Lab 7

Spam classification using logistic regression

Consider the email spam data set. This consists of 4601 email messages, from which 57 features have been extracted. These are as follows:

- 48 features, giving the percentage of words in a given message which match a given word on the list. The list contains words such as “business”, “free”, “george”, etc. (The data was collected by George Forman, so his name occurs quite a lot.)
- 6 features, giving the percentage of characters in the email that match a given character on the list. The characters are ; ([! $ #
- Feature 55: The average length of an uninterrupted sequence of capital letters
- Feature 56: The length of the longest uninterrupted sequence of capital
- Feature 57: The sum of the lengths of uninterrupted sequence of capital

1. Download the data at http://www.cse.scu.edu/~yfang/coen140/spambase.zip. The data is split into a training set (of size 3065) and a test set (of size 1536).

2. Please normalize the features by standardizing the columns so they all have mean 0 and unit variance.

3. Build and fit a logistic regression model using gradient descent. Report the error rate on the training and test sets.