Project: Spam Filtering

S. Hoa Nguyen

1 Data source

http://archive.ics.uci.edu/ml/datasets/Spambase

2 Dataset descriptions

The data set *spambase.data* contains descriptions of over 4,000 emails, each email has 56 attributes (including the frequency of individual words, the average length of the longest length of uninterrupted sequences of characters printed, ...). Emails are classified into one of two classes: it is spam or not. Exact descriptions of this collection of data are in the *spambase_document.doc*

Task:

Induce the prediction model for e-mail classification problem.

3 Methodology

Feature selection:

a) Apply one of feature selection algorithms available, eg. in WEKA system to filtr redundat attributes.

b) Prepare the data set with relevant attributes.

Classification algorithms

Instruction:

- **Normalize the values** of features.
- Create a neural network and use **a back propagation algorithm** to train the network.
- **Test the accuracy** of the calculated network. Describe your experimental results and write the conclusion.

4 Testing model

The data set is available at *Data Source*. Randomly split it in two subsets:

- Training data: 80% of whole data set.
- Testing data: the remaining part.

5 Final result: Confusion matrix

Test the calculated neural network. As the final result create the *confusion matrix* and

a) calculate the **classification accuracy** of the network.

b) find the class with **highest** and the class with **lowest accuracy**.