24-787

Assignment 1

Programming Question Report

1. **Can it learn on all oracles or only a few?**

   Yes, it can learning on different data set. Because I set my loop limit for input to be varied with the size of data, so I can fit in different dataset.

2. **How many examples does it usually need to be given in order to learn well?**

   It can't be given too many training example because it will generate more error rate for some reasons like overfitting. Learned by this code, I think it's better to have less than 100 training examples to have more accurate prediction.

3. **Did you implement any innovative techniques beyond the bare requirements for the algorithm?**

   No, I didn't. But I think the error rate can be reduced by pruning the tree node from the bottom.

4. **Are you satisfied with the results?**

   Yes, it runs pretty well and get the error rate successfully. And it also makes me understand more about the relation between the accuracy of prediction and the number of training data.

5. **What other modifications could you make to improve algorithm performance?**

   With ID3, I think it will results in too much overfitting if we have too many features. It should properly be used with pruning nodes, which means we should reduce some features to make the train data and true value get closer.

6. **Extra Credit: Implement a measure to combat overfitting; demonstrate an improvement in performance and explain why your approach works.**

   To combat overfitting, the most direct way is to take off some of the features by pruning node from the bottom. The reason is that the training sample is part of the real data, and they are not totally the same. If we utilize too many features, we will make their difference higher.

Oracle	#Training sample	#Test sample	Error rate (%)
1	50	100	15.00
2	50	100	3.00
3	50	100	5.00
5	50	100	9.00
7	50	100	0
8	50	100	0
12	50	100	0
15	50	100	12.00

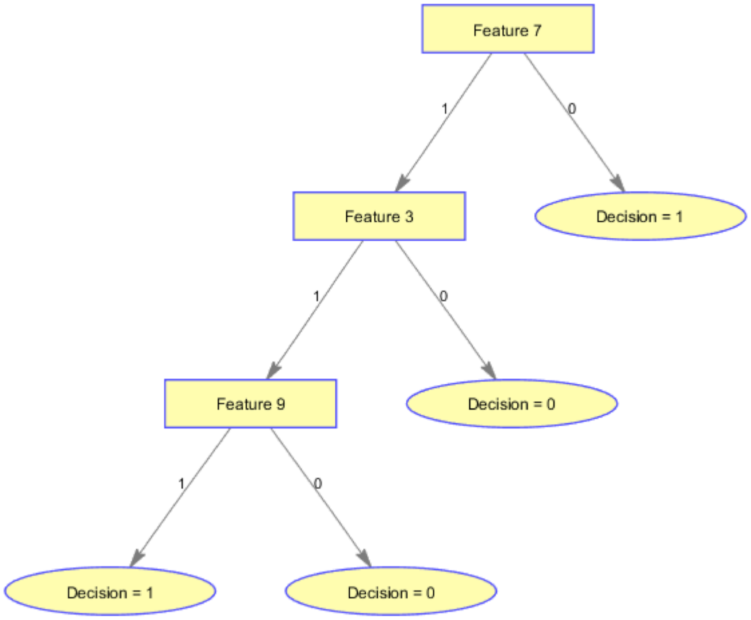
Error rate varies with different oracles

Oracle	#Training sample	#Test sample	Error rate(%)
3	25	100	6.00
3	50	100	5.00
3	75	100	5.00
3	100	100	19.00
3	125	100	19.00
3	150	100	19.00
3	175	100	19.00
3	200	100	19.00

Error rate varies with different number of training sample (oracle3)

Oracle	#Training sample	#Test sample	Error rate(%)
5	25	100	13.00
5	50	100	9.00
5	75	100	7.00
5	100	100	2.00
5	125	100	2.00
5	150	100	2.00
5	175	100	31.00
5	200	100	11.00

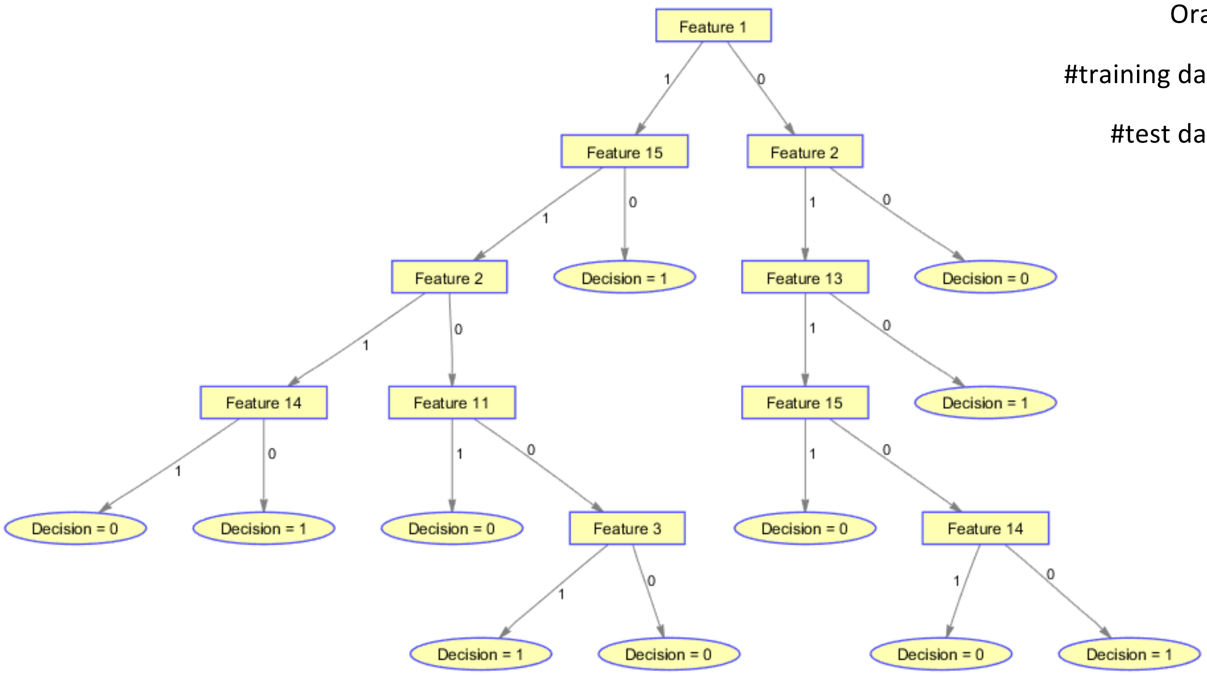
Error rate varies with different number of training sample (oracle5)



Oracle = 3

#training data=50

#test data=100



Oracle = 5

#training data=100

#test data=100