AI_534
MACHINE LEARNING

IA3 Report

CHI-CHIEH WENG
TSU-CHING LIN

Part 0 Preprocessing

(a)

Ten most frequent words of the train set

Positive

['the', 'to', 'to', 'for', 'thanks', 'jetblue', 'southwestair', 'united', 'thank', 'and']

Negative

['to', 'the', 'flight', 'united', 'on', 'and', 'you', 'for', 'my', 'usairways']

Ten most frequent words of the dev set

Positive

['the', 'to', 'you', 'for', 'thanks', 'jetblue', 'southwestair', 'and', 'united', 'flight']

Negative

['to', 'the', 'united', 'flight', 'and', 'for', 'you', 'on', 'usairways', 'my']

We believe that some words has class informative. For example, 'thanks', and 'thank'. These two words are usually used for positive emotions.

(b)

Ten most frequent words of the train set

Positive

['you', 'thanks', 'thank', 'the', 'jetblue', 'united', 'to', 'southwestair', 'for','americanair']

Negative

['to', 'the', 'flight', 'you', 'united', 'on', 'and', 'for', 'usairways', 'my']

Ten most frequent words of the dev set

Positive

['you', 'thanks', 'the', 'jetblue', 'to', 'thank', 'for', 'united', 'southwestair', 'usairways']

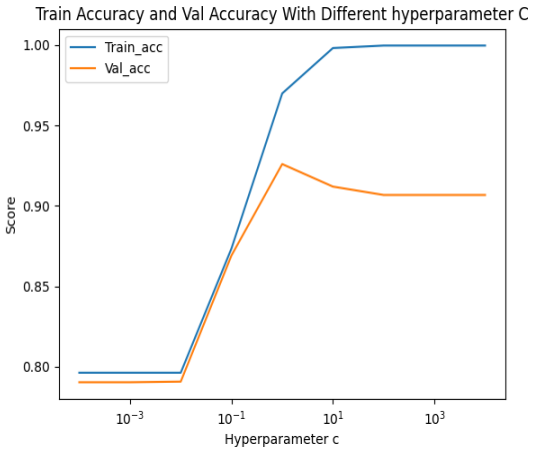
Negative

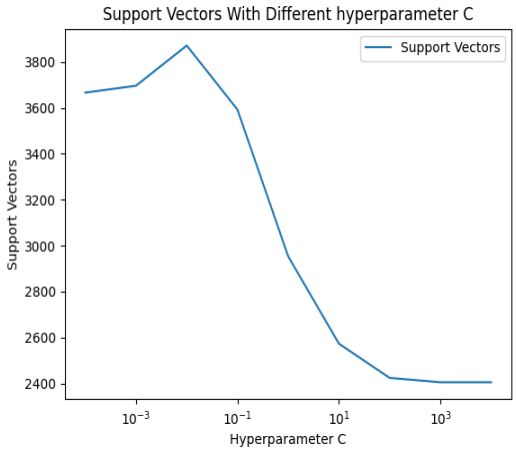
['to', 'the', 'you', 'united', 'flight', 'and', 'for', 'on', 'usairways', 'americanair']

Compared with part 0 (a), we think the first 10 words are similar, but in a different order, which means that the importance of the words has changed. We think that TfidfVectorizer judges positive and negative words better than CountVectorizer. For example, the order of 'thanks' and 'thank' is more to the front.

AI_534
MACHINE LEARNING

Part 1 Linear SVM





(1)

When the hyperparameter c is 1, we obtain the best validation performance of 0.926.

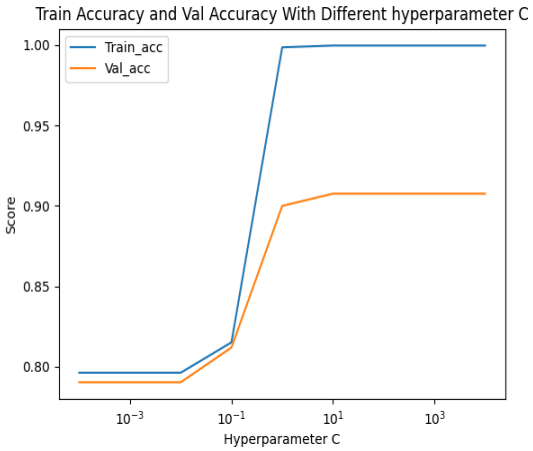
(2)

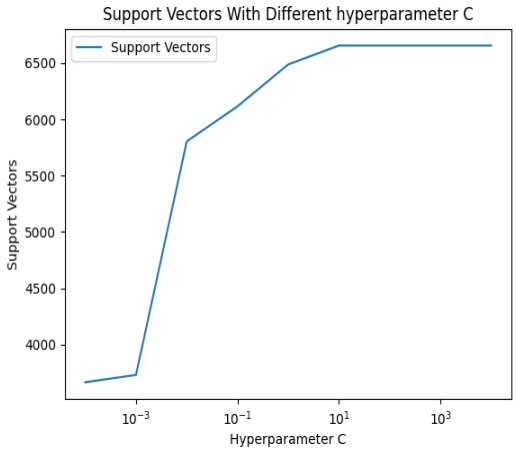
As c increases, we theoretically expect to see an upward trend in training and verification performance, and the trends we observe are in line with our expectations. However, we find that the trend of verification performance reaches a peak when c is 1 and then starts a small decline. In addition, we also find that when C is getting larger, it means less error tolerance, and also support vectors will be less, which is closer to the concept of hard-margin SVM, but easier to overfit. Like the performance trends in the figure, the training and verification performance trends will eventually overfit.

(3)

We theoretically expect that the number of support vectors decreases when c becomes large. According to the trend in the figure, the result is like our expectation, except that the number of support vectors has a tendency to increase and then decrease, but overall it is decreasing.

Part 2 Quadratic SVM





(1)

When the hyperparameter c is 10, we obtain the best validation performance of 0.9076.

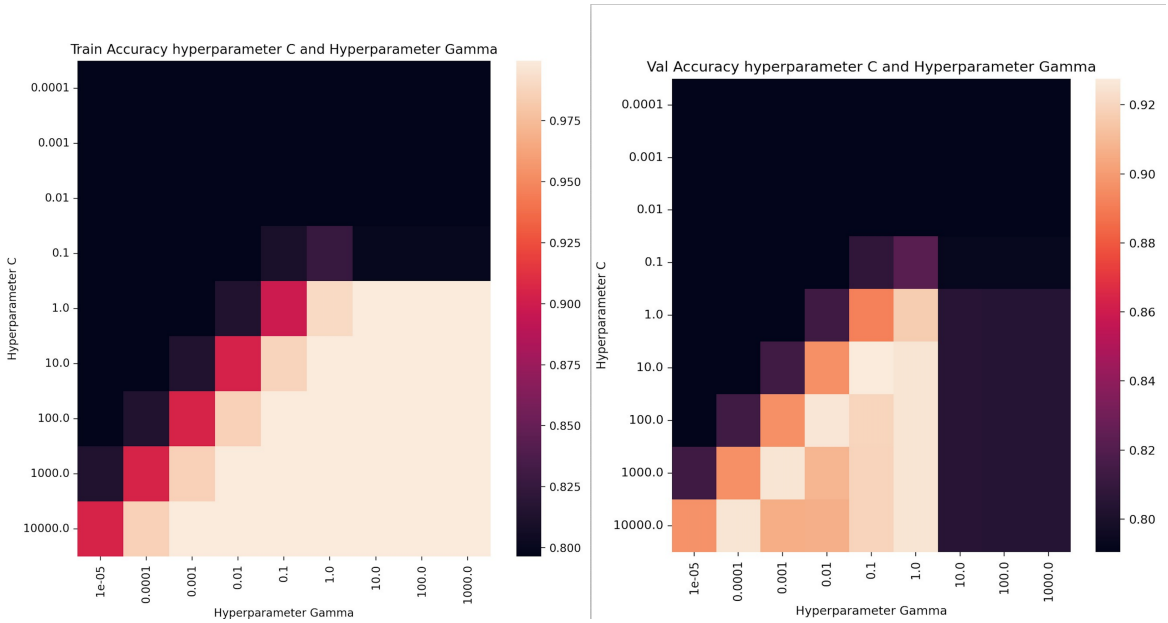
(2)

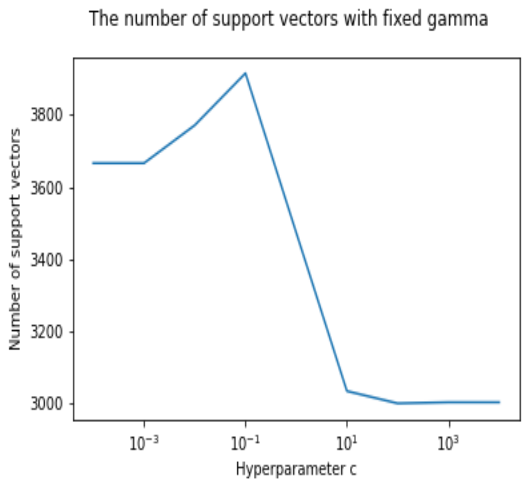
As with the results in Part 1, we theoretically expect to see an upward trend in training and verification performance as c increases, and the trend we observe is consistent with our expectations.

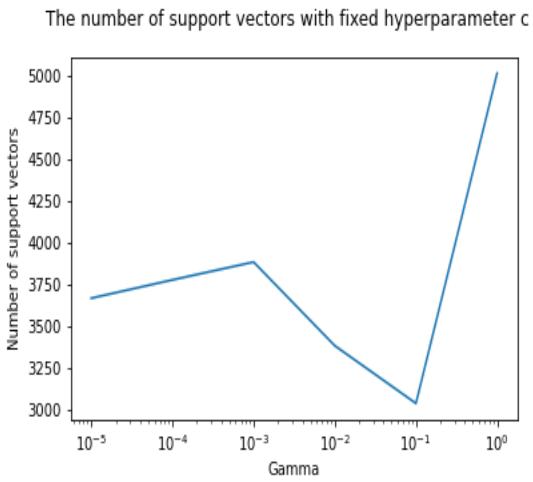
(3)

We theoretically expect that the number of support vectors decreases when c becomes large. However, the result is different from our expectations. We think the reason may be that different kernels lead to different decision boundaries, so as C becomes larger, the support vector also becomes larger.

Part 3 SVM with RBF kernel







(1)

The best validation performance of RBF kernel is 0.9267. For this performance in our experience, the c is 10 and the gamma is 0.1.

(2)

Theoretically, when we increasing the c with fixed gamma, we expect the training and validation performance will be up trend. And the result is fit our expectations.

(3)

Theoretically, when we decrease the gamma with fixed c, we expect the training performance and validation performance will decrease. For training performance,

the trend is down, but for verification performance, the trend is up and then down, but overall it is down, so the results are in line with our expectations.

(4)

Theoretically, the number of support vectors decreases when increasing c using a fixed gamma. Our results are as expected, although there is a small increase, the overall decrease.

(5)

Theoretically, as c increases by a fixed gamma, the number of support vectors increases. Our results are as expected.

Part 4 Final discussion question

The main sources of error in these three models are estimation errors and Bayes errors. In our experience, we found the difference between train accuracy and validation accuracy is very large, about 5 to 10 percent. In addition, even when we use more complex models (polynomial kernel and RBF kernel), the validation accuracy is still poor.