Lab3 : Decision Tree

December 2, 2019

1 ID3 algorithm

1.1 Entropy

Entropy is a measure of homogeneity of data. If the data is completely homogeneous the entropy is 0, if the data is equally divided then it has entropy of 1.

$$H(X) = -\sum_{x \in X} p(x) log_2 p(x) \tag{1}$$

Example:

$X_1 = [1, 1, 1, 0, 0, 1, 0, 1, 0, 1]$

$$\begin{aligned} &H(X) = \\ &\quad -p(X=0)log_2 p(X=0) - p(X=1)log_2 p(X=1) = \\ &\quad -0.4log_2(0.4) - 0.6log_2(0.6) = \\ &\quad 0.9709 \end{aligned} \tag{2}$$

1.2 Conditional entropy

$$H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) log_2 \frac{p(x, y)}{p(x)} \tag{3}$$

Equation from lecture

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X = x) \tag{4}$$

$$H(Y|X = x) = -\sum_{y \in Y} p(y|x) log_2 \frac{p(x, y)}{p(x)} \tag{5}$$

Example 1:

$X = [1, 0, 1, 0, 0, 1, 1, 0, 1, 1]$

$Y = [1, 1, 1, 0, 0, 1, 0, 1, 0, 1]$

$$\begin{aligned}
H(Y|X) = & \\
& -p(X = 0, Y = 0) log_2 \tfrac{p(X=0,Y=0)}{p(X=0)} \\
& -p(X = 0, Y = 1) log_2 \tfrac{p(X=0,Y=1)}{p(X=0)} \\
& -p(X = 1, Y = 0) log_2 \tfrac{p(X=1,Y=0)}{p(X=1)} \\
& -p(X = 1, Y = 1) log_2 \tfrac{p(X=1,Y=1)}{p(X=1)} = \\
& -0.2 log_2 \tfrac{0.2}{0.4} \\
& -0.2 log_2 \tfrac{0.2}{0.4} \\
& -0.2 log_2 \tfrac{0.2}{0.6} \\
& -0.4 log_2 \tfrac{0.4}{0.6} \\
= & \quad 0.9509
\end{aligned} \tag{6}$$

Example 2:

$X = [1, 0, 1, 0, 1, 1, 0, 0, 0, 1]$

$Y = [1, 1, 1, 0, 0, 1, 0, 1, 0, 1]$

$$
\begin{aligned}
& H(Y|X) = \\
& \quad -p(X=0,Y=0)log_2\tfrac{p(X=0,Y=0)}{p(X=0)} \\
& \quad -p(X=0,Y=1)log_2\tfrac{p(X=0,Y=1)}{p(X=0)} \\
& \quad -p(X=1,Y=0)log_2\tfrac{p(X=1,Y=0)}{p(X=1)} \\
& \quad -p(X=1,Y=1)log_2\tfrac{p(X=1,Y=1)}{p(X=1)} = \\
& \quad -0.3log_2\tfrac{0.3}{0.5} \\
& \quad -0.2log_2\tfrac{0.2}{0.5} \\
& \quad -0.1log_2\tfrac{0.1}{0.5} \\
& \quad -0.4log_2\tfrac{0.4}{0.5} \\
& = \quad 0.8464
\end{aligned} \tag{7}
$$

1.3 Information gain

$$IG(Y,X) = H(Y) - H(Y|X) \tag{8}$$

1.4 Data set

Given is a data set of consisting 10 objects. Each object is described by 2 binary attribiutes X_1 and X_2 and one decision attribute cl.

	X_1	X_2	cl
0	1	1	1
1	0	0	1
2	1	1	1
3	0	0	0
4	0	1	0
5	1	1	1
6	1	0	0
7	0	0	1
8	1	0	0
9	1	1	1

1.5 Tree construction

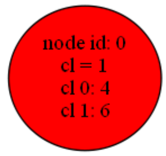

Figure 1: Tree before any split. Response of the tree is majority class cl = 1. In this node we have 4 object in class 0 and 6 object in class 1. Classification error is 4.

Create split on attribute which has the larger value of information gain.

$IG(Y, X_1) = 0.0199$

$IG(Y, X_2) = 0.1245$

First split is on X_2.

Table 1: Data with $X_2 = 0$

	X_1	X_2	cl
1	0	0	1
3	0	0	0
6	1	0	0
7	0	0	1
8	1	0	0

Table 2: Data with $X_2 = 1$

	X_1	X_2	cl
0	1	1	1
2	1	1	1
4	0	1	0
5	1	1	1
9	1	1	1

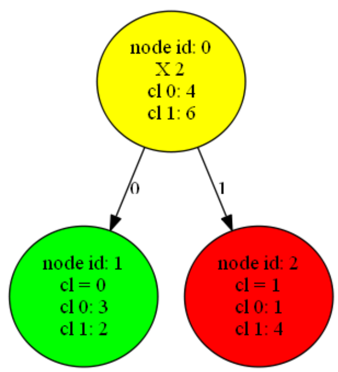

Figure 2: Tree after first split on X_2. If object on attribute X_2 has 0 then we go left (node 1) and classify all the objects to class 0 (green), because we have 3 objects from class 0 and 2 from class 1. If object has value equal 1 on attribute X_2 we go right (node 2). In this node we classify all objects to class 1 (red), because we have only 1 object in class 0 and 4 objects in class 1. The classification error of this tree is equal to 3 (sum of the errors in the leaves).

Let us firstly consider left branch where $X_2 = 0$. Now our data are in the Table 1. $IG(Y, X_1) = 0.4199$

$IG(Y, X_2) = 0.0$

Second split is on X_1.

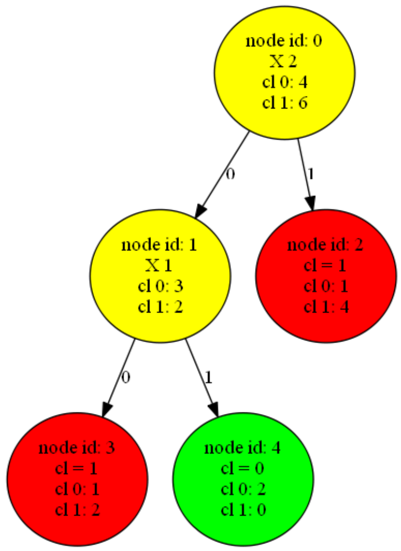

Figure 3: Tree after split on node 1 and X_1.

Table 3: Data with $X_2 = 0$ and $X_1 = 0$

	X_1	X_2	cl
1	0	0	1
3	0	0	0
7	0	0	1

Table 4: Data with $X_2 = 0$ and $X_1 = 1$

	X_1	X_2	cl
6	1	0	0
8	1	0	0

Let us consider right branch where $X_2 = 1$. Now our data are in the Table 2 .

$IG(Y, X_1) = 0.7219$

$IG(Y, X_2) = 0.0$

Split is on X_1.

Table 5: Data with $X_2 = 1$ and $X_1 = 0$

	X_1	X_2	cl
4	0	1	0

Table 6: Data with $X_2 = 1$ and $X_1 = 1$

	X_1	X_2	cl
0	1	1	1
2	1	1	1
5	1	1	1
9	1	1	1

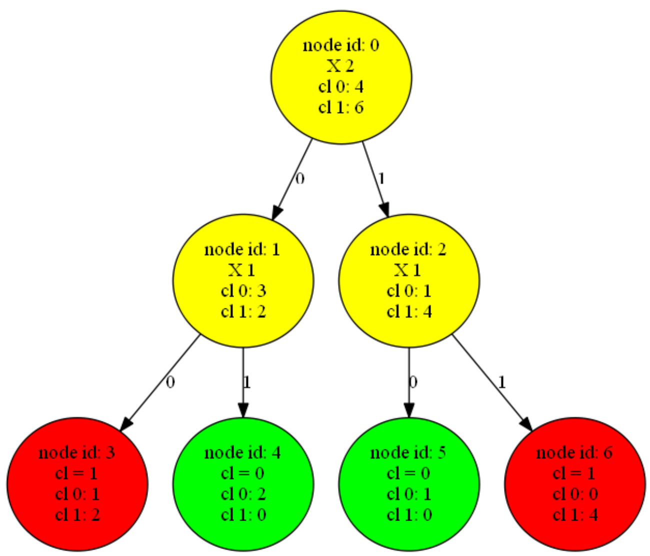

Figure 4: Tree after all splits.

2 Cost-complexity pruning

$$minimize\ cost = \frac{E}{N} + \alpha L \tag{9}$$

where:

- E - number of classification errors;
- N - number of cases in data set;
- α - complexity parameter;
- L - number of leaves.

For each possible pruned tree, we calculate a cost-complexity value and choose the tree with the minimal cost.

2.1 Example

Let's add one more parameter X_3 so now our data are shown in the Table 8 and tree is shown in the Figure 5. α parameter is set to 0.05. For this tree $E = 1$, $N = 10$, and $L = 5$, so $cost = 0.1 + 0.05 * 5 = 0.35$

Table 7: data with parameter X_3

	X_1	X_2	X_3	cl
0	1	1	1	1
1	0	0	1	1
2	1	1	1	1
3	0	0	1	0
4	0	1	1	0
5	1	1	0	1
6	1	0	0	0
7	0	0	0	1
8	1	0	0	0
9	1	1	0	1

Table 8: Cost-complexity for every prunned tree.

Figure	E	N	α	L	cost
5	1	10	0.05	5	0.35
6	2	10	0.05	4	0.4
7	1	10	0.05	4	0.3
8	2	10	0.05	3	0.35
9	2	10	0.05	3	0.35
10	3	10	0.05	2	0.4
11	4	10	0.05	1	0.45

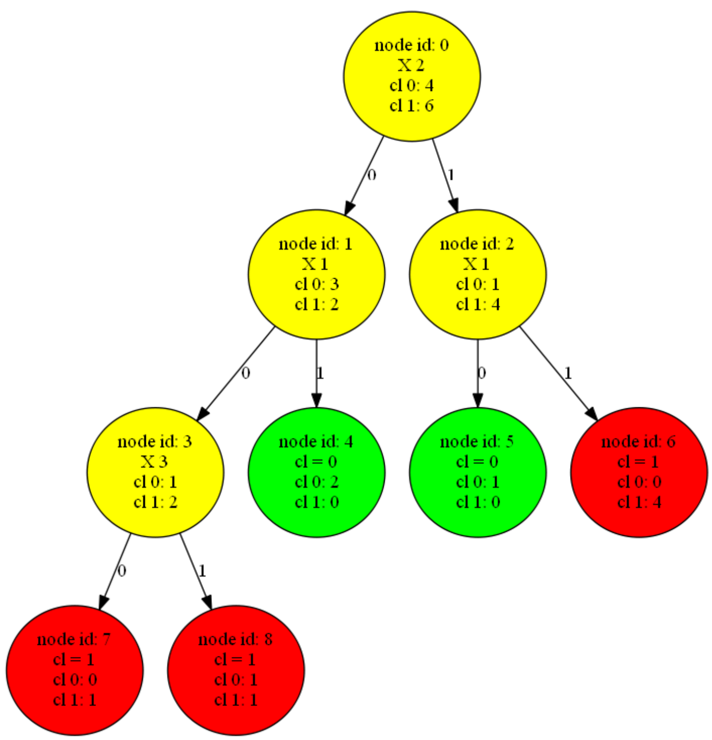

Figure 5: The whole tree without pruning.

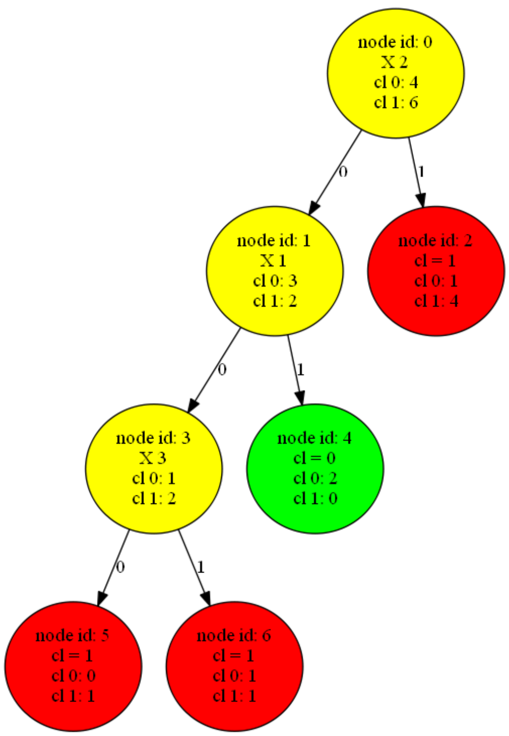

Figure 6: Tree with pruned node 2.

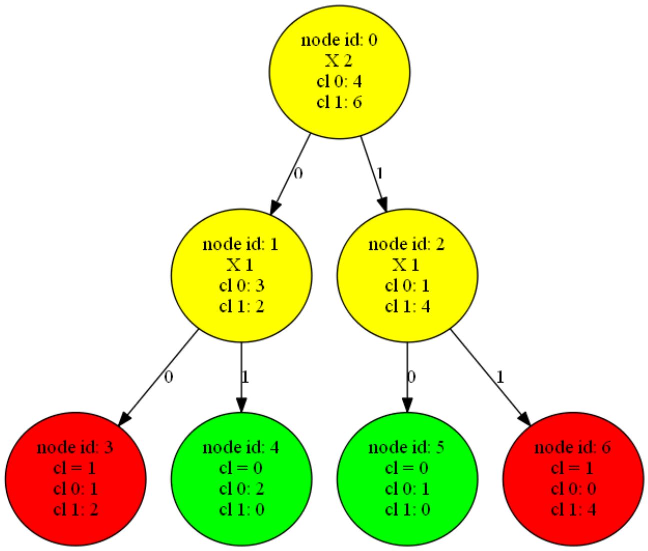

Figure 7: Tree with pruned node 3.

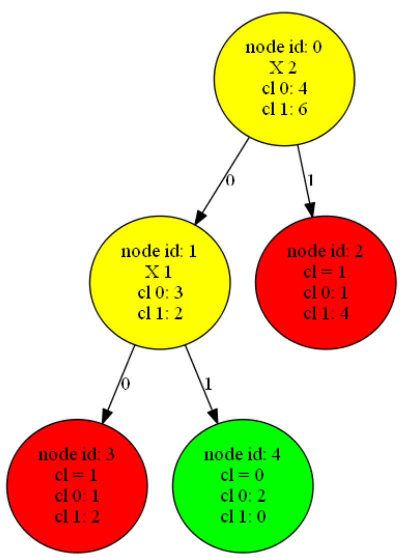

Figure 8: Tree with pruned nodes 2 and 3.

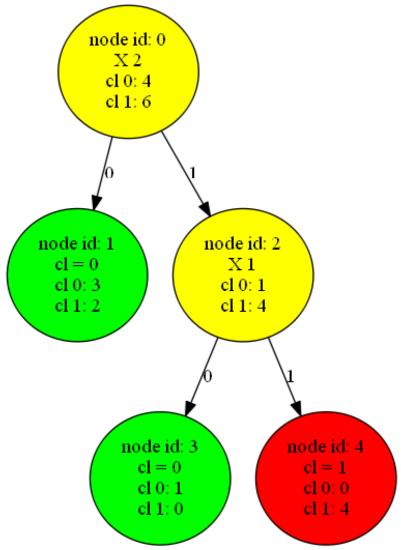

Figure 9: Tree with pruned node 1.

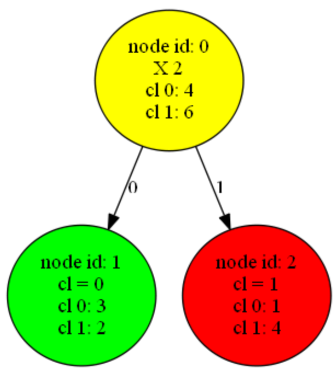

Figure 10: Tree with pruned nodes 1 and 2.

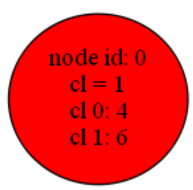

Figure 11: Tree with pruned node 0.