CO 544 Machine Learning and Data Mining

Lab 01

E/15/373

8. Note down the results in the below table.

	Correctly classified instances	Incorrectly classified instances
Training Set	143	12
Cross validation(10 folds)	130	25

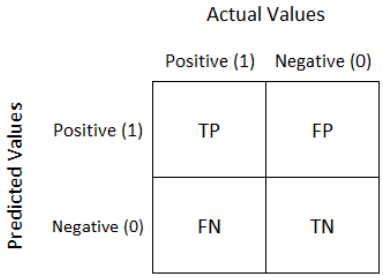
9. Interpret the results from the 'confusion matrix' in the classifier output.

True Positive (TP): Observation is positive (Die), and is predicted to be positive (Die).

False Negative (FN): Observation is positive (Die), but is predicted negative (Live).

True Negative (TN): Observation is negative (Live), and is predicted to be negative (Die).

False Positive (FP): Observation is negative (Die), but is predicted positive (Live).



Confusion matrix (Training set)

a	b	<-- classified as
22	10	a = DIE
2	121	b = LIVE

Here the confusion matrix shows us that True Positives of training set is 22 instances, False Positive is 10 instances, False Negative is 2 instances and True Negative is 121 instances so we can calculate the accuracy of this training set as

$$\text{accuracy} = \frac{True\ Positive + True\ Negative}{True\ Positive + \text{False Negative} + \text{False Positive} + True\ Negative}$$

$$= 92.581\%$$

Confusion matrix (Cross-validation (10 Folds))

a	b	<-- classified as
14	18	a = DIE
7	116	b = LIVE

Here the confusion matrix shows us that True Positives of training set is 14 instances, False Positive is 18 instances, False Negative is 7 instances and True Negative is 116 instances so we can calculate the accuracy of this training set as

$$\text{accuracy} = \frac{True\ Positive + True\ Negative}{True\ Positive + \text{False Negative} + \text{False Positive} + True\ Negative}$$

$$= 83.871\%$$

When using cross validation the accuracy decreased because we could obtain more realistic output while using cross validation. Because cross validation divide the data set into n number of parts, run the algorithm and take the average of it so the accuracy shown here is more realistic than the usual training set.

10. Change the parameters as below and compare the results with the results of the above model with default values. Confidence factor: 0.5 and Min number of folds: 2

- After changing Confidence factor to 0.5 and Min number of folds: 2 (Training set)

  Correctly classified instances: 149

  Incorrectly classified instances: 6

<u>Confusion matrix</u>

a	b	<-- classified as
28	4	a = DIE
2	121	b = LIVE

$$\text{accuracy} = \frac{True\ Positive + True\ Negative}{True\ Positive + \text{False Negative} + \text{False Positive} + True\ Negative}$$

$$= 96.129\%$$

So the accuracy of output is increased when compared to Training set default values. Error percentage also decreased where in default values the mean absolute error is 0.1272 and when the values are changed the error is 0.0836. Size of tree increased when the values are changed, in default values the size of the tree is 21 and number of leaves are 11 after making the confidence factor to 0.5 and min number folds to 2 the size of tree increased to 31 and no of leaves increased to 16.

- After changing Confidence factor to 0.5 and Min number of folds: 2 (Cross-validation (10 Folds))

Correctly classified instances: 126

Incorrectly classified instances: 29

Confusion matrix

a	b	<-- classified as
18	14	a = DIE
15	108	b = LIVE

accuracy = $\frac{True\ Positive + True\ Negative}{True\ Positive + \text{False Negative} + \text{False Positive} + True\ Negative}$

= 81.2903%

So the accuracy of output is decreased when compared to Training set default values. Error percentage also decreased where in default values the mean absolute error is 0.2029 and when the values are changed the error is 0.1965. Size of tree increased when the values are changed, in default values the size of the tree is 21 and number of leaves are 11 after making the confidence factor to 0.5 and min number folds to 2 the size of tree increased to 31 and no of leaves increased to 16.