

COL774 Assignment 2

Aniruddha Deb

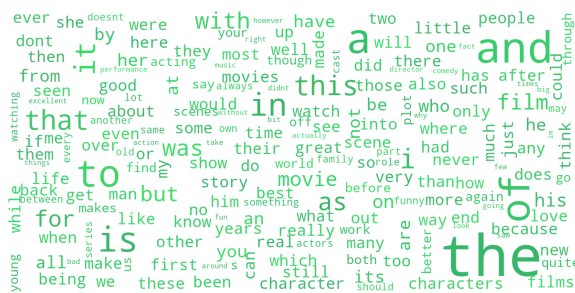
2020CS10869

October 2022

1 Naive Bayes

- (a) Naive Bayes was implemented using the **Multinomial Event Model** as discussed in class, which uses the frequency of the words rather than just their presence. The vocabulary was learnt from the training dataset, and any words not in the vocabulary in the test dataset were ignored. With this implementation, the following results were obtained:

- An accuracy of 79.313% on the test dataset, with 7501/10000 positive examples and 4396/5000 negative examples correctly classified
- The following word clouds were obtained. Note that there are a lot of stopwords, which we remove in part (d)



Positive reviews



Negative reviews

- (b) With Random Guessing, we'd obtain a test set accuracy of approximately 50%. By simply predicting each sample as positive, this would jump to 66.6%, as the number of positive reviews in the test set is twice the number of negative reviews.

Our Algorithm gives a 30% increase in accuracy over random guessing and a 14% increase in accuracy over simply predicting each review as positive.

- (c) The confusion matrices are as follows:

NaiveBayes			Random			AllPositive		
	AP	AN		AP	AN		AP	AN
PP	7501	604	PP	5000	2500	PP	10000	5000
PN	2499	4396	PN	5000	2500	PN	0	0

TODO part 2 of this: wdyd "highest value of diagonal entry"?

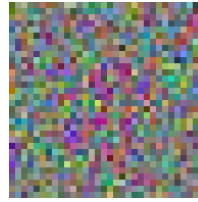
The pattern is that the column sums for the actuals always add up to the number of examples of that category in the training set. This is because the actual column measures the number of actual examples in the dataset, and across all predictions, would sum to the total number of that category in the dataset.

2 Binary SVM

- (a)
- With a threshold of $\alpha_i > 10^{-6}$ for the support vectors, we obtain **1528** support vectors, which is **38.2 %** of the training set
 - The test set accuracy we obtain is **79 %** (1580/2000 test images were classified correctly)
 - The top 5 support vectors are:



The weight image is:



- (b)
- With a threshold of $\alpha_i > 10^{-6}$ for the support vectors and the given parameters, we obtain **1769** support vectors, which is **44.2 %** of the training set. Out of these, **1143** support vectors overlap with the linear ones
 - The test set accuracy we obtain is **85.75 %** (1715/2000 test images were classified correctly)
 - The top 5 support vectors are:



- iv. The Gaussian Kernel allows for more fine-grained boundaries to be drawn between the datapoints. TODO this
- (c)
- For the Scikit-learn implementation, we obtain **1811** support vectors with a gaussian kernel and **1494** support vectors with a linear kernel. The pairwise overlaps are given below:

		QP	
		lin	rbf
Scikit	lin	1494	1123
	rbf	1130	1560

- The normed difference of weights is $\|w_{sk} - w_{qp}\| = 1571.61$, and the difference of biases is $b_{sk} - b_{qp} = -1.58$. Note that $b_{sk} = 0$.
- The test set accuracy we obtain is **86.95 %** (1739/2000) for the RBF kernel and **79.1 %** (1582/2000) for the linear kernel
- The times are given below:

Scikit RBF	13.8 s
Scikit linear	26.4 s
QP RBF	58.1 s
QP linear	47.7 s

3 Multiclass SVM