SpamDetector

Project for Data Mining and Machine Learning

Francesco Campilongo

Contents

1 Introduction **2**

2 Datasets **3**
2.1 Data Preprocessing 3

3 Design **4**
3.1 Implementation 4
3.1.1 Libraries 4
3.1.2 Functions 4
3.1.3 Main code 4

4 Tests and results **5**

Chapter 1

Introduction

The purpose of this documentation is to prove the machine learning algorithms utility for the detection of spam, useless messages, into dataset of emails.
This is a basic text classification problem, and so the algorithm used are very known in the lecture.
The project is carried out on python.

Chapter 2

Datasets

The datasets used for this project are two:

- E-mail dataset
- SMS dataset

Those two datasets where found on Kaggle.com both of them in a .csv format.

2.1 Data Preprocessing

The datasets found were already pretty good for the type of result this project wants to achieve. The E-mail dataset is characterize from three column a "label", "text" and "label_num", the second column contains the actual message, the first and third column are the same, but in the first one there are the actual words "ham", for the messages which are not classified as spam, and "spam", instead in the third column there is there are "0" for the ham messages and "1" for the spam messages;
The first column has been discharged since not useful for the execution of the classification algorithm.
The SMS file had need a little more work to obtain a dataset usable, because the it has five columns named "v1", "v2", "", "" and "", the first contains the words "ham", for the messages which are not classified as spam and "spam"; the second column contains the actual messages and the last three columns are basically blank, so in order to obtain a good dataset to work with the last three columns has been discharged the first two are taken in consideration, but in the first one the "ham" word is changed with a "0" and the "spam" word with a "1", in order to have the same kind of datasets between the E-mail and the SMSs.

Chapter 3

Design

The application design is pretty basic, it is written in python (version 3) does not involve any classes, just the creation of some functions in order to keep the code clean and more understandable.

3.1 Implementation

3.1.1 Libraries

The libraries involved into this application are two:

- Pandas
- Sklearn

Pandas gives the possibility to use a dataset in .csv (and many others) from the disk, implementing it on a dataframe into the main memory to use all the machine learning algorithm that are imported from Sklearn.

3.1.2 Functions

Support Vector Machine

3.1.3 Main code

Chapter 4

Tests and results