**BUSINESS DATA MINING**

**(IDS 572)**

**Solutions to Homework 3**

**Group Members**

* Amey Pophali (apopha2@uic.edu)
* Karthik Varanasi (vvaran3@uic.edu)
* Mrinal Dhawan (mdhawa3@uic.edu)

**Problem 1 –**

**To see whether a classifier is actually working, we should compare it to a constant classifier**

**which always predicts the same class, no matter what the input features actually are.**

* **What fraction of the e-mails are actually spam?**

0.3940448

vari = table(spam$spam)

varip <- ((vari)/nrow(spam))\*100

varip

email spam

60.59552 39.40448

* **What should the constant classifier predict?**A constant classifier predicts the same target value irrespective of the input. Here The constant classifier should predict that all the values fall into a single category “E-mail”.
* **What is the error rate of the constant classifier?**

The error rate of the constant classifier can be found by comparing the predicted value to the actual value of the variable. The error is 39.40448%.

**Problem 2 –**

**Divide the data set at random into a training set of 2301 rows and a testing set of 2300**

**rows. Check that the two halves do not overlap (use intersect () function), and that they have the right number of rows. What fraction of each half is spam? (Do not hand in a list of 2301 row numbers.)**

**Problem 3 –**

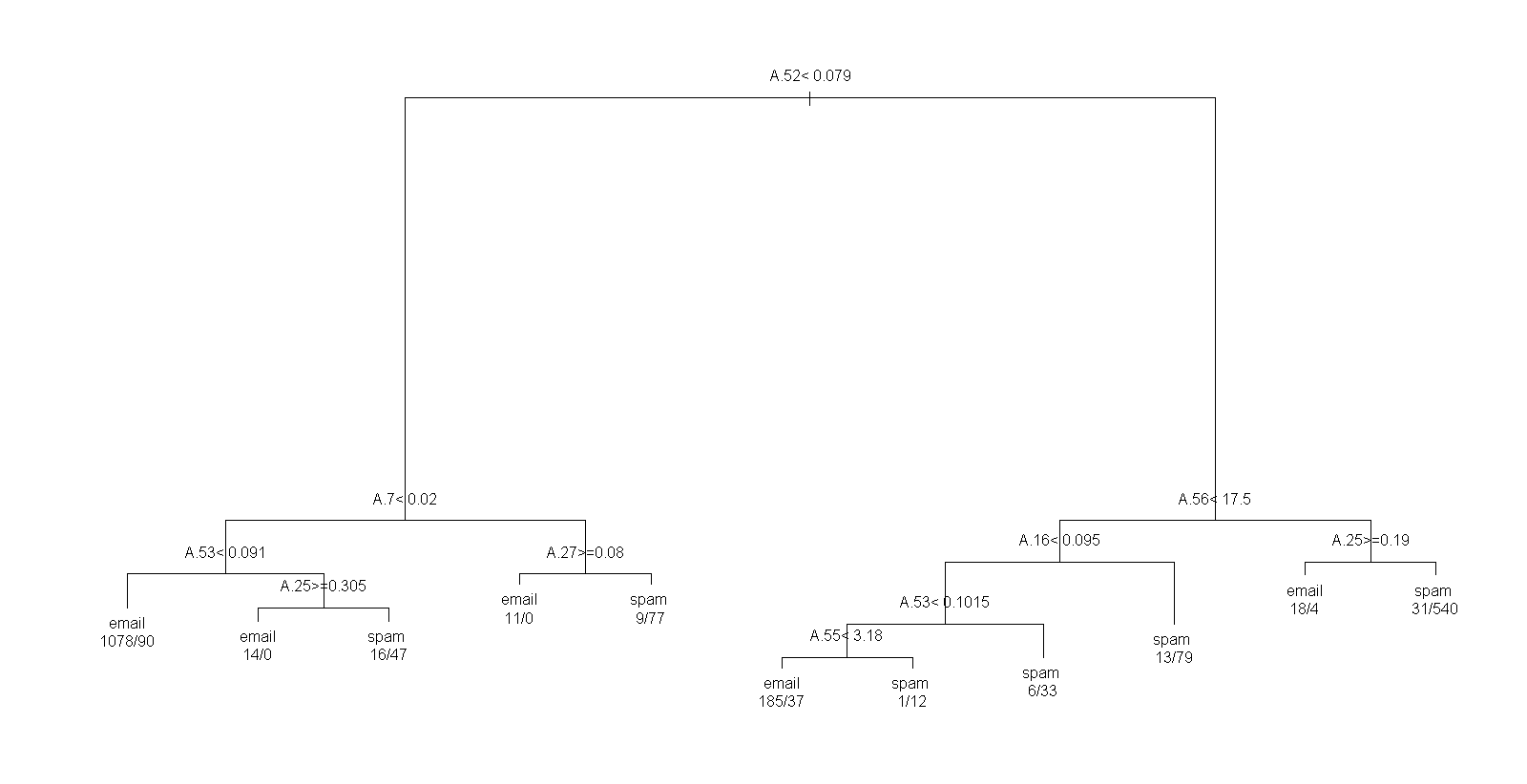
**Remember to show your work by including your code.**

* **Fit a classification tree to the training data. Prune the tree by cross-validation (see below).**

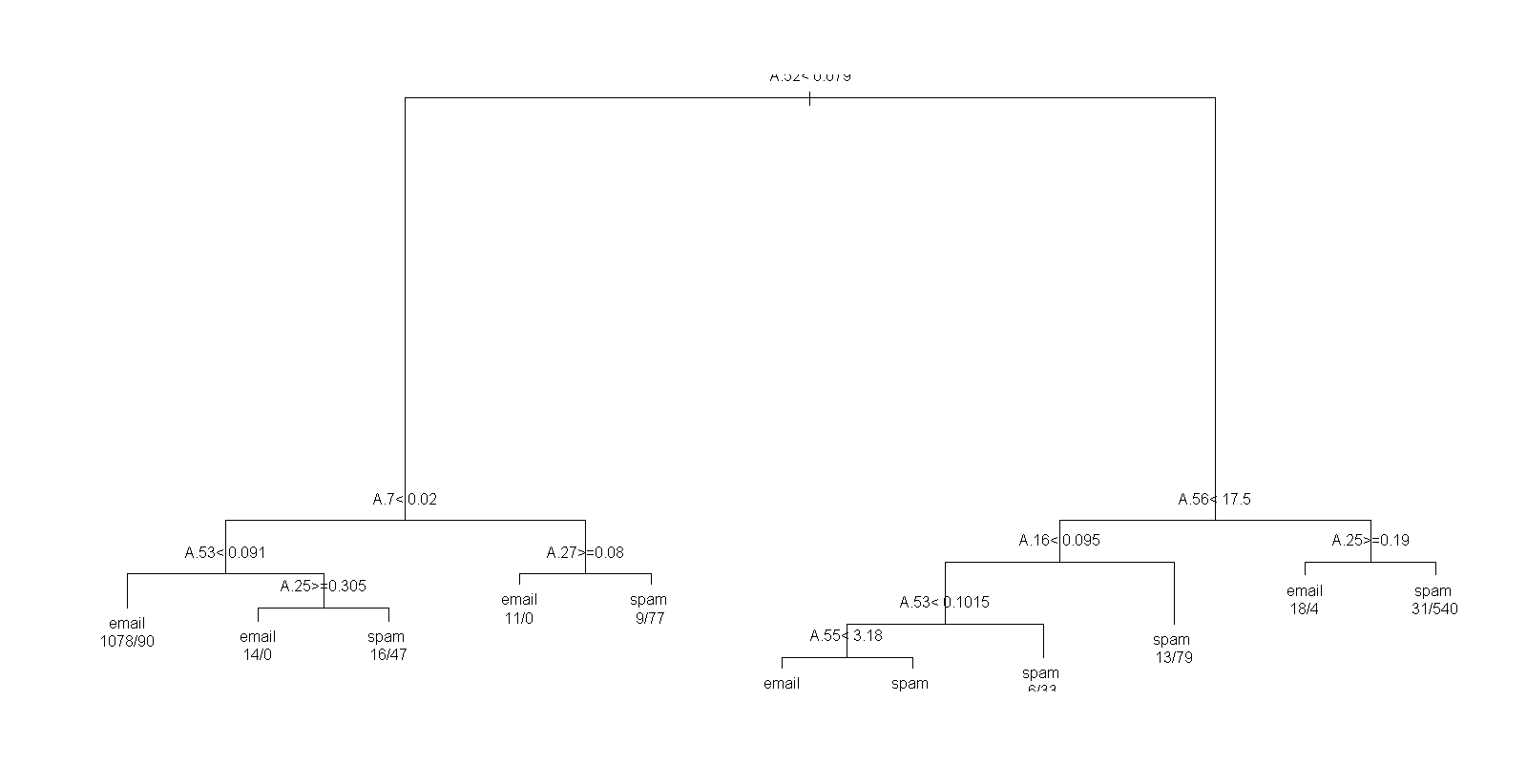
**Include a plot of the CV error versus tree size, a plot of the best tree, and its error rate on the**

**testing data. Which variables appear in the tree?**

**Original Tree:**



**Pruned Tree:**



* **Use bagging to fit an ensemble of 100 trees to the training data. Report the error rate of the**

**ensemble on the testing data. Include a plot of the importance of the variables, according to**

**the ensemble.**

* **Which (if any) of these methods out-performs the constant classifier?**

**Problem 4 –**

**Pick the prediction method from the previous problem with the lowest error rate.**

* **What fraction of the spam e-mails in the training set did it not classify as spam?**
* **What fraction of the genuine e-mails in the testing set did it classify as spam?**
* **What fraction of e-mails it classified as spam were actually spam?**