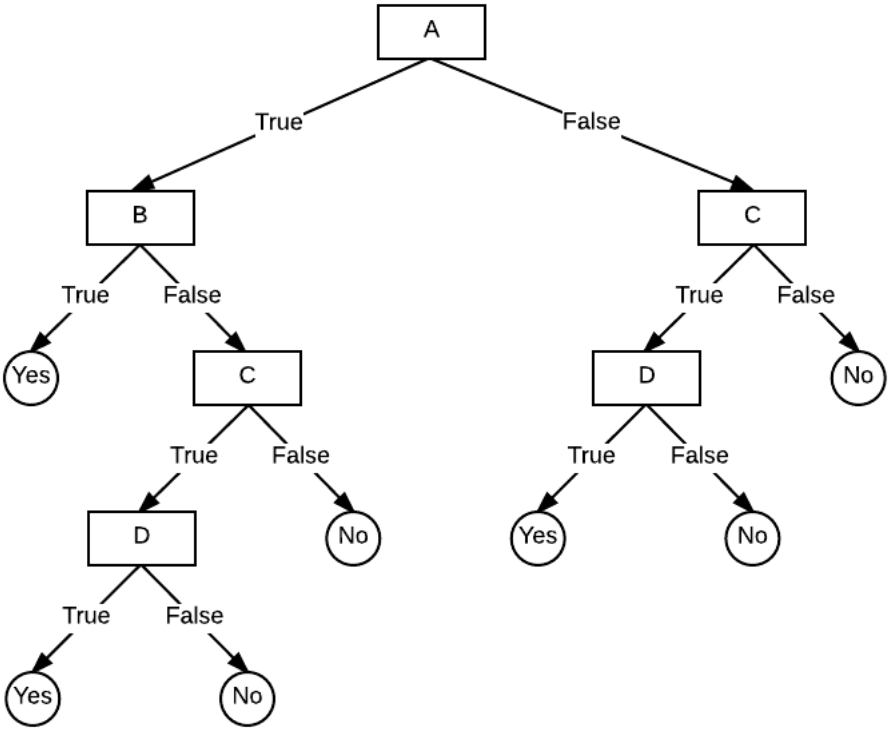

Figure 5: $[A \wedge B] \vee [C \wedge D]$

3 Question 3

(a) The entropy of the dataset is calculated as follows:

$$
\begin{aligned}
I &= -p^+ \log_2 p^+ - p^- \log_2 p^- \\
&= -\frac{3}{6} \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6} \\
&= 1
\end{aligned}
$$

(b) The ID3 algorithm would choose the attribute with smallest impurity which is calculated as below:

$$
\begin{aligned}
I'(a_i) = &\frac{|T|}{|T+F|} \left(-p_T^+ \log_2 p_T^+ - p_T^- \log_2 p_T^-\right) \\
&+ \frac{|F|}{|T+F|} \left(-p_F^+ \log_2 p_F^+ - p_F^- \log_2 p_F^-\right)
\end{aligned}
$$

where $i = 1, 2$.

$$
\begin{aligned}
I'(a_1) &= \frac{3}{6}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) + \frac{3}{6}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) = 0.9183 \\
I'(a_2) &= \frac{4}{6}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) + \frac{2}{6}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) = 1.0
\end{aligned}
$$

Since $I'(a_1) < I'(a_2)$, *ID3* algorithm would choose a_1 next.

4 Question 4

TODO

4.1 Introduction

This is time for all good men to come to the aid of their party!

Outline lorem

4.2 Data

4.3 Implementation

In this section we describe the results.

4.4 Results

We worked hard, and achieved very little.

4.5 Discussion

lorem

References

5 Appendix