Udacity Machine Learning Nanodegree

Capstone Proposal

Spam Email Detection

Gaurav Jain
February 2020

- ### Domain Background

  Nowadays, along with desired emails we get a lot of unsolicited emails like promotional, social, and other secondary types. Hence it becomes hard to distinguish between spam and genuine emails and create a lot of noise in our inbox. There are cases where an attacker can send you spam mails and can trick you to expose your private data using phishing or malicious attachments. Most of the time spam mails do not greet you by name, they simply start with Hello, Hi, Dear, etc. Sometimes the subject of spam emails also doesn't make any sense and usually contains words such as free, money, discount, help etc. In some cases we can even see that email address is really weird e.g webscrapper57@hackerdev.com. My curiosity behind this problem is that, a mail can be spam for one person and for another It can be a legitimate email so it should learn from the user's emails and behave as per that user.

- ### Problem Statement

  The problem is to create or use the existing machine learning model to predict whether an email is spam or not.

- ### Datasets and Inputs

  We'll take 30228_easy_ham.tar.bz2 and 30228_spam.tar.bz2 from Apache Spam Assassin's public datasets https://spamassassin.apache.org/old/publiccorpus/.
  There are approximately 2500 ham and 500 spam emails in the dataset. Compressed dataset's approx size is 3 MB so I'll attach with the final source code as well. We'll create a final dataset for training and testing using the above dataset, and that final set will have two columns that will be fed to the model.

- ## Solution Statement

  We'll parse the data, sanitize/tokenize it and then we'll apply LogisticRegression on it. We'll also make use of hyperparameters to tune the model and get the best result.

- ## Benchmark Model

  For benchmarking we'll use the simple Gaussian Naive Bayesian model and will try to get the improved result against it. Reasoning behind using Naive Bayesian classifier is that it is fast for both training and prediction.

- ## Evaluation Metrics

  Evaluation metrics would be an accuracy score along with Precision & Recall calculated using True/False Positive/Negative.

- ## Project Design

  We will simplify the data set by removing headers/metadata, urls and other unnecessary information. Once we have the sanitized data, we'll have two columns in the final dataset. Column1 would be the actual email message and column2 would be a label that represents the email as spam or not.
  We'll split the data in training and testing set 70%, 30% respectively.
  After doing this we'll have a clear data structure defined for data and label for both training and testing set. We'll use the LogisticRegression algorithm which seems to be best for this problem. We'll play with different hyperparameters to tune the model and submit the one that performs best.

- ## References

  https://en.wikipedia.org/wiki/Linear_regression

  https://en.wikipedia.org/wiki/Binary_classification

  https://en.wikipedia.org/wiki/Supervised_learning