Machine Learning
IA-3
Nov 10 2022 *Aviyan Khadka*

Part 0: Preprocessing

Question 1: Please report the ten most frequent words (frequency is measure by the total counts of each word are maximal) for the positive tweet and negative tweet respectively. Do you find these words to be informative of the classes?

Positive	on	my	and	flight	you	for	the	thanks	thank	to
Negative	in	is	my	for	you	and	on	flight	to	the

Table 1: Top 10 most frequent words using CountVectorizer

There seems to be a rough framework to distinguish between the the positive and the negative tweets. The positive tweets have repeated word like ”thanks” which is absent in the negative. The presence of these distinct words are definitively helpful in separating the two types of tweets.

Question 2: Please report the ten words with the highest total TF-IDF’s for the positive tweet and negative tweet respectively. How do they compare to the list in 0(a)? Which one do you think are more informative and why?

Positive	my	you	thanks	it	thank	great	for	and	to	the
Negative	it	is	flight	for	the	my	you	on	to	and

Table 2: Top 10 most frequent words using TfidfVectorizer

Similar to the CountVectorizer list, there is a repetition of the distinct words in the positive tweets that are absent in the negative. However, the words list for Tfid is a lot more comprehensive than the previous as it have words like ”great” in addition to the words like ”thanks”. The presence of these more distinct words make the Tfid data more information that CountVectorizer data.

Part 1: Linear SVM

Q1: What is the best validation performance you are able to achieve with linear SVM? What c value is used?

The best validation performance was achieved with 92.52% accuracy at c=0

Q2: What trend do you theoretically expect to see for training and validation performance as we increase c? Plot the training and validation accuracy against different values of c. Does the trend you observe match your expectation?

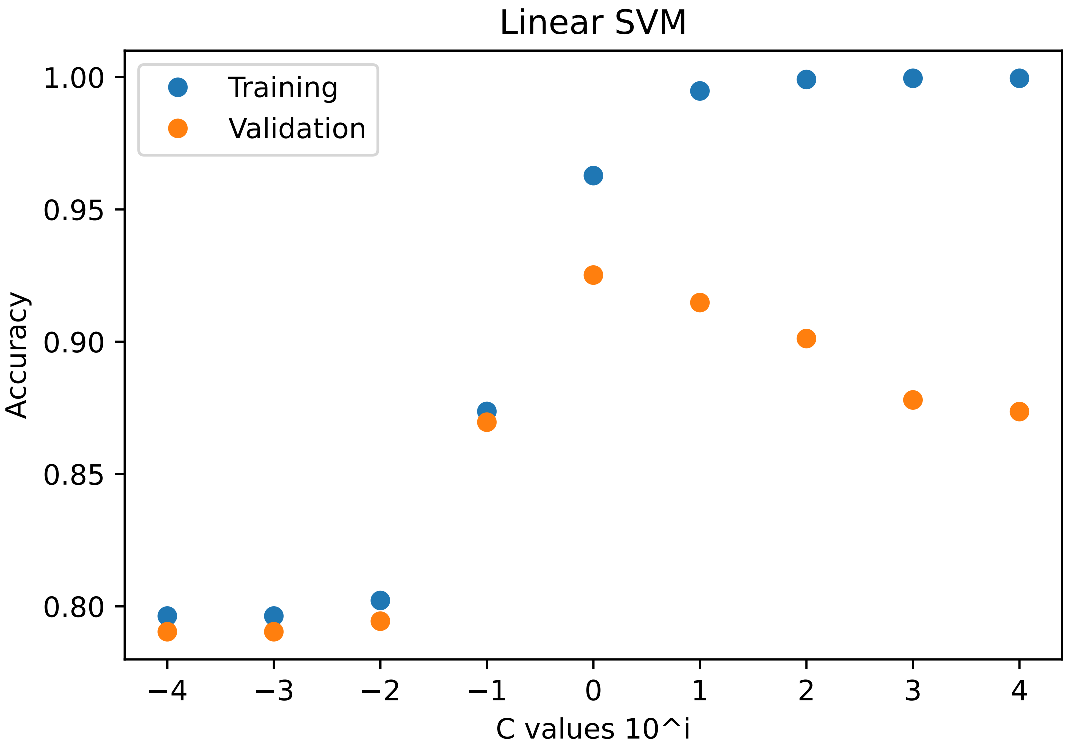


Figure 1: Linear SVM training and validation accuracy

In our analysis, we start with a c value of 10^{-4} and increase it by 1 unit until 10^{4}. Under the theoretical assumption, the svm margin grows thinner as we increase the value of c. This thinning of the linear separator will result in a low to almost no misclassification and therefore, the accuracy will be 100%.When the accuracy is very high, there is a high chance of overfitting and therefore, model might not be a good fit for the validation data leading to a lower accuracy.

So, when doing our analysis, we saw that, as the training accuracy was going up to a 100%, the validation accuracy was tapering off after increasing upto a certain c value. This tapering off matches our expectation that the overfit model lead to a lower accuracy

Q3: What relationship do we theoretically expect between c and the number of support vectors? Plot the number of support vectors against the different values of c. Does the trend you see match your theoretical expectations?

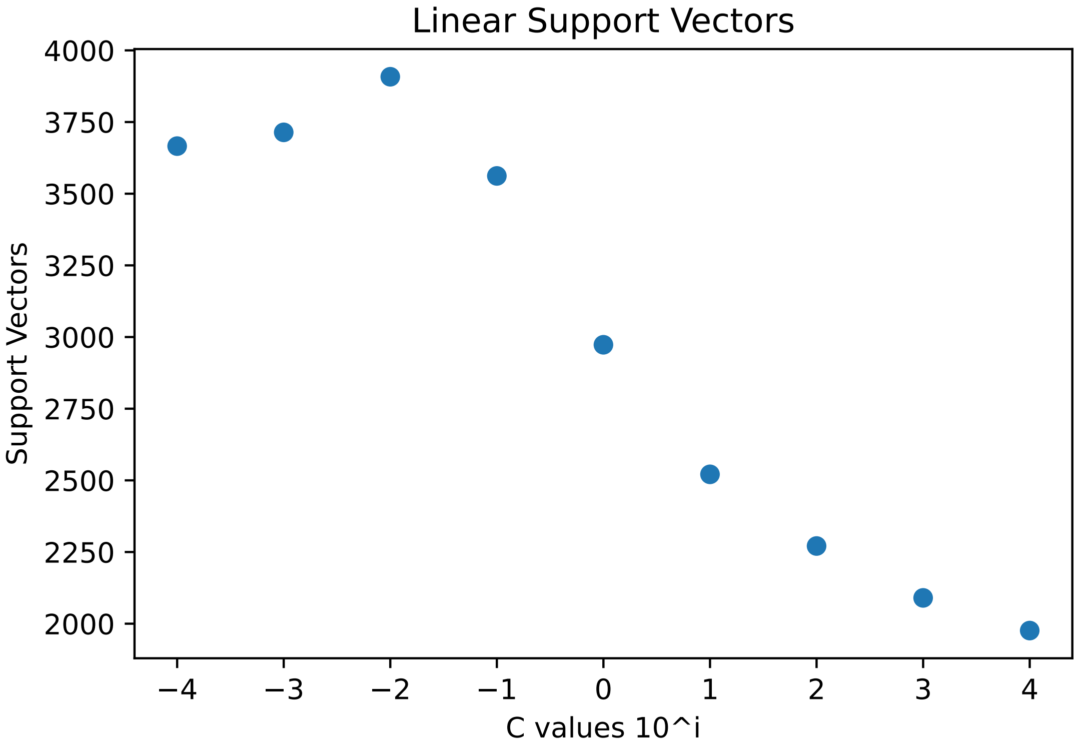


Figure 2: Number of support vectors while changing C

Under the theoretical assumption, the number of support vectors should be decreasing with with an increasing c. Our analysis matches that expectation as the number of support vectors decreases when the value of c increase from 10ˆ-2. However, there is a slight increase in support vectors from 10^{-4} to 10^{-2} which may be explained by some processing error.

Part 3: Quadratic SVM

Q1. What is the best validation performance you are able to achieve with the quadratic kernel? whatc value is used?

The best validation performance was acheived with a accuracy of 91.56% at $c = 10^3$

Q2. What trend do you theoretically expect to see for training and validation performance as we increase c? Plot the training and validation accuracy against different values of c. Does the trend you observe match your expectation?

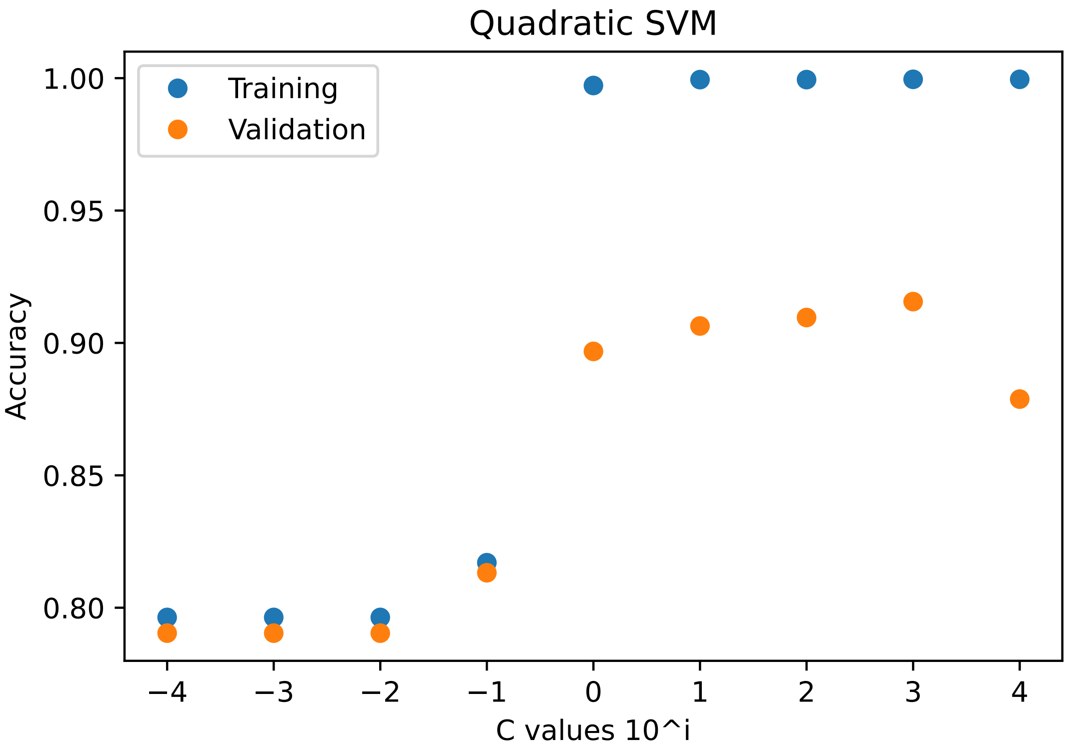


Figure 3: Quadratic SVM training and validation accuracy

Similar to the linear SVM, the accuracy will increase as we increase the value of c for the training data. However, there will always be a risk of overfitting our model under high value of c resulting in lower accuracy for the testing data. Therefore, aligning with our expectation, there is an increase in training accuracy with an increase in c but the validation accuracy will decrease after a certain value of c which is 10^3 in our current model.

Q3. What relationship do we theoretically expect between c and the number of support vectors? Plot the number of support vectors against the different values of c. Does the trend you see match your theoretical expectations?

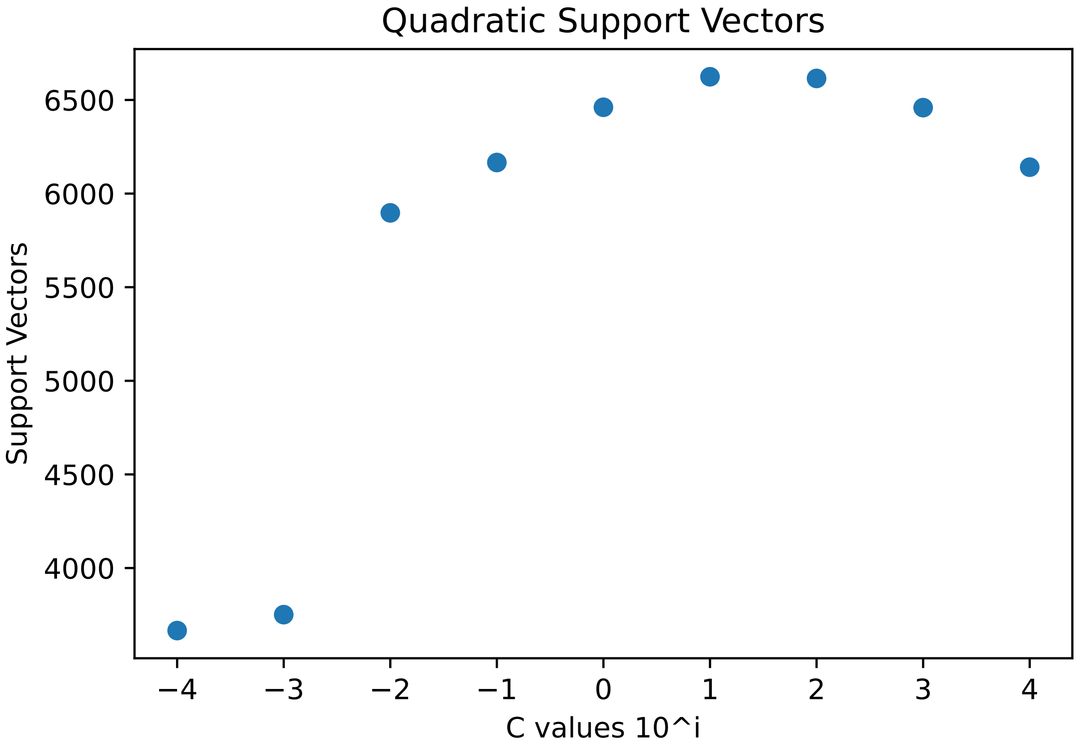


Figure 4: Number of support vectors while changing C

The number of support vectors should decrease with an increase in the value of c. However, in our analysis there is an increase in the number of support vectors until c=10. After 10, there is a gradual decrease in the support support vectors aligning with our expectation. The increase might be caused due to the presence of influential features in our feature vector.

Part 4: SVM with RBF Kernel

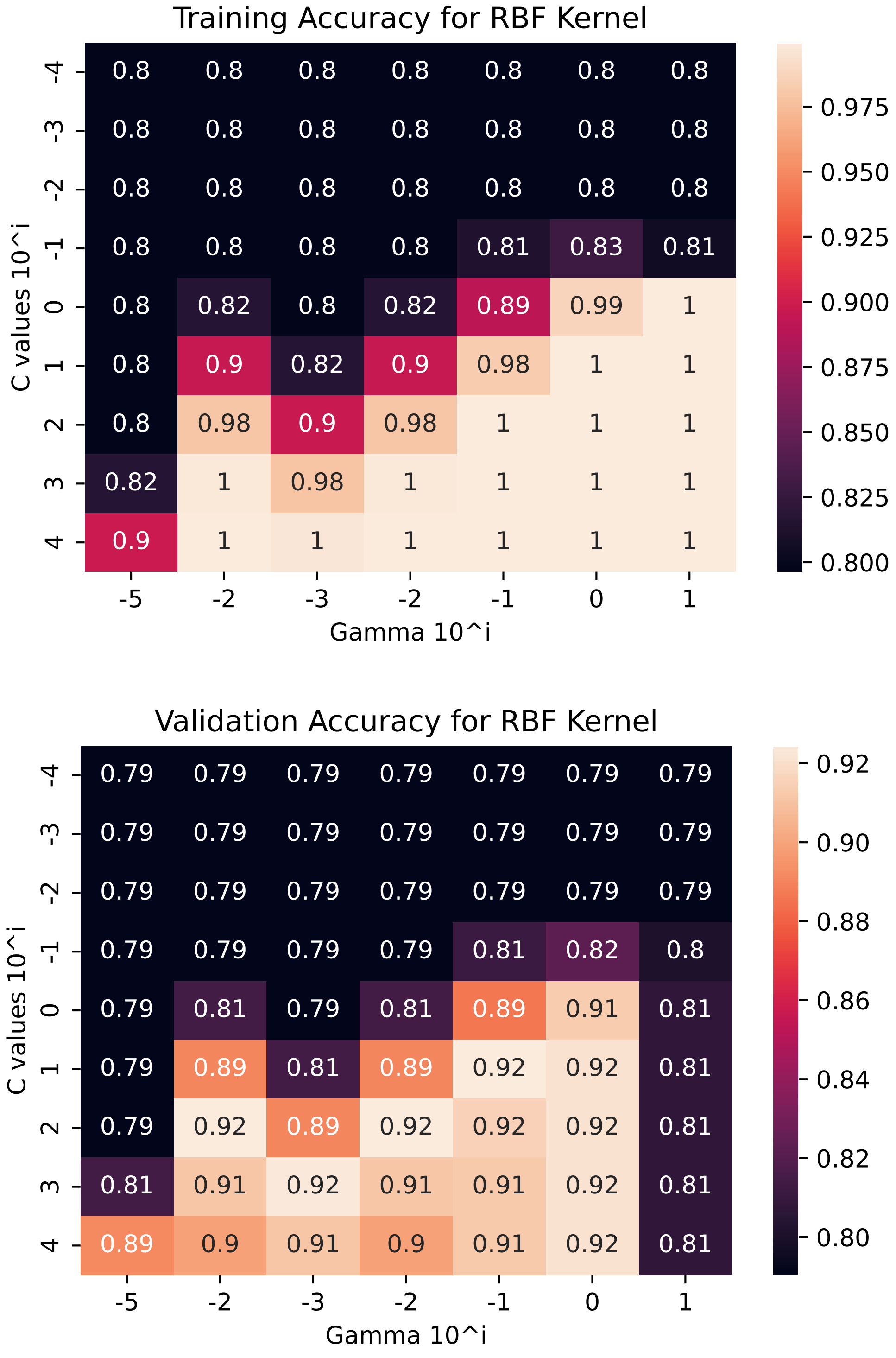


Figure 5: Caption

Q1: What is the best validation performance you are able to achieve for RBF kernel? What c and γ parameters are used?

The best validation performace was acheived with 92.44% accuracy at c= 10 and γ=0.1.

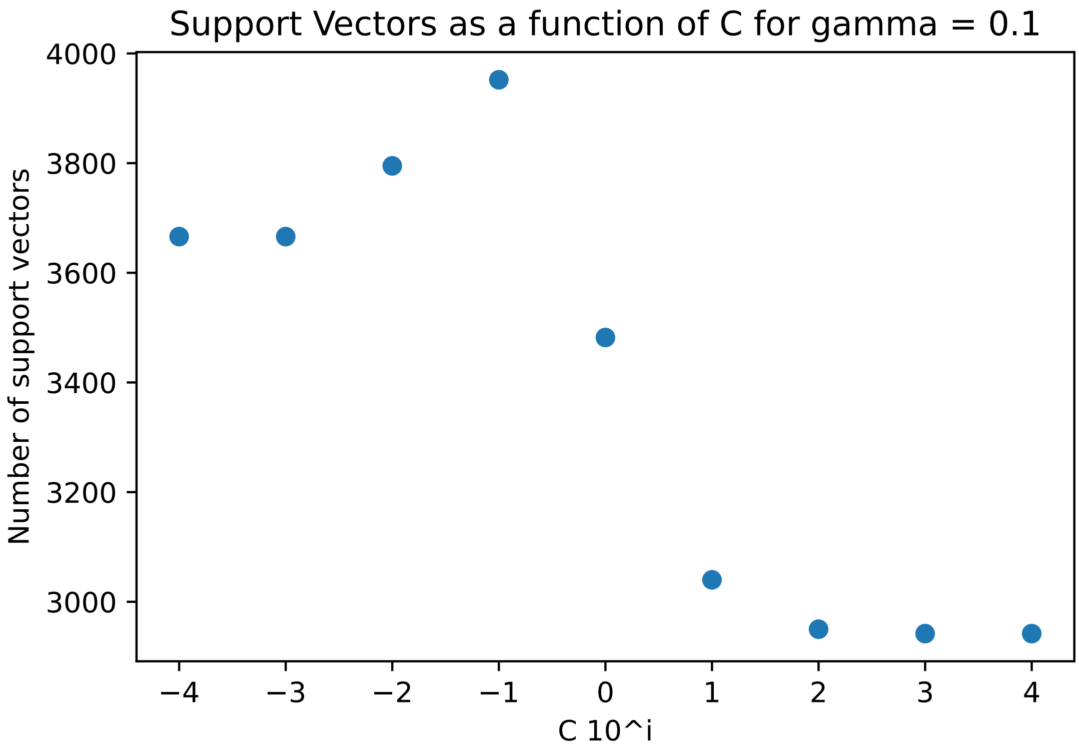
Q2: What trend do you theoretically expect to see for training and validation performance as we increase c with fixed γ? Does the trend you observe match your expectation?

Similar to linear and quadratic, there is an increase in accuracy with increase in c for a fixed gamma, and similar to the previous models, we can expect overfitting leading to decrease in accuracy. This trend has been verified by our analysis wit the training accuracy increasing and the validation accuracy decreasing after a certain c value.

Q3 What trend do you theoretically expect to see for training and validation performance as we decrease γ with fixed c? does the trend you observe match your expectation?

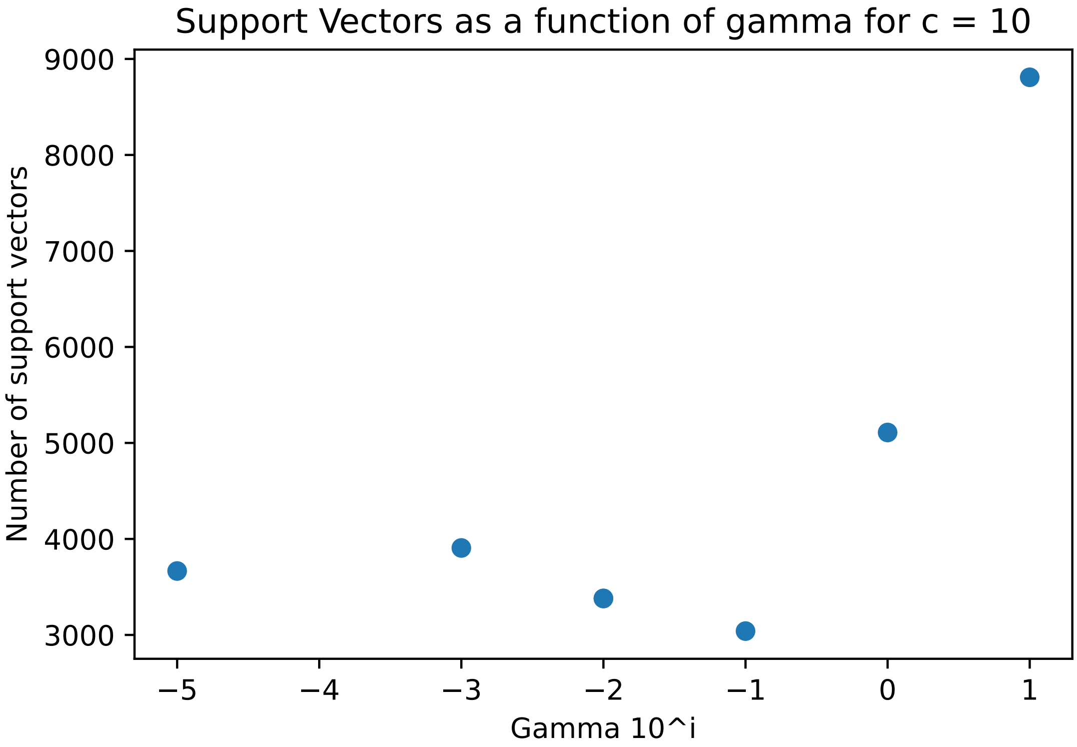
Under the theoretical assumptions, high and low values of gamma result in underfitting and overfitting of the data respectively, resulting in lower validation accuracy.The training accuracy will however increase with increasing gamma due to the overfit. For our validation data, the intermediate values will be the ones that provide us with the highest accuracy. Therefore, in our analysis, the lower and higher values of gamma have resulted in lower validation accuracy accuracy than the intermediate values.

Q4: With fixed γ What relationship do we theoretically expect between c and the number of support vectors? Plot the number of support vectors as a function of c value for γ = 0.1. does the trend you see match your theoretical expectations?



Increasing c values will result in a decrease of the support vectors. This trend has been seen in our analysis with a support vector number decreasing after c=10-̂1. However, there is a slight increase in the number for 10^{-4} to 10^{-1}, due to presence of a few influential weight in our features.

Q5: With fixed c, what relationship do we theoretically expect between γ and the number of support vectors? Plot the number of support vectors as a function of γ for c = 10. Does the trend you see match your theoretical expectations?



Under small and large values of gamma, for a fixed c the number of support vectors should increase. However, for intermediate values of gamma, there should be a decrease in the number. Our analysis matches with the theoretical expectation with 0.01 and 0.1 have low support vectors while gamma values 10^{-5} and 10 have comparatively larger support vector numbers.