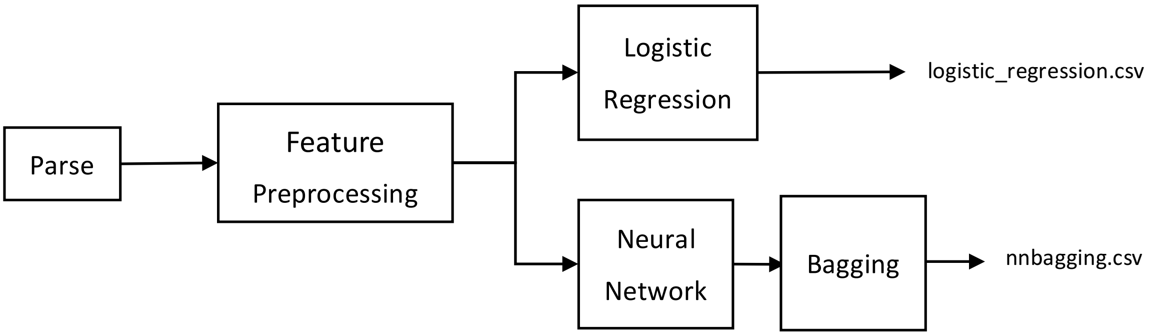
# Machine Learning Homework 2 Report

Student ID: r05922063  Name: 陳啟中

## Overview



## Parse

In parsing phase, the purpose is to read spam_train.csv and spam_test.csv and interpret them into three matrices: training feature matrix, training label matrix and testing feature matrix.

Feature matrices, including training and testing, consist of rows of feature sets. Each row represents an instance. In a row, there are totally 57 columns. Training label matrix consists of rows corresponding to rows in training feature matrix. Each row records whether the mail is a spam or not.

## Feature Preprocessing

By observing the range of features, we can find the last 3 features are relatively more huge than the first 54 features. To reduce the side effect of scale problem, the program standardize these 3 features to make them smaller. Besides, the program adds additional 3 features by dividing one of capital_run_length features by another.

## Logistic Regression

My logistic regression model takes 60 features, 60 weights and a bias term to perform linear combination and output the prediction result. The following are some details:

## Training Data Random Shuffling

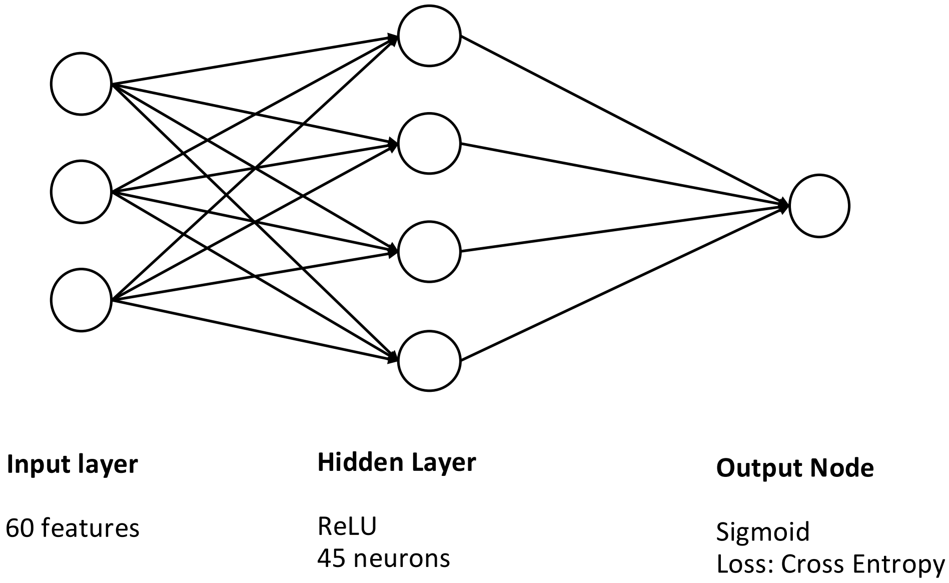
Upon several iterations of training, the program shuffles the order of training data randomly. This prevents the training model from learning the bias obtained by the certain order of data.

## Early stopping

To prevent overfitting, early stopping is an easy and efficient way. The key point is to find the ending point just before overfitting occurs. How? I firstly split the training data into 80% subset training data and 20% validation data. Upon each epoch (1000000 iterations), the program calculates the testing MSE on validation data. Once the testing MSE grows, we then find the stopping point.

# Neural Network

## Topology

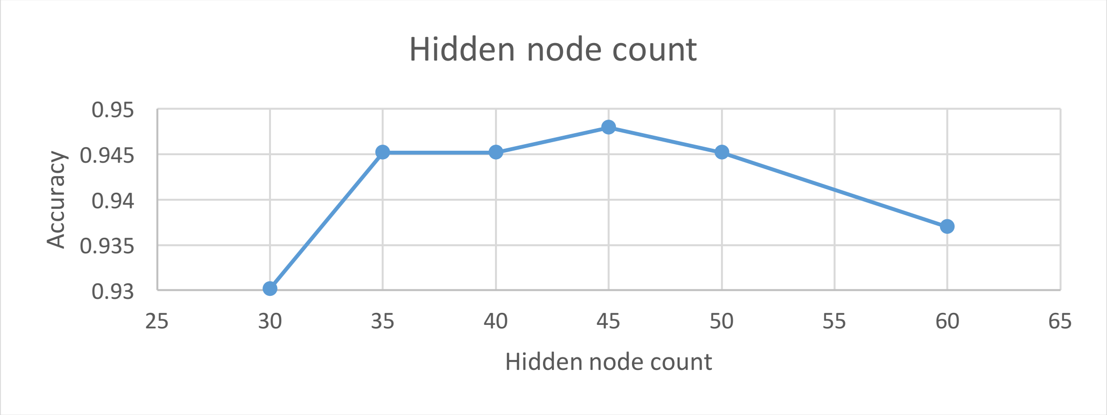


## Description

Between input layer and hidden layer, there are configurable synapses (weight and bias). Each hidden node perform linear combination and then rescale the result by ReLU function, which is faster to converge. Then, the output node performs sigmoid function on the linear combination of hidden node outputs. In backpropagation phase, the program uses cross entropy as loss function.

## Hidden node count discussion

Hidden node count is a tuning parameter. If the count is too low, the neural network cannot extract hidden features from data. If the count is too high, it makes the model overfitting to training data. I make some trials on different hidden node counts to find the optimal one. After the experiment, I find 45 hidden neurons is the best parameter. The following shows the relationship between hidden node count and validation accuracy:



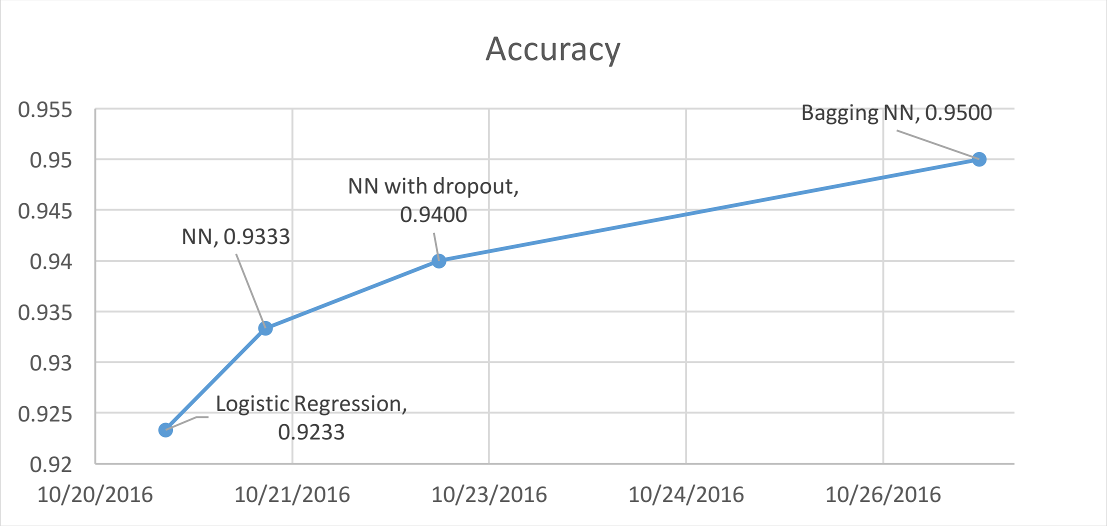
## Dropout

Dropout is an efficient way on neural network to prevent overfitting by randomly dropping out some nodes in training phase. After some tuning, I set 0.2 dropout probability on hidden layer for better result.

## Bagging

Bagging is also a method to prevent overfitting. The first step is sampling the training data with or without replacement. The second step is to train the subset of training data by neural network. Repeat the two steps generate several neural networks. Finally, evaluate on testing data by averaging the results of the neural network set. The sampling method I use is n' = n with replacement. The bag size is set to 8.

# Milestone


Accuracy
10/20/2016
10/21/2016
10/23/2016
10/24/2016
10/26/2016
Logistic Regression, 0.9233
NN, 0.9333
NN with dropout, 0.9400
Bagging NN, 0.9500