Machine Learning: Exercise Set V

Georgios Manos
csd4333

19 November 2022

1 Entropy and Decision Trees (Theoretical)

First, we need to calculate the required probabilities and use those results to calculate the entropy at the root, as well as the information gained for each attribute.

For simplicity purposes, I denode as Y the virus variable ($Y = 0$ represents not virus), BL is for Body Length, bl is for bold letters and sa is for susp. address ($sa = 0$ represents susp. address = 'no').

Since BL is a continuous variable, we will use its mean as the splitting factor. The mean is equal to $\frac{1}{4}(23 + 18 + 43 + 68) = 38$. Calculating the useful probabilities:

$$P(Y = 0) = 0.75,\ P(Y = 1) = 0.25$$
$$P(BL < 38) = 0.50,\ P(BL \geq 38) = 0.50$$
$$P(bl = 0) = 0.75,\ P(bl = 1) = 0.25$$
$$P(sa = 0) = 0.50,\ P(sa = 1) = 0.50$$

and the conditional ones:

$$P(Y = 0|BL < 38) = 1,\ P(Y = 1|BL < 38) = 0$$
$$P(Y = 0|BL \geq 38) = 0.50,\ P(Y = 1|BL \geq 38) = 0.50$$

$$P(Y = 0|bl = 0) = \frac{2}{3},\ P(Y = 1|bl = 0) = \frac{1}{3}$$
$$P(Y = 0|bl = 1) = 1,\ P(Y = 1|bl = 1) = 0$$

$$P(Y = 0|sa = 0) = 0.5,\ P(Y = 1|sa = 0) = 0.5$$
$$P(Y = 0|sa = 1) = 1,\ P(Y = 1|sa = 1) = 0$$

So, the entropy of Virus at the root is:

$$H(Y) = -\sum_{i=1}^{\#classes} P(Y = c_i)\log_2(P(Y = c_i)) = -0.75\log_2(0.75) - 0.25\log_2(0.25) \approx 0.81$$

We also need to compute the conditional entropies in order to calculate the information gained for each variable, and choose the best one for the root of our tree.

$$H(Y|BL) = -P(BL < 38)[P(Y = 0|BL < 38)\log_2(P(Y = 0|BL < 38))$$

$$+P(Y = 1|BL < 38)\log_2(P(Y = 1|BL < 38))] - P(BL \geq 38)[P(Y = 0|BL \geq 38)\cdot$$

$$\cdot \log_2(P(Y = 0|BL \geq 38)) + P(Y = 1|BL \geq 38)\log_2(P(Y = 1|BL \geq 38))]$$

$$= -0.5 \cdot (1 \cdot \log_2(1) + \lim_{x \to 0} x \cdot \log_2(x)) - 0.5(0.5 \cdot \log_2(0.5) + 0.5\log_2(0.5))$$

$$\implies H(Y|BL) = 0.5$$

Also:

$$H(Y|bl) = -P(bl = 0)[P(Y = 0|bl = 0)\log_2(P(Y = 0|bl = 0))$$

$$+P(Y = 1|bl = 0)\log_2(P(Y = 1|bl = 0))] - P(bl = 1)[P(Y = 0|bl = 1)\cdot$$

$$\cdot \log_2(P(Y = 0|bl = 1)) + P(Y = 1|bl = 1)\log_2(P(Y = 1|bl = 1))]$$

$$= -0.75[\frac{2}{3}\log_2(\frac{2}{3}) + \frac{1}{3}\log_2(\frac{1}{3})] - 0.25[1\log_2(1) + \lim_{x \to 0} x \cdot \log_2(x)]$$

$$\implies H(Y|bl) \approx 0.68$$

Finally:

$$H(Y|sa) = -P(sa = 0)[P(Y = 0|sa = 0)\log_2(P(Y = 0|sa = 0))$$

$$+P(Y = 1|sa = 0)\log_2(P(Y = 1|sa = 0))] - P(sa = 1)[P(Y = 0|sa = 1)\cdot$$

$$\cdot \log_2(P(Y = 0|sa = 1)) + P(Y = 1|sa = 1)\log_2(P(Y = 1|sa = 1))]$$

$$= -0.5[0.5\log_2(0.5) + 0.5\log_2(0.5)] - 0.5[1\log_2(1) + \lim_{x \to 0} x \cdot \log_2(x)]$$

$$\implies H(Y|sa) = 0.5$$

Finally, we can calculate the information gained for each variable:

$$IG(BL) = H(Y) - H(Y|BL) \approx 0.81 - 0.5 = 0.31$$

$$IG(bl) = H(Y) - H(Y|bl) \approx 0.81 - 0.68 = 0.13$$

$$IG(sa) = H(Y) - H(Y|sa) \approx 0.81 - 0.5 = 0.31$$

Susp address and Body Length have the same information gained score, so we can choose any. I will choose Body Length for this exercise as the root of my decision tree.

Our tree now looks like this:

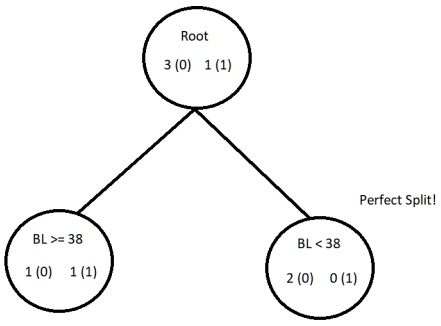

Figure 1: First part of our under-construction decision tree.

For $BL < 38$ we have a perfect split, so that node is a leaf. For $BL \geq 38$, we repeat the same procedure as before. Our table now is as follows:

	Bold letters	**susp. adress**	**Virus**
s3	0	Yes	¬Virus
s4	0	No	Virus

Figure 2: Updating our table given $BL \geq 38$.

The Bold Letters column has the same value on the whole table, so we cannot calculate the information gained from that variable (or we can assume it is 0). Hence, the variable can be discarded. If done so, we are left with just 1 variable which we can use to further split the tree. Note that if we had chosen susp. address on the previous step, bold letters wouldn't be all zero's and we would have to repeat the procedure anyways ☺.

We reach our final form for our tree:

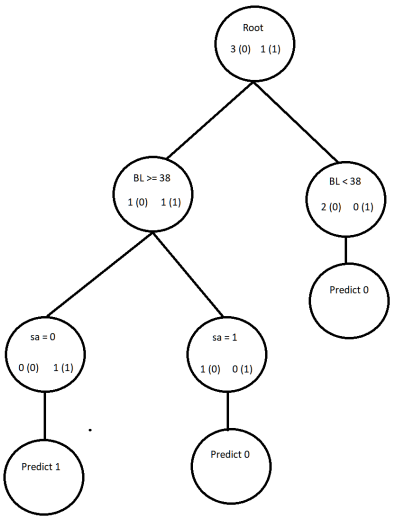

Figure 3: Fully constructed decision tree.

Using it to predict on our test set, we have the following results:

	Body Length	**Bold letters**	**susp. adress**	**Virus**	Predictions
s5	20	1	Yes	Virus	¬Virus
s6	25	1	No	¬Virus	¬Virus
s7	60	0	Yes	Virus	¬Virus
s8	35	0	No	¬ Virus	¬Virus

Figure 4: Prediction results from our decision tree.

All samples ended being classified as not virus. The accuracy is 0.5, pretty much the same with the random classifier.

2 Random Forest (Programming)

2.1 Part A

The code once again is configurable on top level. One can change the global variables to change the program's behavior. To produce the bonus part results simply change MAX_FEATURES to 'None', MIN_SAMPLES_LEAF_VALUES to '[1]' and PERMUTATIONS to 'False'. Feel free to comment out lines 104 and 105 as they set limits to the plotted graph, done for illustration purposes.

I'm also comparing my results to SKLearn's Random Forest Classifier and they both achieve similar accuracy on the same test set. Finally, there was no need to set 'random_state' to 'None' as it is set by default on SKLearn's Decision Tree Classifier implementation.

2.2 Part B

The discrepancy between the average tree accuracy and the forest accuracy slightly decreases as the parameter value changes from 1 to 10. Also note that the distribution on the later is more similar to a Gaussian than the 1st one. This is due to that each tree is more constrained on its output as it has more samples on its leaves and on large value cases could underfit the data.

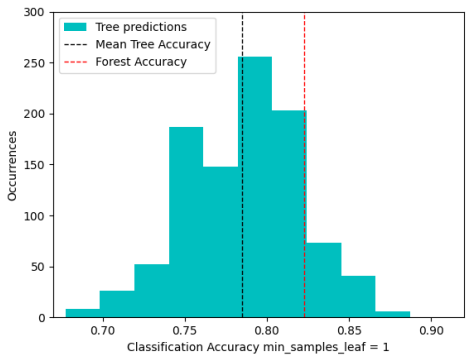

Figure 5: Decision Trees average accuracy vs Random Forest accuracy for min_samples_leaf = 1

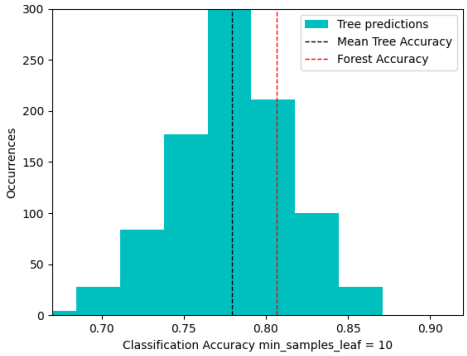

Figure 6: Decision Trees average accuracy vs Random Forest accuracy for min_samples_leaf = 10

2.3 Bonus

Comparing with the accuracy obtained from a basic decision tree and my Random Forest algorithm produces the following result. The accuracy difference is actually much smaller than previously. This happens because the trees are more similar, considering the same amount of variables and trained on the exact same samples without permutations.

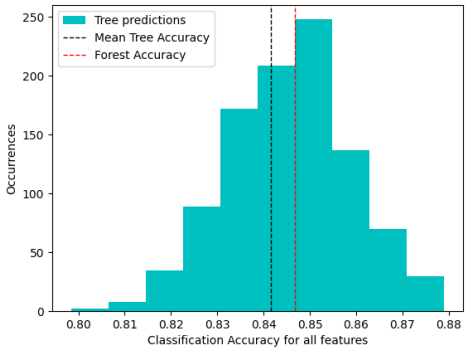

Figure 7: Basic Decision Trees average accuracy vs Random Forest accuracy