ArXiv Article Classification

Fundamentals of Machine Learning

Matteo Esposito (20173298)

University of Montreal

0 Introduction

This report will cover the efforts related to the classification of articles on the popular scientific research paper distribution website, `arxiv.org`.

More specifically, given the abstract of several thousand papers, our goal was to classify them into one of the following categories: astro-ph.GA, math.AP, astro-ph.CO, math.CO, stat.ML, cs.LG, gr-qc, astro-ph, astro-ph.SR, hep-th, physics.optics, hep-ph, cond-mat.mtrl-sci, cond-mat.mes-hall, quant-ph

1 Feature Design

Since at the lowest level we were interested in the count of unique words in our training set abstracts, our feature design could be seen as a dataset condensing/cleaning exercise. All efforts to tidy the dataset were done to reduce the number of unique words/strings (grouping of string characters between spaces) down from over 65,000 to 21,253 which speeds up computation times significantly.

Note: All cleaning functions were implemented using the following syntax (adapted from the TA's notebook):

```
df[t] = df[t].apply(lambda x : re.sub("<regex pattern>", " ", x))
```

More specifically, the following changes were done:

1. Remove newline characters
2. Remove punctuation
3. Remove dollar signs and any mathematical symbols (very important given that many abstracts had mathematical terms and functions)
4. Remove standalone integers
5. Make all strings lowercase
6. Strip all strings of spaces (using x.strip())
7. Remove all common stopwords

2 Algorithms

The algorithms considered and implemented in this kaggle were the following:

- Random Classifier
- Bernoulli Naive Bayes Classifier using base python + numpy
- Multinomial Naive Bayes Classifier from the sklearn library including the use of CountVectorizer, TfidfTransformer and Pipeline.
- Support Vector Machine Classifier from the sklearn library including the use of CountVectorizer, TfidfTransformer and Pipeline.

3 Methodology

4 Results

Figure 1: Random Classifier Accuracy Comparison

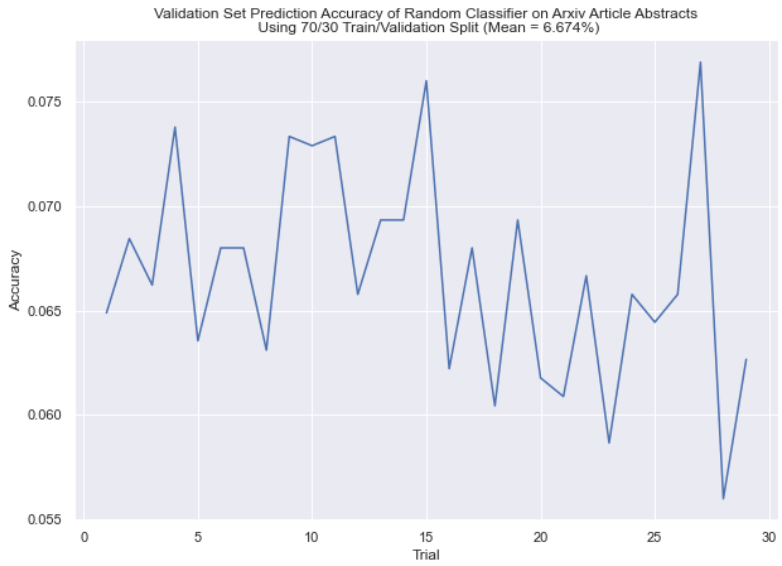


Figure 2: Bernoulli NB and Multinomial NB Accuracy Comparison

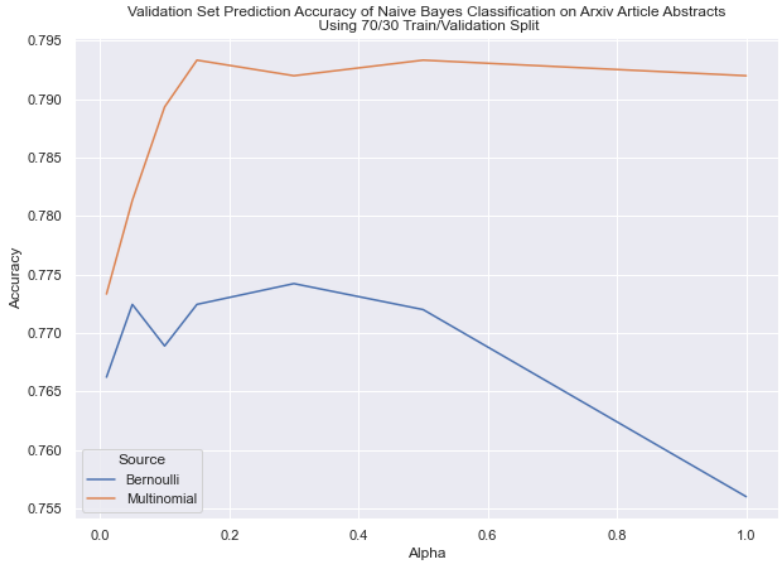
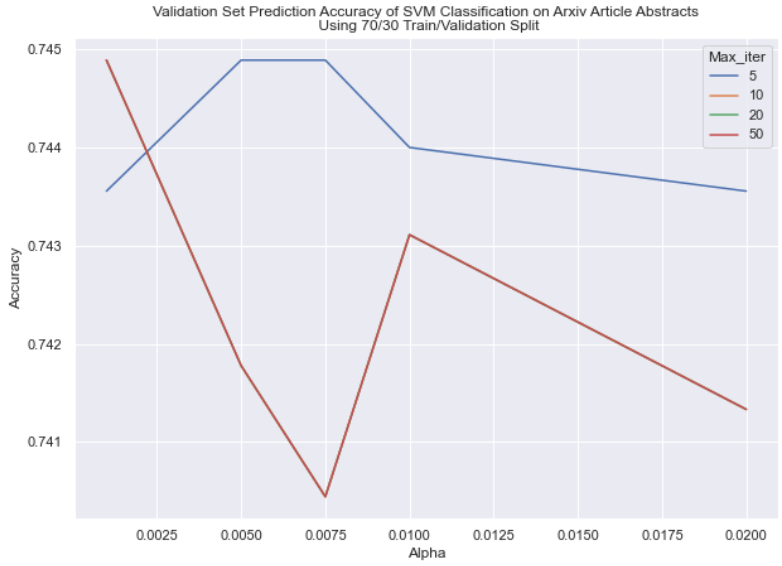


Figure 3: SVM Grid Search Results



5 Discussion

6 References