Assignment 3 (Due Date is 23:59pm 21st Oct 2022)

Data Introduction

4601 email messages were sent to "George" at HP-Labs. He labeled 1813 of these as spam, with the remainder being good email (ham). The goal is to build a customized spam filter for George. The feature set tracks 57 of the most commonly used, non-trivial words in the corpus, using a bag-of-words model. Recorded for each email message is the relative frequency of each of these words and tokens. Included as well are three different recordings of capitalized letters.

The data has 59 columns:

1. The first column "spam" is a logical variable, for which "TRUE" is spam, "FALSE" is ham (good email).

2. The second colum "testid" is also a logical variable. An optional split into training data (FALSE) and test data (TRUE) data.

3. The remainder of the columns are features used to build a predition model.

As guided by the following questions, please use various methods to help George to build a spam filter, which can predict whether future emails are spam or ham using these 57 predictors.

Questions

Question 1 [20 marks]: Fit a neural network model to this set of spam data. Present the values for weights parameters; and calculate the test mse on the test data set and training mse on the training data set, respectively.

Question 2 [20 marks]: Apply support vector machine (SVM) to classify this set of data. Calculate the training mse and test mse, respectively.

Question 3 [20 marks]: Apply flexible discriminant analysis (FDA) on this set of data. Calculate the training mse and test mse, respectively.

Question 4 [20 marks]: Apply penalized discriminant analysis (PDA) on this set of data. Calculate the training and test mse, respectively.

Question 5 [20 marks]: In terms of trainig mse and test mse, please compare the performances of these four methods (neural network, SVM, FDA and PDA), and provide reasons for your analysis, i.e. for each method, why the performances are good or bad.

Note: for Questions 1 - 4, please provide R codes.