Q1

1. 27 + 37 + 1 + 24 + 2 + 1 + 29 + 3 + 1 + 1 + 1 + 24 + 32 + 2 + 1 + 24 + 1 + 3 + 31 + 1 + 1 + 27 + 2 + 1 + 2 + 21 = 300

The size of the training set used is 300.

1. [[27 0 0 0 0 0 0 0 0 0] row 1

[ 0 37 0 0 0 0 0 0 0 0] row 2

[ 0 1 24 0 0 0 2 1 0 0] row 3

[ 0 0 0 29 0 3 0 1 1 1] row 4 The class in the 4th row had the highest number of errors

[ 0 0 0 0 24 0 0 0 0 0] row 5

[ 0 0 0 0 0 32 0 0 0 2] row 6

[ 0 0 0 0 0 1 24 0 0 0] row 7

[ 0 0 0 0 1 3 0 31 0 0] row 8

[ 0 1 0 0 0 0 1 0 27 2] row 9

[ 0 0 0 0 1 2 0 0 0 21]] row 10

1. the overall percentage correct = (27 + 37 + 24 + 29 + 24 + 32 + 24 + 31 + 27 + 21) / 300 x 100% = 92%

Q2

8 hyper-rectangles

The number of hyper-rectangles is equal to the number of leaf nodes that indicate the spam class.

Q3

This model has 16 parameters.

Extension Question

This article briefly described three methods of estimating error rates in classification, Train-and-test (percentage split), Cross-validation and Bootstrap, these three methods are still correct and still have a wide range of applications in the field of machine learning.

However, one thing that is not mentioned in the article is that Train-and-test and Cross-validation are more suitable for large data sets, and Bootstrap works well when used with small samples. Bootstrap method can avoid the sample reduction problem caused by Cross-Validation by resampling, it also can be used to create randomness of data.

The article also mentioned that Breiman et al. (1984) note that there are practical difficulties in applying the bootstrap to decision trees, however, with the development of machine learning technology, this information has become obsolete. For instance, the training algorithm for Random Forests applies the general technique of bootstrap aggregating, it’s first step is to randomly select k new sample sets from the training data set using the Bootstrap method, and construct k decision trees from this.