EECS 440 Machine Learning Programming Problem 1 Writeup

Sixiao Zhang sxz603, Jianzhe Zhang jxz851, Junyi Zheng jxz990

(a) What is the CV accuracy of the classifier on each dataset when the depth is set to 1?

When the depth is set to 1 and with information gain, the CV accuracy of the classifier on each dataset is shown in Table. 1.

Table. 1 CV accuracies of each dataset with depth=1 and information gain

	Voting	Spam	Volcanoes
Accuracy	0.9887	0.6639	0.7243

(b) For *spam* and *voting*, look at first test picked by your tree. Do you think this looks like a sensible test to perform for these problems? Explain.

The first test of spam is OS. we don't think OS is a sensible test to classify whether an email is spam or ham. Every OS can send spam emails. There is no evidence that spam emails come more from a particular OS.

The first test of voting is Repealing-the-Job-Killing-Health-Care-Law-Act. Because we are all international students, we don't know much about the perspective of the Democrats and the Republicans. But in our view, we think this could be a sensible test because whether to repeal the law influences the benefit of different groups of people, and the Democrats and the Republicans might be representatives of the different groups, thus they would vote opposite to the other party.

(c) For *volcanoes* and *spam*, plot the CV accuracy as the depth of the tree is increased. On the x-axis, choose depth values to test so there are at least five evenly spaced points. Does the accuracy improve smoothly as the depth of the tree increases? Can you explain the pattern of the graph?

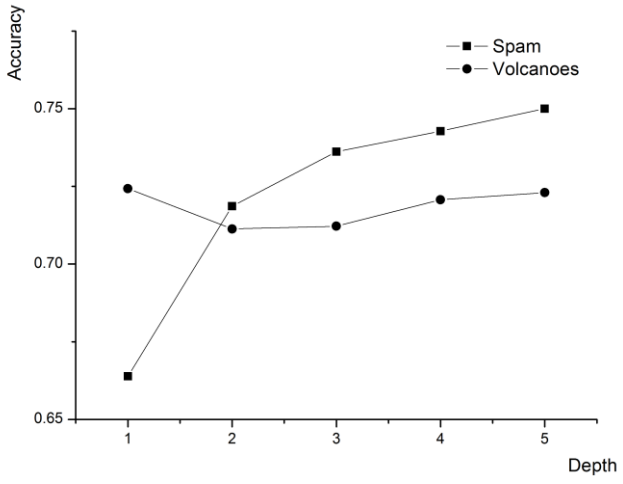

Fig. 1 CV accuracy for different depths of volcanoes and spam

We use depths 1, 2, 3, 4, 5 to test the CV accuracy for information gain of volcanoes and spam. The results are shown in Fig. 1.

The accuracy of volcanoes has no significant improvement as the tree goes deeper, whereas the accuracy of spam has significant improvement. Actually, the accuracy of volcanoes is always 0.7230 as the tree goes deeper.

The reason why there is no improvement in volcanoes is that the first attribute, image_id, contributes a lot to the classification, which means that the image_id can classify the dataset well enough that there are only a few pieces of data that are misclassified.

For spam, the case is that no attribute can classify the dataset very well. It needs to consider several attributes together to make a correct prediction. So when the tree goes deeper, there are more attributes taken into consideration, thus the accuracy increases significantly.

(d) Pick 3 different depth values. How do the CV accuracies change for gain and gain ratio for the different problems for these values?

We use the volcanoes set to test the CV accuracies for gain and gain ration. The results are shown in Table. 2.

Table. 2 CV accuracies for gain and gain ratio with 3 depth values on volcanoes

	1	3	5
Information gain	0.7243	0.7122	0.7230
Gain ratio	0.7243	0.7122	0.7230

Table. 2 shows that the CV accuracies for gain and gain ratio are the same in each depth.

(e) Compare the CV accuracies and the accuracy on the full sample for depths 1 and 2. Are they comparable?

We use the volcanoes set to test the accuracies. The results are shown in Table.3.

Table. 3 Accuracies for CV and full sample for depth 1 and 2 on volcanoes set

	1	2
CV	0.7243	0.7113
Full sample	0.7297	0.8597

From Table.3, we can see that, the accuracies of CV and full sample are nearly the same when the depth is 1. But when the depth becomes 2, the accuracy of full sample increases a lot whereas the accuracy of CV has not increased. This is because when the depth increases, the network that is trained and tested on full sample turns out to be memorization. It memorizes better when the tree is deeper, so the accuracy increases significantly.

Other findings:

1. Time & memory consuming

When our code runs on the voting set, it finishes in a second. When it runs on the volcanoes set, it takes several minutes to build a tree. When it runs on the spam set, it spends around two hours to build a tree when the depth is set less than 5.

As we see, the voting set has 11 discrete attributes, and 440 instances. The volcanoes set has 226 attributes, and 2231 instances. The spam set has 19 attributes, and 74739 instances.

According to above numbers, the time consuming increases significantly with the number of instances increases. Also, the memory consuming increases as it has to save more instances at one time.

2. Code complexity

During training, most of the time is consumed in the continuous_data() in IG_v5.py. There is a nested for loop in this function, which is the mean reason why the time needed when training the spam set increases significantly compared to the other two sets.