Assignment: CS 5785 Assignment 4 Written

Author: John Waldo

Net ID: jw922

Reflection

I thought the random forest image processing was amazing. It was incredible to see that with just 1 percent of the pixels, it could recreate much of the image. I tried running the same script with 2,5,10 percent of pixels was able to notice how much better it got. It made me realize how services such Apple FaceID could be so accurate with larger training sets.

Decision trees

Suppose we modify the tree-growing algorithm presented in class to use a new impurity function. Define $f(r)$ =min$(r, 1-r)$. Then we will define the impurity of a set of examples as:

$$I(\{y_1, ..., y_n\}) = f(p)$$

where p is the fraction of positive examples in $\{y_1, ..., y_n\}$. Let us call this the min-error impurity function.
As usual, for a split where p_1 positive and n_1 negative examples reach the left branch and p_2 positive and n_1 negative examples reach the right branch, the weighted impurity of the split will be

$$(p_1 + n_1) * f(\frac{p_1}{p_1 + n_1}) + (p_2 + n_2) * f(\frac{p_2}{p_2 + n_2})$$

a) Suppose that each branch of this split is replaced by a leaf labeled with the more frequent class among the examples that reach that branch. Show that the number of training mistakes made by this truncated tree is exactly equal to the weighted impurity given above. Thus, using the min-error impurity is equivalent to growing the tree greedily to minimize training error.

Let us begin at node 0 with, $p_0 = P$ and $n_0 = N$. Furthermore, let us define our child nodes 1, 2.

Then if we were to label each child node as a leaf node labeled by the most frequent class we would have. Node$_1$ =max(p_1, n_1) and Node$_2$ =max(p_2, n_2).

Therefore node1 would misclassify min(p_1, n_1) and node2 would misclassify min(p_2, n_2) and total training mistakes is min(p_1, n_1)+ min(p_2, n_2)

$$\begin{aligned}
mistakes &= min(p_1, n_1) + min(p_2, n_2) \\
&= \frac{p_1 + n_1}{p_1 + n_1} min(p_1, n_1) + \frac{p_2 + n_2}{p_2 + n_2} min(p_2, n_2) \\
&= (p_1 + n_1) min(\frac{p_1}{p_1 + n_1}, \frac{n_1}{p_1 + n_1}) + (p_2 + n_2) min(\frac{p_2}{p_2 + n_2}, \frac{n_2}{p_2 + n_2}) \\
&= (p_1 + n_1) min(\frac{p_1}{p_1 + n_1}, 1 - \frac{p_1}{p_1 + n_1}) + (p_2 + n_2) min(\frac{p_2}{p_2 + n_2}, 1 - \frac{p_2}{p_2 + n_2}) \\
&= (p_1 + n_1) f \frac{p_1}{p_1 + n_1}) + (p_2 + n_2) f(\frac{p_2}{p_2 + n_2})
\end{aligned}$$

b) Suppose the dataset looks like the following. There are three $\{0, 1\}$-valued attributes, and one $\{-, +\}$-valued class label y. Which split will be chosen at the root when the Gini index impurity function is used? Which split will be chosen at the root when min-error impurity is used? Explain your answers

a_1	a_2	a_3	y
0	0	0	$+$
1	1	0	$+$
0	1	0	$+$
1	0	1	$-$
0	0	1	$-$
0	1	0	$-$
1	1	0	$-$
1	1	1	$-$
1	0	0	$-$
1	1	0	$-$

(1)

If we were to split on a_1 we would get,

$$\begin{aligned}\mathbb{P}(y = +|a_1 = 0) &= 2/4\\ \mathbb{P}(y = -|a_1 = 0) &= 2/4\\[1em] \mathbb{P}(y = +|a_1 = 1) &= 1/6\\ \mathbb{P}(y = -|a_1 = 1) &= 5/6\end{aligned}$$

therefore,

$$\begin{aligned} gini &= (4/10) * (1 - (2/4)^2 - (2/4)^2) + (6/10) * (1 - (1/6)^2 - (5/6)^2)\\ &= (2/5) * (1 - 1/2) + (3/5) * (1 - 26/36)\\ &= (1/5) + (30/180)\\ &= 11/30\\[1em] min - error &= 4 * min(1/2, 1/2) + 6 * min(1/6, 5/6)\\ &= 4/2 + 6/6\\ &= 3 \end{aligned}$$

If we were to split on a_2 we would get,

$$\begin{aligned}\mathbb{P}(y = +|a_2 = 0) &= 1/4\\ \mathbb{P}(y = -|a_2 = 0) &= 3/4\\[1em] \mathbb{P}(y = +|a_2 = 1) &= 2/6\\ \mathbb{P}(y = -|a_2 = 1) &= 4/6\end{aligned}$$

therefore,

$$\begin{aligned} gini &= (4/10) * (1 - (1/4)^2 - (3/4)^2) + (6/10) * (1 - (2/6)^2 - (4/6)^2) \\ &= (2/5) * (1 - 9/16) + (3/5) * (1 - 20/36) \\ &= 6/40 + 16/60 \\ &= 5/12 \end{aligned}$$

$$\begin{aligned} min - error &= 4 * min(1/4, 3/4) + 6 * min(2/6, 4/6) \\ &= 4/4 + 12/6 \\ &= 3 \end{aligned}$$

If we were to split on a_3 we would get,

$$\begin{aligned} \mathbb{P}(y = +|a_3 = 0) = 3/7 \\ \mathbb{P}(y = -|a_3 = 0) = 4/7 \\ \\ \mathbb{P}(y = +|a_3 = 1) = 0/3 \\ \mathbb{P}(y = -|a_3 = 1) = 3/3 \end{aligned}$$

therefore,

$$\begin{aligned} gini &= (7/10) * (1 - (3/7)^2 - (4/7)^2) + (3/10) * (1 - 0 - 1) \\ &= (7/10) * (1 - 25/49) \\ &= 12/35 \end{aligned}$$

$$\begin{aligned} min - error &= 7 * min(3/7, 4/7) + 3 * min(0, 1) \\ &= 3 \end{aligned}$$

Thus, if we were splitting using gini impurity, we would split on a_1 and if we were splitting on min-impurity all three splits are equivalent.

c) Under what general conditions on $p_1, n_1, p_2,$ and n_2 will the weighted min-error impurity of the split be strictly smaller than the min-error impurity before making the split (i.e., of all the examples taken together)?

From part 1 we have that the impurity of after the split is,

$$\begin{aligned} &(p_1 + n_1) * f(\frac{p_1}{p_1 + n_1}) + (p_2 + n_2) * f(\frac{p_2}{p_2 + n_2}) \\ &= min(p_1, n_1) + min(p_2, n_2) \end{aligned}$$

And the impurity before the split is,

$$\begin{aligned} f(p) &= min(\frac{P}{P + N}, 1 - \frac{P}{P + N}) \\ &= min(\frac{P}{P + N}, \frac{N}{P + N}) \\ (P + N) * f(p) &= min(P, N) \end{aligned}$$

Therefore, the weighted impurity after the split will be strictly smaller iff,

$$\begin{aligned} min(P, N) &> min(p_1, n_1) + min(p_2, n_2) \\ & \qquad\qquad\qquad\qquad\qquad or \\ min(p_1 + p_2, n_1 + n_2) &> min(p_1, n_1) + min(p_2, n_2) \end{aligned}$$

d) What do your answers to the last two parts suggest about the suitability of min-error impurity for growing decision trees?

This suggests that min-error impurity is not a suitable choice for growing a decision tree. We want our decision trees to separate the data in minimal depth. However, as we can see from the examples above, min-error impurity tends to focus on finding splits that separate the most populated classes with highest priority. It does not always find the split that best separates the majority of the data.