CS5691: Pattern Recognition and Machine learning
Assignment 3

1 Choice of dataset

The dataset chosen for this assignment was, the **Spam Mails Dataset** from Kaggle. This dataset contains 5171 unique emails out of which 3672 emails are labelled "ham" and 1499 emails are labelled "spam". 80 % of the dataset is used for training and the remaining 20 % is used for testing.

2 Preprocessing the data

The raw email text was split into a list of words and preprocessed using the following steps

- Punctuation removal
- Digits removal
- Stopword removal

2.1 Punctation removal

Punctations such as "," , ".", "!", "?" are removed from the list of words (i.e tokens)

2.2 Digits removal

Digits are removed from the dataset. This prevents the occurence of different digits in the dataset from influencing our prediction.

2.3 Stopword removal

Stopwords such as "a", "an", "the", "he", "him", etc. are removed from the dataset since the don't contribute much to the prediction.

3 Feature extraction

To extract the binary features from the obtained tokens, a dictionary is created. A dictionary is a list of unique words or tokens in the train dataset. Now the feature vector of each email is given by:

$$f[i] = \left\{ \begin{array}{ll} 1, & \text{if } dictionary[i] \text{ is present in the email} \\ 0, & \text{otherwise} \end{array} \right\} \tag{1}$$

Where **f** is the d dimensional feature vector. 39133 unique words were extracted from the test dataset. Thus d = 39133.

4 Algorithms used

The following algorithms were trained on the train dataset and the corresponding accuracy scores were measured.

4.1 Logistic regression

The logisitc regression algorithm was performed on the train dataset. The weights were initialised as 0, and gradient ascent was performed

$$w_{t+1} = w_t + \alpha \times \sum_{i=1}^{n} x_i(y_i - s(w_t^T x_i)) \tag{2}$$

$$s(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

where s(x) is the sigmoidal function.
Here the learning rate α was chosen to be 0.01 and the algorithm was run for 200 iterations.
This was tested int test dataset and an accuracy of 96.5 % was obtained.