CS 760

Homework 4

Submitted by: Sneha Rudra

Part 1

- The file 'bayes.py' implements a program to learn naïve bayes and TAN. 'bayes.py' is callable from the command line and accepts the following command line arguments- <train-set-file> <test-set-file> <n|t>

  where l is the learning rate, h is the number of hidden units, e is he number of training epochs, <train-set-file> is training set file name and <test-set-file> is the test set file name.

- Script file 'bayes' can be used to accept the above command line arguments, to invoke bayes.py and the python interpreter. For example: If bayes.py needs to be executed to learn naïve bayes for the lymph dataset using the bayes script, then the following command should be executed from the terminal (in the appropriate path where bayes.py, the training set arff file and test set arff file are located):

  [sneha@royal-18] (25)$ bayes lymph_train.arff lymph_test.arff n

bayes.py reads the training set and test set files in the ARFF format. It can parse ARFF files that are similar in format to the example ARFF files provided (lymph_train.arff, lymph_test.arff, vote_train.arff, vote_test.arff, chess-KingRookVKingPawn.arff).

Part 2

For this part, use stratified 10-fold cross validation on the chess-KingRookVKingPawn.arff data set to compare naïve Bayes and TAN. Be sure to use the same partitioning of the data set for both algorithms. Report the accuracy the models achieve for each fold and then use a paired t-test to determine the statistical significance of the difference in accuracy. Report both the value of the t-statistic and the resulting p value. You can use a t-test calculator for this exercise.

$$\text{Test Accuracy} = \frac{Number\ of\ correctly\ classifed\ test\ instances}{Total\ number\ of\ test\ instances}$$

Fold #	Naïve Bayes Test Accuracy	TAN Test Accuracy
1	0.751572	0.852201
2	0.801887	0.886792
3	0.867925	0.889937

4	0.902516	0.946541
5	0.871069	0.949686
6	0.852201	0.924528
7	0.839623	0.883648
8	0.886792	0.933962
9	0.899371	0.949686
10	0.916168	0.928144

Paired t-test was performed using online t-test calculator: https://www.graphpad.com/quickcalcs/ttest1.cfm to determine the statistical significance of the difference in accuracy. The following result was obtained:

P value and statistical significance:

The two-tailed P value equals 0.0001

By conventional criteria, this difference is considered to be extremely statistically significant.

Confidence interval:

The mean of Group One minus Group Two equals -0.05560012080

95% confidence interval of this difference: From -0.07568411381 to -0.03551612779

Intermediate values used in calculations:

t = 6.2625

df = 9

standard error of difference = 0.009

Review your data:

Group	Group One	Group Two
Mean	0.85891236390	0.91451248470
SD	0.05056741114	0.03394798823
SEM	0.01599081946	0.01073529648
N	10	10

Report both the value of the t-statistic and the resulting p value

t-statistic	6.2625
Two tailed p-value	0.0001