**Homework 5: Decision Trees and Random Forests**

*Antonio Rohit de Lima Fernandes*

**Explanation of Decision Tree Techniques:**

I implemented the decision tree as a class, as recommended by the spec. It was rewarding to build it completely from scratch without any help, but took a little while, and hence is minimalistic as of now. There is no pruning, dealing with attributes, heuristics for faster training, or Adaboost.

Splitting Criteria

* Implemented both the entropy and gini impurity techniques
* Both give similar results
* Used mean of the features as a threshold for splitting
* Using median took less computational intensity but worse results

Stopping Criteria

* Multiple stopping criterion: Purity or maximum depth

Cross Validation

* Ten-fold cross validation performed

**Explanation of the Features Added**

* Added simplistic word count features (Viagra, sex, thanks, please)
* I wrote analyse\_mail.py to view the word counts of both ham and spam mail, and hence get an intuition of what features to use
* Also added a ‘total unique word count’ feature because it seemed like spam had a lot more unique words than ham

**Report of the Results**

* Performance of Decision Tree

|  |
| --- |
| ==================================================  CROSS VALIDATION USING CUSTOM FUNCTION  ==================================================  DEPTH: 1  Depth: 1 Accuracy: 72.88% (+/- 0.11)  --------------------------------------------------  DEPTH: 5  Depth: 5 Accuracy: 76.65% (+/- 0.10)  --------------------------------------------------  DEPTH: 10  Depth: 10 Accuracy: 77.56% (+/- 0.11)  --------------------------------------------------  DEPTH: 25  Depth: 25 Accuracy: 80.43% (+/- 0.12)  --------------------------------------------------  DEPTH: 50  Depth: 50 Accuracy: 85.94% (+/- 0.24)  --------------------------------------------------  Best Depth Value: 50 Accuracy for that Depth: 85.938%  --------------------------------------------------  ==================================================  CROSS VALIDATION USING SCIKIT LEARN  ==================================================  DEPTH: 1  Depth: 1 Accuracy: 77.00% (+/- 0.35)  --------------------------------------------------  DEPTH: 5  Depth: 5 Accuracy: 83.30% (+/- 0.50)  --------------------------------------------------  DEPTH: 10  Depth: 10 Accuracy: 83.77% (+/- 0.58)  --------------------------------------------------  DEPTH: 25  Depth: 25 Accuracy: 82.65% (+/- 0.56)  --------------------------------------------------  DEPTH: 50  Depth: 50 Accuracy: 82.38% (+/- 0.38)  --------------------------------------------------  Best Depth Value: 10 Accuracy for that Depth: 83.772%  -------------------------------------------------- |

* Performance of Random Forest

|  |
| --- |
| ==================================================  CROSS VALIDATION USING RANDOM FORESTS  ==================================================  Num Trees: 25  --------------------------------------------------  DEPTH: 1  Depth: 1 Accuracy: 73.46% (+/- 0.11)  --------------------------------------------------  DEPTH: 5  Depth: 5 Accuracy: 76.83% (+/- 0.11)  --------------------------------------------------  DEPTH: 10  Depth: 10 Accuracy: 79.65% (+/- 0.11)  --------------------------------------------------  DEPTH: 25  Depth: 25 Accuracy: 79.81% (+/- 0.11)  --------------------------------------------------  DEPTH: 50  Depth: 50 Accuracy: 84.93% (+/- 0.20)  --------------------------------------------------  Best Depth Value: 50 Accuracy for that Depth: 84.932%  --------------------------------------------------  ==================================================  CROSS VALIDATION USING SCIKIT-LEARN  ==================================================  DEPTH: 1  Depth: 1 Accuracy: 71.47% (+/- 0.31)  --------------------------------------------------  DEPTH: 5  Depth: 5 Accuracy: 82.60% (+/- 0.70)  --------------------------------------------------  DEPTH: 10  Depth: 10 Accuracy: 85.27% (+/- 0.30)  --------------------------------------------------  DEPTH: 25  Depth: 25 Accuracy: 85.52% (+/- 0.49)  --------------------------------------------------  DEPTH: 50  Depth: 50 Accuracy: 85.53% (+/- 0.52)  --------------------------------------------------  Best Depth Value: 50 Accuracy for that Depth: 85.528%  -------------------------------------------------- |

* Best Kaggle Score

78.2%

**Show Splits (Decision Tree)**

* Feature 38 (“sex”): 0.044
* Feature 5 (“prescription”): 0.028
* Feature 8 (“featured”): 0.0224
* Feature 6 (“creative”): 0.0124
* Feature 39 (“viagra”): 0.0042

**Most Common Splits at the Root (Random Forest)**

* All my trees (40) split on feature 35 (“at”): 0.044

**References**

- Reading material provided in HW