Faculty of Engineering
Department of Systems Design Engineering

SYDE 675: Pattern Recognition

Assignment 2

Prepared by
Rahij Imran Gillani
rgillani@uwaterloo.ca

University of Waterloo
200 University Avenue West
Waterloo, Ontario, Canada

TABLE OF CONTENTS

Question 1

Suppose that instead of selecting a node using information gain (IG) in a binary decision tree, we select a node randomly from nodes with IG>0:

a) Show that each leaf of the tree contains at least one training data.

As we will be selecting a node randomly with IG> 0, when the data is split into two parts there must exist at least one training data in the two splits. Otherwise a split won't occur and a leaf won't form.

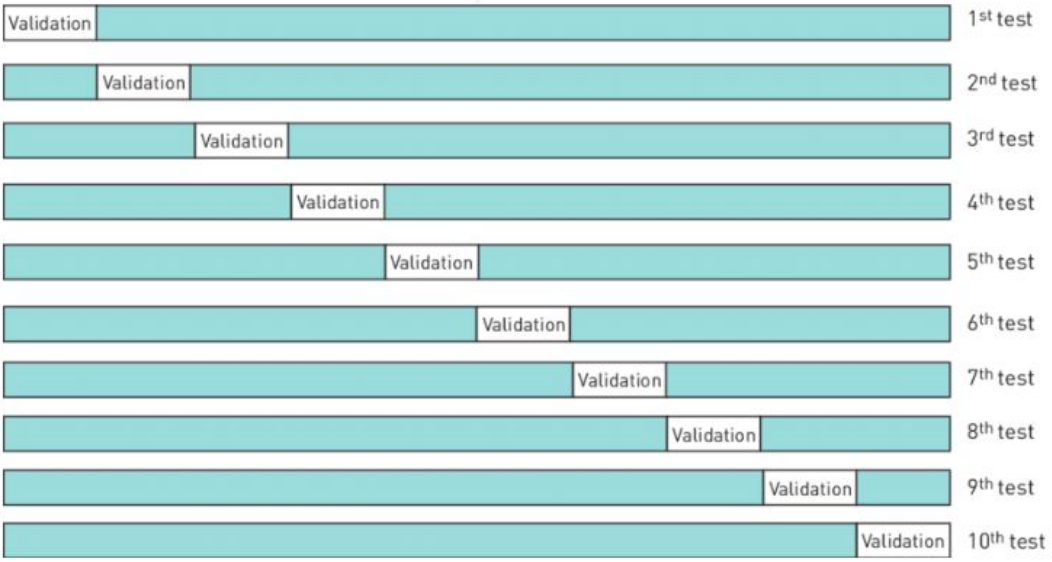
b) If we have n training data, what is the maximum number of leaves in the constructed decision tree? Compare the result with the state that we used IG for selecting nodes.

For n training data, the maximum number of leaves possible would be n leaves in the case when every training data has a unique class label. If we use IG for selecting nodes, the number of leaves will remain the same as every training data will have a different label and it won't be possible to reduce that.

Question 2

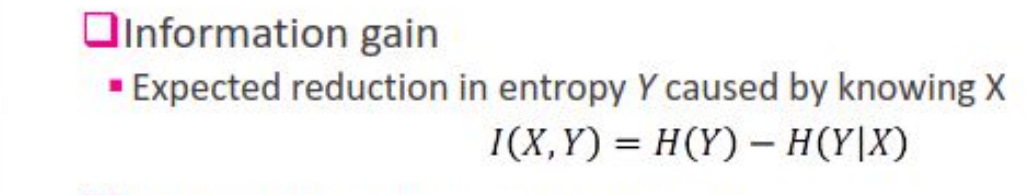
Part (A)

For this question we construct a ID3 Decision Tree Classifier for 2 datasets, namely the categorical 'Tic-Tac-Toe Endgame" and the continuous "Wine dataset".
After creating the Classifier, we were to perform the 10-times-10-fold cross validation, and based on that we were to calculate the mean accuracy and variance.



10-Fold Cross Validation

The **ID3 Function** in my code implements the ID3 Algorithm based on the **Information Gain** method, which follows the following equation:



Entropy:

$$H(X) = -\sum_{x_i \in X} P(x_i) \log_2 P(x_i)$$

Mean Tic-Tac Accuracy = 0.8479692982456141
Mean Tic-Tac Variance = 0.001625573128270237

BEST CONFUSION MATRIX TIC-TAC

	+ive	-ive
+ive	61	3
-ive	2	30

Analysis:
The accuracy of the Tic-Tac dataset after averaging 100 trees is 84.79%, and the best confusion matrix has an accuracy of 94.79%. There are more False negatives than there are False positives.

Mean Wine Accuracy = 0.9288235294117647
Mean Wine Variance = 0.0042920628817975995

BEST CONFUSION MATRIX WINE

	1	2	3
1	9	0	0
2	0	4	0
3	0	0	4

Analysis:
The accuracy of the Wine dataset after averaging 100 trees is 92.88%, and the best confusion matrix has an accuracy of 100%.

Part (B)

The **ID3_gain_ratio Function** in my code implements the ID3 Algorithm based on the **Gain Ratio** method, which follows the following equation:

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)}$$

$$SplitInformation(S,A) = -\sum_{v \in Values(A)} \frac{|S_v|}{|S|} \log_2 \frac{|S_v|}{|S|}$$

Mean Tic-Tac Accuracy Gain Ratio = 0.8628980263157895
Mean Tic-Tac Variance Gain Ratio = 0.0014552060862476918

BEST CONFUSION MATRIX TIC-TAC Gain Ratio

```
          +ive  -ive
+ive [[62    2]
-ive  [ 2   30]]
```

Mean Wine Accuracy Gain Ratio = 0.937875816993464
Mean Wine Variance Gain Ratio = 0.0031437684224016414

BEST CONFUSION MATRIX WINE Gain Ratio

```
     1  2  3
1 [[ 3  0  0]
2  [ 0 10  0]
3  [ 0  0  5]]
```

Analysis:
The accuracy of both, the Tic-Tac dataset and the Wine dataset is better using Gain Ratio instead of IG, but only just. The results are similar because all the attributes have similar number of values, hence Gain Ratio has very little benefit.

Question 3

Part (A) Effect of Attribute Noise

Wine Data

CvC

ACCURACY	VARIANCE
0.94	0.003550882139348114

DvC

NOISE	ACCURACY	VARIANCE
5%	0.9270588235294117	0.005102926224956213
10%	0.9236274509803921	0.0044185259942757064
15%	0.9116666666666666	0.005634895766585499

CvD

NOISE	ACCURACY	VARIANCE
5%	0.9254901960784314	0.0038307915758896143
10%	0.9113725490196078	0.005022700670682215
15%	0.9048039215686274	0.0044501377675253105

DvD

NOISE	ACCURACY	VARIANCE
5%	0.9131045751633987	0.004618756674783204
10%	0.8891830065359477	0.006484492075697381
15%	0.8733006535947713	0.006116419966679483

Tic-Tac Data

CvC

ACCURACY	VARIANCE
0.8646524122807018	0.0012441226278758848

DvC

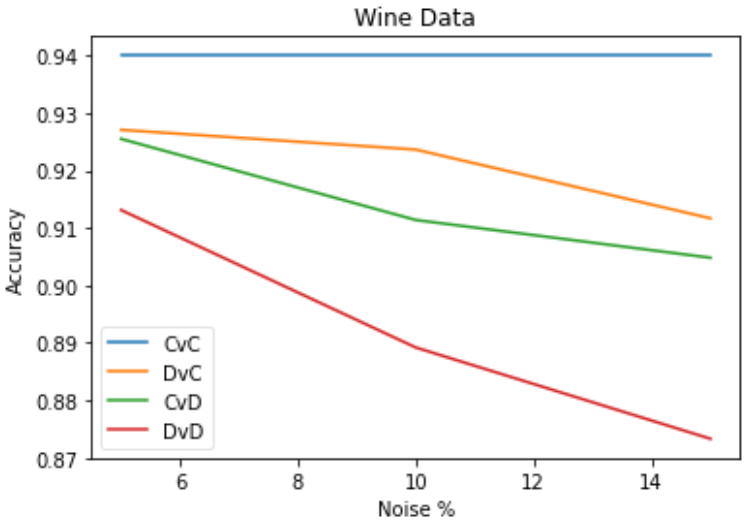
NOISE	ACCURACY	VARIANCE
5%	0.8481524122807017	0.0015761550128404896
10%	0.8308311403508772	0.0013402629607186822
15%	0.8110942982456141	0.001735414112996307

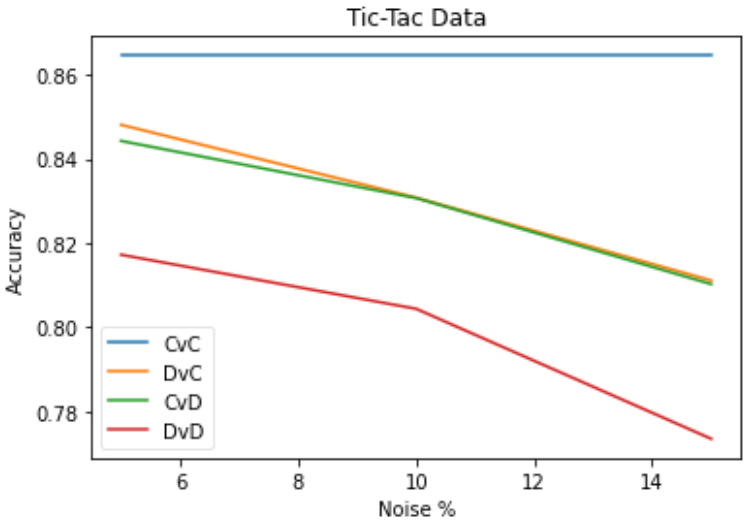
CvD

NOISE	ACCURACY	VARIANCE
5%	0.8443267543859649	0.001346434869382887
10%	0.8307127192982456	0.00123602562807787
15%	0.8102456140350878	0.0013160845933364112

DvD

NOISE	ACCURACY	VARIANCE
5%	0.8172291666666667	0.0014211763848010928
10%	0.8043618421052632	0.0019533615102723923
15%	0.7734232456140351	0.001717142000423207





Analysis:

For both plots, the DvD accuracy is lowest which intuitively also makes sense as we have introduced noise to both the training and testing dataset;hence, it would be better if we were to handle the noise in someway in a dataset in which both the training and testing dataset are noisy.

Moreover for both the Wine and Tic-Tac dataset the DvC and CvD accuracies are relatively closer to each other but CvD has a lower accuracy than DvC in both. Which makes sense because as the testing set is smaller, any noise introduced into that will have a greater effect.

Part (B) Effect of Class Label Noise

Wine Data

No Noise

ACCURACY	VARIANCE
0.9497385620915032	0.0025493998034943834

Contradiction

NOISE	ACCURACY	VARIANCE
5%	0.9212745098039216	0.003950637575291555
10%	0.9066013071895425	0.004982737408688966
15%	0.8748366013071895	0.0082037891409287

Misclassification

NOISE	ACCURACY	VARIANCE
5%	0.8862745098039215	0.00658165662779273
10%	0.8249346405228758	0.008413912384125762
15%	0.7833006535947712	0.008315539963261991

Tic-Tac Data

No Noise

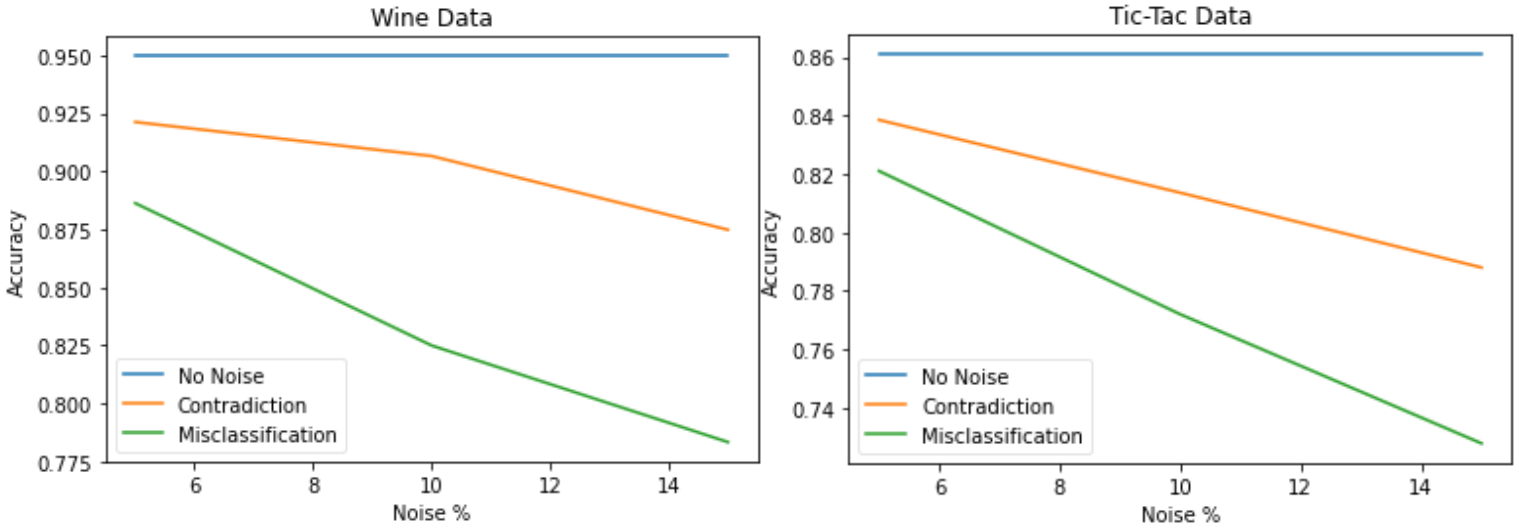
ACCURACY	VARIANCE
0.860875	0.0014072864679516772

Contradiction

NOISE	ACCURACY	VARIANCE
5%	0.838469298245614	0.0017904727223761164
10%	0.8135668859649123	0.0024816621111784396
15%	0.7879857456140351	0.001652182223135965

Misclassification

NOISE	ACCURACY	VARIANCE
5%	0.8210054824561404	0.002247193668965451
10%	0.7719912280701754	0.0020712869152046784
15%	0.7278267543859649	0.0021396940837565396



C) How do you explain the effect of noise on ID3 when there is not any pruning variable?

Noise has a big effect on ID3 without any pruning variable as the tree tries to fit to the data available; hence, a little bit of noise has a drastic effect on the accuracy.

D) In comparison with attribute noise and class noise, which is more harmful? Why?

For both plots we can see that Misclassification is **more harmful** than Contradiction, which makes sense as with Contradiction there is a chance that the true class label may be chosen when a tree is being formed, but with misclassification the correct label is replaced with the wrong one, hence the correct one may never be chosen now.