CSCU9YE Assignment Report

Student no: 2519302

November 17th 2018

Contents

1 Introduction

BIG ASS TODO

2 Specification

Talk through each section of the flowchart

3 Pre-Processing

This section will aim to cover the different aspects of pre-processing employed in this project. Within the datasets used in this project, we can see that the general steps one would need to take are as follows:

- Lemmatization
- Removal of stop words
- Removal punctuation

Along with the aforementioned steps, the text will be processed to handle case. This is done by converting the text to lower case before any text normalization is done.

3.1 Lemmatization

Lemmatization is a form of text normalization where the aim is to remove inflection and/or derive the base word from a family of words in the dataset. The base of a word, is also known as the **lemma** or the **dictionary form**. An example of which is shown in *Table 1* below.

Such mapping then allows us to find all relevant documents using a specific word.

Word		**Lemma**
Cleaning	$\rightarrow$	Clean
Cleaner	$\rightarrow$	Clean
Cleanliness	$\rightarrow$	Clean

Table 1: Showing the mapping of words to its lemma

3.2 Removal of Stop Words

Stop words are words such as "the", "he" or "in". These words do not contribute to the overall meaning of the text and are thus removed during the pre-processing stage. This also tends to be done in order to improve indexing times for larger datasets.

3.3 Removal of Punctuation

Punctuation, like stop words are removed due to their lack of contribution to the overall understanding of the text.

4 Machine Learning Methods

This section will aim to provide some knowledge on the machine learning models employed in this project.

4.1 Support Vector Machine

A Support Vector Machine (SVM) is a discriminative classification method. This method produces a classification by finding an optimal hyperplane that segregates the data points and thus returns classifications [1]. This can be visualized easiest on a 2-Dimensional plot as can be seen below in *Figure 1*:

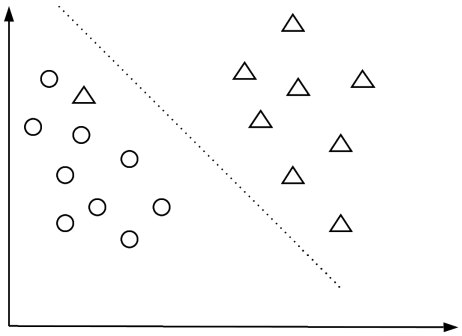

Figure 1: SVM on 2-D plot area

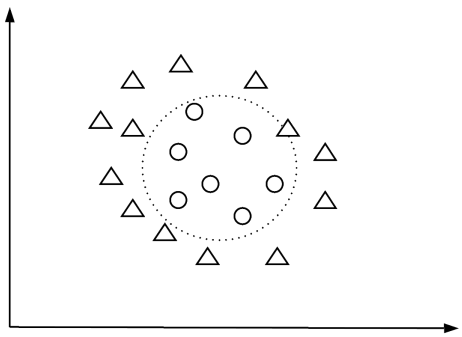

Figure 2: Kernel feature of SVM

Figure 2 above shows an SVM making a classification on data where it would seem like there is no *linear* hyperplane to be found. In order to make a classification, the SVM employs the kernel method [2]. This transformation, introduces a new dimension, by way of suggesting that for every **x** and **x'**, there is a function *k* such that *k* is equivalent to the sum of the squares of **x** and **x'** [3]. This transformation then allows the model to find an optimal hyperplane within this third dimension. Transforming this back into a 2-Dimensional plot, the hyperplane is mapped as a circular boundary around the classified data points.

4.2 Random Forest

Random Forest is a supervised learning algorithm. The way it works is by

References

[1] S. Patel, "Chapter 2 : Svm support vector machine theory." https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72, May.

[2] S. Ray, "Understanding support vector machine algorithm from examples (along with code)." https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/, 2017 (accessed November 17, 2018).

[3] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The annals of statistics*, pp. 1171–1220, 2008.