INTRODUCTION TO DATA SCIENCE PROJECT

THE ULTIMATE SPAM DETECTOR

BUILDING TEAM SPIRIT TOGETHER

SOCIETE GENERALE
Corporate & Investment Banking

1. SPAM

From Wikipedia, the free encyclopedia

Spam or **SPAM** may refer to:

- Spamming, unsolicited or undesired electronic messages
  - Email spam, unsolicited, undesired, or illegal email messages
  - Messaging spam, spam targeting users of instant messaging (IM) services, sms or private messages within websites
- Spam (food), a canned pork meat product

As you probably have guessed, in this project we will not be talking about canned pork but rather about spams in emails.

You recently have been hired by a big multinational holding, Awesome corp, and his president has a huge problem, his mail-box is continually targeted by spamming process and is close to be unusable.

Of course, he is your boss and you can’t really tell him that spam usually result from a very bad usage of his mail address, and as a data scientist it is your job to create a spam filter so good that it will predict whether a received email is a spam or not.

1.1. PREREQUISITES

In order to test your algorithms you will use a data-set given to you by the university of Irvine, California : UCI

http://mlr.cs.umass.edu/ml/datasets/Spambase

this data-set will allow you to test your algorithms, implement them, train your programs, compare their accuracy.

To share your work with your manger (me) and the rest of your team, you will set a github project.

The language in which you will implement your work is for you to choose, but just know that I really encourage you to use python.

At the end of your project, you will present me all of your work in a 30 minutes report

1.2. UNDERSTANDING THE DATA

The first part of this work is to understand the data you will be given, pay attention to the description of the data on the UCI website, your goal is to clean the data of every features that could lead to misfiting your algorithm, for example the non-spam data comes from work related emails and often contains the word george and the area code 650 but for a stronger and wider spam filter you will want to omit those cases.

It is also your job to understand every columns of the data (all the 57) omit the one that will not matter, because those are real data they will contain gaps and missing values, so it’s also your job to clean those and present them in the best shape for your algorithm.

1.3. ALGORITHM CHOICE

For each of the algorithms you will implement you will compute metrics and statistics about their accuracy and present them in your reports (I like nice graphs). You will implement several algorithms with sklearn and play with the tunning (depending of the algorithms you chose)

Keep in mind that we search an accurate algorithm but also a fast one, so the final decision will have to be supported by evidences and do not hesitate to search the internet for more metrics to use.

1.4. TO GO FURTHER

As a proof that your algorithm works with data unrelated with the one from your tests you will apply it to, whether data that you will have generated yourself or real data that you can find in the internet.