Homework 1

Name: Mikhail Podvalkov

Date:10/07/21

Calculating Information Gain

Taking first attribute as outlook-values are sunny, rainy, overcast

$$Entropy = \frac{-y}{n+y}\log_2\frac{y}{y+n} - \frac{n}{n+y}\log_2\frac{n}{y+n} = \frac{-9}{5+9}\log_2\frac{9}{9+5} - \frac{5}{5+9}\log_2\frac{5}{9+5} = 0.94$$

Outlook	Play Golf
Sunny	Yes
Sunny	Yes
Sunny	No
Sunny	Yes
Sunny	No

Outlook	Play Golf
Rainy	No
Rainy	No
Rainy	No
Rainy	Yes
Rainy	Yes

Outlook	Play Golf
Overcast	Yes
Overcast	Yes
Overcast	Yes
Overcast	Yes

Outlook	n	Y
Sunny	2	3
Rainy	3	2
Overcast	0	4

$$Entropy(Sunny) = \frac{-2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$$

Entropy for Rainy the same and Entropy for overcast =0

$$Entropy(Outlook) = \frac{5}{14} * 0.97 + \frac{5}{14} * 0.97 + 0 = 0.69$$

$$Gain = 0.94 - 0.69 = 0.25$$

Temperature

Temperature	n	y
Hot	2	2
Mild	2	4
Cool	1	3

$$Entropy(Temperature) = \frac{4}{14} * 1 + \frac{6}{14} * 0.92 + \frac{4}{14} * 0.81 = 0.91$$

$$Gain = 0.94 - 0.91 = 0.03$$

Humidity

Humidity	n	y
High	4	3
Normal	1	6

$$Entropy(\text{Humidity}) = \frac{7}{14} * 0.98 + \frac{7}{14} * 0.59 = 0.78$$

$$Gain = 0.94 - 0.78 = 0.16$$

Windy

Windy	n	y
False	2	6
True	3	3

$$Entropy(\text{Windy}) = \frac{8}{14} * 0.81 + \frac{6}{14} * 1 = 0.89$$

$$Gain = 0.94 - 0.89 = 0.05$$

The Highest gain is Outlook.

Decision Trees

For each classification task, what is your learned decision tree's accuracy over the training set? Over the test set? Suggest why these accuracies might differ in the ways you observe.

The accuracy is 49% and 51% This accuracy might be different because we use different feature in entropy and it can influence on our tree.

In the above pruning step, you used an independent set of pruning data rather than the test data. Explain why this is better than using the test data for pruning if we want the best possible estimate of the accuracy of our final decision tree. (Feel free to conduct the experiment to find out what happens if you use the test data for pruning.)

We use separate data because the decision tree need to be generalized the data so we if we use test data accuracy become 100%/

Suppose we want to optimize the accuracy of the learned decision tree and are willing to forgo having a precise estimate of its actual accuracy. How would you recommend splitting the available examples into training, pruning, and testing sets in this case?

I recommend to split it in 70% percent is training data, 20% percent is purring data and 10% test data.