THE UNIVERSITY OF HONG KONG

MSc in E-Commerce and Internet Computing

ECOM7126 Machine Learning for Business and E-Commerce
(2023-24)

Assignment 4 – Spam Email Prediction using Neural Network

A cyber security company wishes to develop a SPAM filter based Artificial Neural Network. A sample of 4,600 emails are collected within a certain period randomly from the company's email server, and 48 of words, 6 characters (or symbols) and other features are randomly picked from these email samples as input to train the Neural Network. The emails are then carefully labeled as SPAM or HAM (not spam). Your job is to build a Neural Network based spam filter and demonstrate the quality of your model.

The dataset consists of 4,600 of labelled instances, in .csv format. The dataset is already randomized and the first 3,600 samples are for training, and the remaining 1,000 samples are for testing. The input features of each email instance consist of the following categories:

word_freq_WORD	is the frequency count (in percentage) of a particular "WORD" = # of occurrence of "WORD"/ total # of words in the email (in %)
char_freq_CH	is the frequency count (in percentage) of a particular character "CH" = # of occurrence of character "CH"/ total # of characters in the email (in %)
capital_run_length_average	average length of uninterrupted sequences of capital letters in the email
capital_run_length_longest	length of longest uninterrupted sequences of capital letters in the email
capital_run_length_total	total number of capital letters in the email
spam	0 = not spam; 1 = spam

You should include the following in your report (plus your Colab notebook with you Python code to get the results):

1. Examine and understanding the feature set provided and handle any issues, if any, in the dataset.
2. Design a multi-layer neural network, train the network with the appropriate validation, then test your final model. You may wish to tune your model if time permits.
3. As a stretched goal (optional), try an ML model which is not a neural network and compare.
4. Discuss what you have learned from this assignment.

Dataset provided: SpamDataset.csv

Deadline: 5 May 2024