CSE 472
Assignment 1: Decision Tree Learning for Cancer Diagnosis

Kazi Ashik Islam
1205007

March 12, 2017

Performance Measures

Test Data	Accuracy(%)	TPR(%)	TNR(%)	PPV(%)	NPV(%)	FPR(%)	FNR(%)	FDR(%)	F1
part 1	93.99	85.96	100	100	90.48	0	14.04	0	0.9245
part 2	94.74	95.52	93.94	94.12	95.38	6.06	4.48	5.88	0.9481
part 3	93.99	91.11	95.45	91.11	95.45	4.55	8.89	8.89	0.9111
part 4	96.99	92.86	98.1	92.86	98.1	1.9	7.14	7.14	0.9286
part 5	97.81	97.14	98.04	94.44	99.01	1.96	2.86	5.56	0.9577
Average	95.5	92.52	97.11	94.51	95.68	2.89	7.48	5.49	0.934

Question Answers

- **Why are you using cross validation? Do the dataset justify it?**

  To create and test the decision tree we need train data as well as test data. To use all the available data effectively we use them as both train and test data. This is achieved by cross validation. In this assignment in particular, we divided the data in five parts, used one part of the data as test data, and the rest as train data. The process is repeated for each part of data.

  We observed different performance measure values for different test data. So the dataset justifies cross validation.

- **Besides accuracy, which of the criteria mentioned above should be used in cross validation for the given data set? Explain.**

  The *F1* score and *FNR* should be used.

  A good classifier should have high *PPV(Positive Predictive Value/ Precision)* and high *TPR(True Positive Rate/Sensivity/Recall/Probability of Detection)*. But maximizing both at the same time is not possible, as maximizing one reduces the other. So, their harmonic mean is a good performance measurement, which is called the *F1* score.

Also, in this particular problem we need to minimize false negative errors because that can be fatal for the patients. So, FNR should also be used in cross validation.