About the Spam dataset.

Link to data description: <http://archive.ics.uci.edu/ml/datasets/Spambase>

Link to actual dataset: <http://archive.ics.uci.edu/ml/machine-learning-databases/spambase/spambase.data>

This is a old dataset (1999) created by HP. The aim is to separate out spam emails from good (non-spam) emails. Thus, each record (or row in the data file) corresponds to one email.

A labeled training dataset was created by the HP people by taking a bunch of emails and having some human editors label each one of them as spam and non-spam so as to form the training set.

After the model is developed, it will be applied on each new email in order to classify it as spam or non-spam.

The last column of the dataset (variable is called "spam") is a binary variable - it is the target. 1 means spam and 0 means non-spam.

HP came up with some features (variables used to do the prediction of the target) that will be useful for forming patterns.

Most of the features are word-occurrence features.

For example, does the email contain a specific word, e.g., "internet"? This feature (column) is called "word.freq.internet" in the data file. The actual feature value is computed by counting the number of occurrences of the word "internet" in an email and normalizing that value to make a number such as "0.88" or "1.23" etc.

This "turning of each email" into a set of (feature, value) pairs is a pre-processing step that is usually executed using some text processing script.

The last two features, 'capital.run.length.longest', 'capital.run.length.total' indicate the occurrences of capitals in an email. This is useful because, in those days (pre year 2000), most advertisement emails tended to have lot of capitalized letters while non-spam emails usually lack many continuously occurring capital letters.