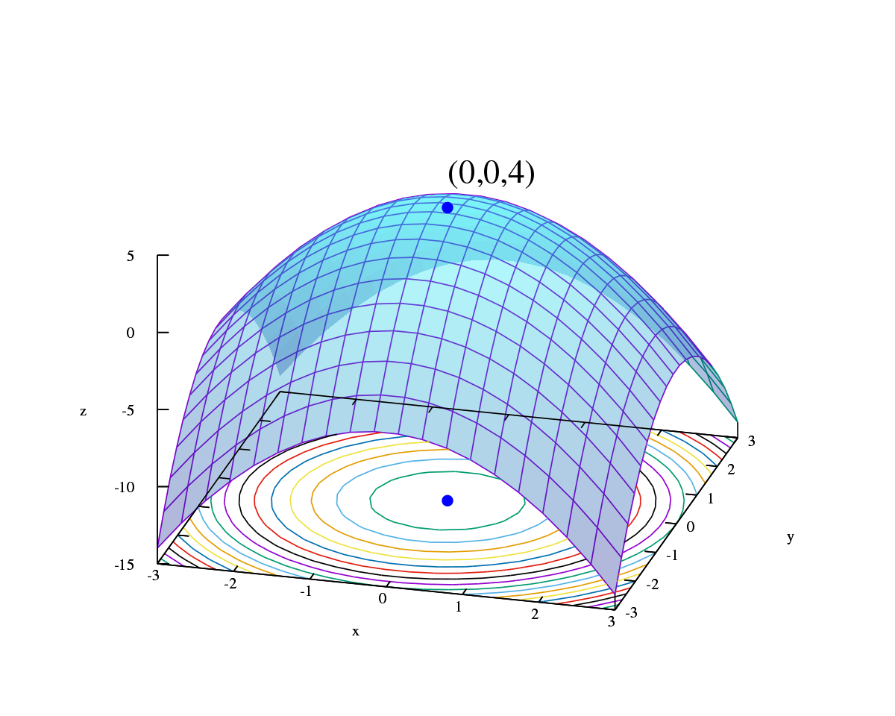
**DAT405: Introduction to data science and AI**

Module 3



BROU BONI Joël

SANOGO Ibrahim Bechir

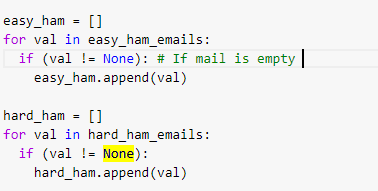
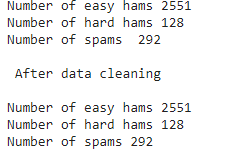
# Preprocessing

This report presents our work on the assignment 4. In this assignment, we worked with Naïve Bayes algorithms using the scikit-learn package. We implemented Naïve Bayes classifiers in Python that classify emails into spam and non-spam (“ham”) classes. To do so we had in our disposition 3 dataset composed with:

- easy-ham: non-spam messages typically quite easy to differentiate from spam messages.

- hard-ham: non-spam messages more difficult to differentiate

-spam: spam messages

 After getting the datasets, we decided to make some data cleaning. Indeed, there was empty emails inside of the files that we got. We decided to take away because it could false the accuracy of our model since they are e-mails of this type in the spams and hard hams files. And since the mail is empty it is impossible to predict with accuracy if it is a ham or a spam. For this reason, we decided to not keep them.

After checking at the cleaned dataset, we see that the email files contain a lot of extra information, besides the actual message, such as the e-mail of the sender the footer of the headers. Regarding the task asked we split the datasets with 70% training set and 30% dataset.

# Part 2:

The goal here was to train our model with the dataset that we made in question 1.

We used two Bayes classifiers, the multinomial and the Bernoulli naïve Bayes. Multinomial Naïve Bayes classifier works on the concept of term frequency which means that how many times does the word occur in a document. This model tells two facts that whether the word occurs in a document or not as well as its frequency in that document. On the other hand, Bernoulli Naïve Bayes classifier works on the binary concept that whether the term occurs in a document or not but unlike Multinomial Naïve Bayes, it does not tell about the term frequency.

So, after training the model and testing it has asked in the report. We get the following result:



As we assumed before, the multinomial model has better accuracy for this application. Indeed, this model seems less biased towards one class and can achieve higher accuracy than the multivariate Bernoulli model, in particular when frequency information must be taken into account in the detection which is the case here.

# Part 3:

The last step of the basic tasks of this assignment was to concatenate the spam dataset and the easy ham dataset and after a spam dataset and the hard ham dataset to compare the accuracy of the two naives bayes algorithms. We got the following results:

|  |  |  |
| --- | --- | --- |
|  | **Bernoulli** | **Multinomial** |
| **Spam versus easy-ham** | 0.89449 | 0.98475 |
| **Spam versus hard-ham** | 0.6904761 | 0.873015 |

We can see that the accuracy of the spam vs easy ham are pretty close for both processes but drops to 69% for the Bernoulli algorithm when we are comparing the hard hams vs the spams. This result seems coherent regarding what we explained before. Indeed, since the Bernoulli Naïve Bayes Classifier is working on the binary concept that whether the term occurs in a document or not but is not telling about the frequency, it is normal that the result are not as good as before because we find the same types of words in these mail categories