Machine Learning Assignment 3

Omkar Nibandhe

Q1 Train decision tree without pruning and observe model overfitting.

1. Size of the tree: **442**
2. Training error $error_D = \frac{incorrectly\ classified\ instances}{Total\ number\ of\ instances}$

$$= \frac{30}{30 + 5670} = \frac{30}{5700} = \mathbf{0.005263}$$

3. Test error $error_S = \frac{incorrectly\ classified\ instances(r)}{Total\ number\ of\ instances(n)}$

$$= \frac{456}{456 + 1968} = \frac{456}{2424} = \mathbf{0.188118}$$

4. 95% confidence interval of the true error on future data

$$error_D = \ error_S \mp 1.96 * \sqrt{(error_S * (1 - error_S))/n}$$

$$= 0.188118 \ \mp 1.96 * 0.00793$$

$$= 0.188118 \ \mp 0.0155579$$

Lower bound = 0.1725601

Upper bound = 0.2036759

Q2 Decision pruning and model selection by cross validation.

Range for MinNumObj = 1 to 5

Range for NumFolds = 2 to 6

MAE value(%error)	**m = 1**	**m = 2**	**m = 3**	**m = 4**	**m =5**
N = 2	0.0454(3.0877)	0.0475(3.1754)	0.049(3.1404)	0.0509(3.1404)	0.0537(3.3509)
N =3	0.0476(2.8947)	0.0488(3)	0.0499(3.0351)	0.0511(3.1053)	0.0531(3.2632)
N =4	**0.0444(2.8246)**	0.0455(2.8947)	0.0473(2.9825)	0.0496(3.2281)	0.0515(3.3158)
N= 5	0.0455(2.9474)	0.0465(3.0351)	0.0482(3.0877)	0.0501(3.1754)	0.0511(3.193)
N= 6	0.0495(3.1579)	0.0496(3.1228)	0.0511(3.1754)	0.0518(3.2105)	0.0534(3.2982)

1. Parameter setting used to retrain a j48 tree on entire training data:
   MinNumObj = 1, NumFolds = 4, reducedErrorPrunning = True, Unprunned = False, numDecimalPlaces = 2, seed = 1, batchSize = 100.
2. Size of tree = **122** and number of leaves = 73
3. Training error $error_D = \frac{incorrect\ classified\ instances}{Total\ number\ of\ instances} = \frac{105}{105+5595} = \frac{105}{5700} = \mathbf{0.018142}$
4. Test error $error_S = \frac{incorrect\ classified\ instances(r)}{Total\ number\ of\ instances(n)} = \frac{481}{481+1943} = \frac{481}{2424} = \mathbf{0.19843}$
5. The 95% confidence interval of the true error on future data

$$error_D = \ error_s \mp 1.96 * \sqrt{(error_s * (1 - error_s))/n}$$

$$= 0.19843 \ \mp 1.96 * \sqrt{0.19843 * (1 - 0.19843)/2424}$$

$$= 0.19843 \ \mp 1.96 * 0.0081$$

$$= 0.19843 \ \mp 0.015876$$

Upper bound = 0.214306

Lower bound = 0.182554

Q3 Compare learning algorithms by cross validations.

Average error of NBC models resulted in 10-fold cross validation:$\frac{1410}{5700} * 100$ = **24.7368% (0.2418 MAE).**

J48, 10-fold cross validation error: $\frac{\mathbf{161}}{\mathbf{5700}} * \mathbf{100}$ =2.8245% (0.0444 MAE)

Comparing results with J48, J48 is better as there is significant difference in average error of 10-fold cross validation.

1. Finally selected model is J48,
   Test error $error_S = \frac{incorrectly\ classified\ instances(r)}{Total\ number\ of\ instances(n)} = \ \mathbf{0.19843}$
2. Test error of other model, NBC is
   $$error_S = \frac{incorrectly\ classified\ instances(r)}{Total\ number\ of\ instances(n)} = \frac{664}{664 + 1760} = \frac{664}{2424} = \mathbf{0.27392}$$
3. Yes, as per selection of J48 the average error is low and hence prediction of the model will work more efficiently than NBC for independent data set.