CSE 472

Assignment 1
Decision Tree Learning for Cancer Diagnosis

Submitted By:

Mohammad Imam Hossain

Id: 1105031

L/T: 4/2

Dept. CSE, BUET.

Outputs:

Evaluation Criterion	Accuracy	Precision	Recall
Information Gain	0.959	0.954	0.929

Question 1:

Why are we dividing the dataset 80% into training and 20% into test data rather than using 100% data for training?

Answer

To avoid **overfitting** problem.

if we supply too much (100%) data as training set, then the model will actually be created perfectly, but just for that data. But our objective here is to use the model to predict the future unknowns. This is why we create a test set and after creating the model, we check to ensure that the accuracy of the model we built doesn't decrease with the test set. This ensures that our model will accurately predict future unknown values.

Question 2:

Do you see evidence of overfitting in some experiments? Explain.

Answer

Given a hypothesis space H, a hypothesis hЄH is said to overfit the training data if there exists some alternative hypothesis h'ЄH, such that h has smaller error than h' over the training examples, but h' has smaller error than h over the entire distribution of instances.

We know that random noise in the training example creates overfitting but in our case there is no scope of that random noise.
if we build the decision tree twice on the same training set then both the decision trees

will be same. So there is no scope of overfitting here.

Another reason of overfitting is because of using a small data set as training set. Here the algorithm will have greater control over this small dataset and it will make sure it satisfies all the data points exactly.
In our case, we are using 80% of the input data as training data set. As a result, the algorithm is forced to generalize and come up with a good model that suits most of the points.

But if we compare two experiments on different training data set then we will see that there exists some cases when the accuracy is high around 97% for my case and some experiments when the accuracy is low around 93%. So I can say there exists some hypothesis that models well than the others.