COL774 Assignment 2

Aniruddha Deb
2020CS10869

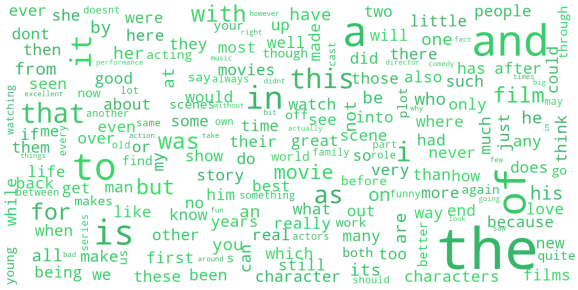
October 2022

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:

i. An accuracy of 79.313% on the test dataset, with 7501/10000 positive examples and 4396/5000 negative examples correctly classified

ii. 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

	AP	AN
PP	7501	604
PN	2499	4396

Random

	AP	AN
PP	5000	2500
PN	5000	2500

AllPositive

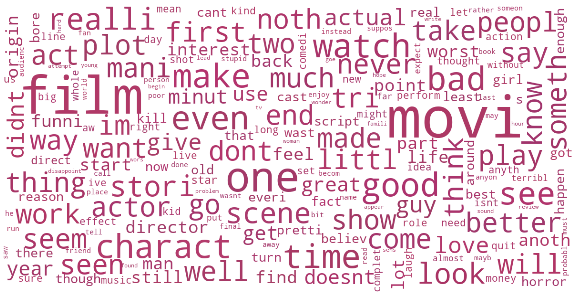
	AP	AN
PP	10000	5000
PN	0	0

The highest sum of diagonal entries is in the Naive Bayes classifier. This implies that the accuracy (number of correctly classified samples) of Naive Bayes is better than randomly or selectively guessing.

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.

(d) Removing stopwords and stemming gives us an accuracy of **80.4%**, with 7678/10000 positive reviews and 4388/5000 negative reviews correctly classified.

Positive reviews



Negative reviews

The word clouds show the frequency of the various word stems rather than the frequency of the words themselves.

(e) Using Bigrams (in addition to removing stopwords and stemming) gives us an accuracy of **84.4 %**, with 8209/10000 positive reviews and 4447/5000 negative reviews correctly classified.

For the additional feature, we use Trigrams. Using trigrams, in addition to bigrams and removing stopwords + stemming gives us an accuracy of **85.2 %**, with 8352/10000 positive reviews and 4423/5000 negative reviews correctly classified.

The additional sets of features do help us improve our accuracy (explain why)

(f) The best-performing (most accurate) model is the Trigram model, which has a precision of 8352/(8352+577) = 0.935, a recall of 8352/(8352+1648) = 0.835 and an F1 score of $2(0.935 \times 0.835)/(0.935 + 0.835) = 0.882$

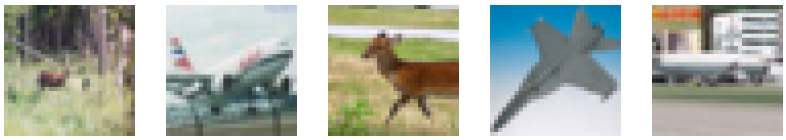
The F1 score is a better metric for this dataset, as there are twice as many positive movie reviews in the test dataset compared to negative reviews. In this case, we can see significant accuracy gains by simply declaring everything as positive instead of proceeding class by class. The F1 score helps to prevent this from happening, and gives a more realistic representation of how accurate our model is, across both classes irrespective of how many samples there are in each class.

2 Binary SVM

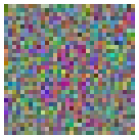
(a) i. With a threshold of $\alpha_i > 10^{-6}$ for the support vectors, we obtain **1528** support vectors, which is **38.2 %** of the training set

ii. The test set accuracy we obtain is **79 %** (1580/2000 test images were classified correctly)

iii. The top 5 support vectors are:



The weight image is:



(b) i. 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

ii. The test set accuracy we obtain is **85.75 %** (1715/2000 test images were classified correctly)

iii. The top 5 support vectors are:

iv. The Gaussian Kernel allows for more fine-grained boundaries to be drawn between the datapoints, which 'fit' the shape of the data boundaries better compared to a linear separator. This is why the accuracy of the gaussian kernel is better than the linear kernel.

(c) i. 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

ii. 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 $b_{sk} = 0$.

iii. The test set accuracy we obtain is **86.95 %** (1739/2000) for the RBF kernel and **79.1 %** (1582/2000) for the linear kernel

iv. The training times are given below:

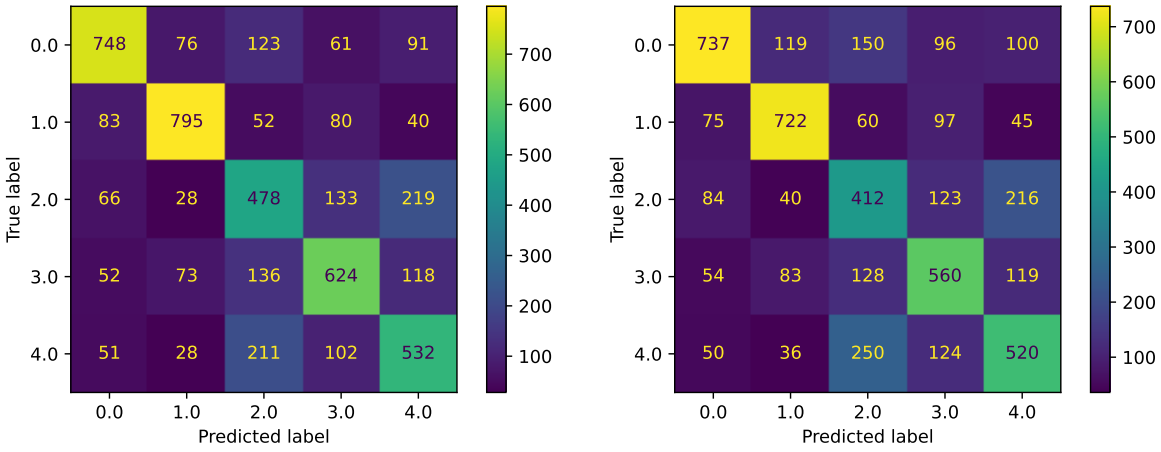
Scikit RBF	13.8 s
Scikit linear	26.4 s
QP RBF	58.1 s
QP linear	47.7 s

3 Multiclass SVM

(a) Using a multi-class variation of the CVXOPT solver with the same parameters as before ($\gamma = 0.001, C = 1$, and breaking ties based on the sum of scores (highest first), we obtain a test set accuracy of **59.02 %** (2951/5000).

(b) Using the Scikit-Learn SVM package, we obtain a test set accuracy of **63.54 %** (3177/5000). The Scikit implementation is much faster, requiring only 187.8 s to train. The CVXOPT implementation, by comparision, takes 494.4 seconds to train.

(c) The confusion matrices for both implementations is as follows:



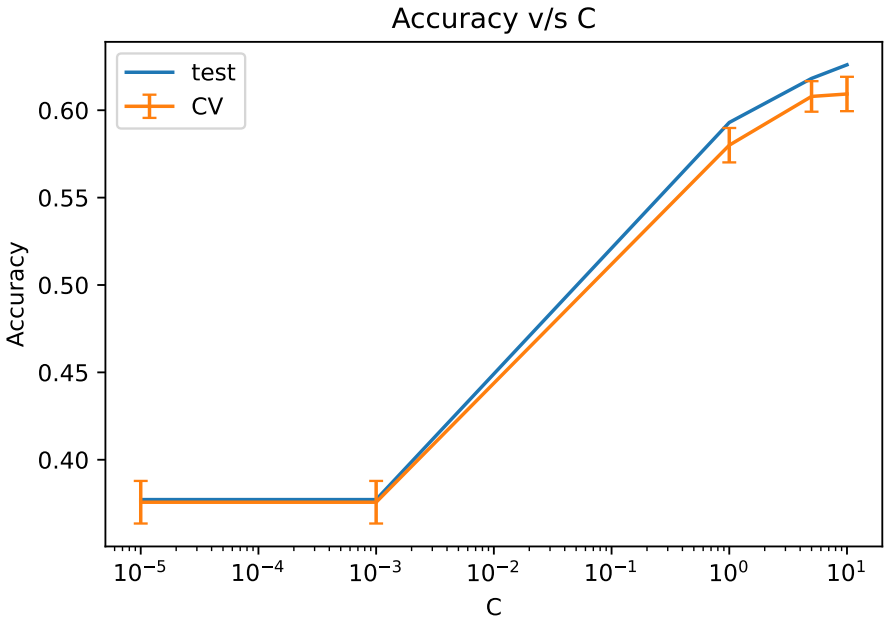
Scikit CVXOPT (QP)

Classes 2 and 4 are most often misinterpreted as each other. Below, we plot 10 randomly chosen examples which are misclassified by Scikit and CVXOPT, with their predicted and original labels respectively

The classes of the CIFAR-10 dataset are Airplane(0), Automobile(1), Bird(2), Cat(3) and Deer(4). The model(s) misclassify several examples in the bird, cat and deer categories amongst each other, which is to be expected as the backgrounds in which these animals are present is similar. Similarly, Airplanes and Automobiles are most commonly confused with each other rather than with birds, cats or deer because they have similar features (large regions of the same colour, with an urban or blue background).

(d) The plot is given on the following page. We see an increasing trend, where the best accuracy is given by the 'hardest' classifier i.e. for $C = 10$. This value also gives the best test set accuracy.

The value of C tells the classifier how much we want to avoid misclassifying each example: larger values of C will mean less misclassifications on the training dataset, but this may lead to overfitting. In our case, for small values of C, the model underfits, and chooses a very broad decision boundary, leading to poor cross-validation/test accuracy


Accuracy v/s C
test
CV
0.60
0.55
0.50
0.45
0.40
Accuracy
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
C