CS 760 HW4: Bayes net

Qihong Lu

December 3, 2016

QUESTION 1

To run the program, type the following command:

```
./bayes <train-set-file> <test-set-file> <n|t>
```

where ’n’ corresponds to the naive bayes algorithm and ’t’ corresponds to the tree augmented bayes net.

Code:

- fitBayesNet.py: fits bayesian network with binary classification
- bayesNetAlg.py: implements bayesian network with binary classification
- util.py: the definitions of some constants and helper functions
- prim.py: obtains max spanning tree with the Prim’s algorithm
- kFoldsCV.py: uses k-Folds cross validation to compare naive bayes and tree augmented bayes

Dependencies:

- pyhton 2.7
- numpy
- scipy
- sys

QUESTION2

For this part, use stratified 10-fold cross validation on the chess-KingRookVKingPawn.arff data set to compare naive 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.

Here's a table summarize the accuracy obtained with 10-Folds cross validation.

	Naive Bayes	TAN
1	0.884375	0.9
2	0.871875	0.921875
3	0.859375	0.934375
4	0.884375	0.925
5	0.915625	0.946875
6	0.875	0.940625
7	0.903125	0.95
8	0.86875	0.91875
9	0.859375	0.90625
10	0.86392405	0.91455696

P value and statistical significance:
- The two-tailed P value is less than 0.0001

Confidence interval:

- The mean of Group One minus Group Two equals -0.0472507910
- 95% confidence interval of this difference: From -0.0639878676 to -0.0305137144

Intermediate values used in calculations:

- t = 5.9312
- df = 18
- standard error of difference = 0.008

Conclusion:

On the chess-KingRookVKingPawn.arff data set, the tree augmented bayes (TAN) is significantly more accurate than the naive bayes algorithm, t(18) = 5.9312, p < 0.0001.