Machine Learning : Foundation and Application
AI42001

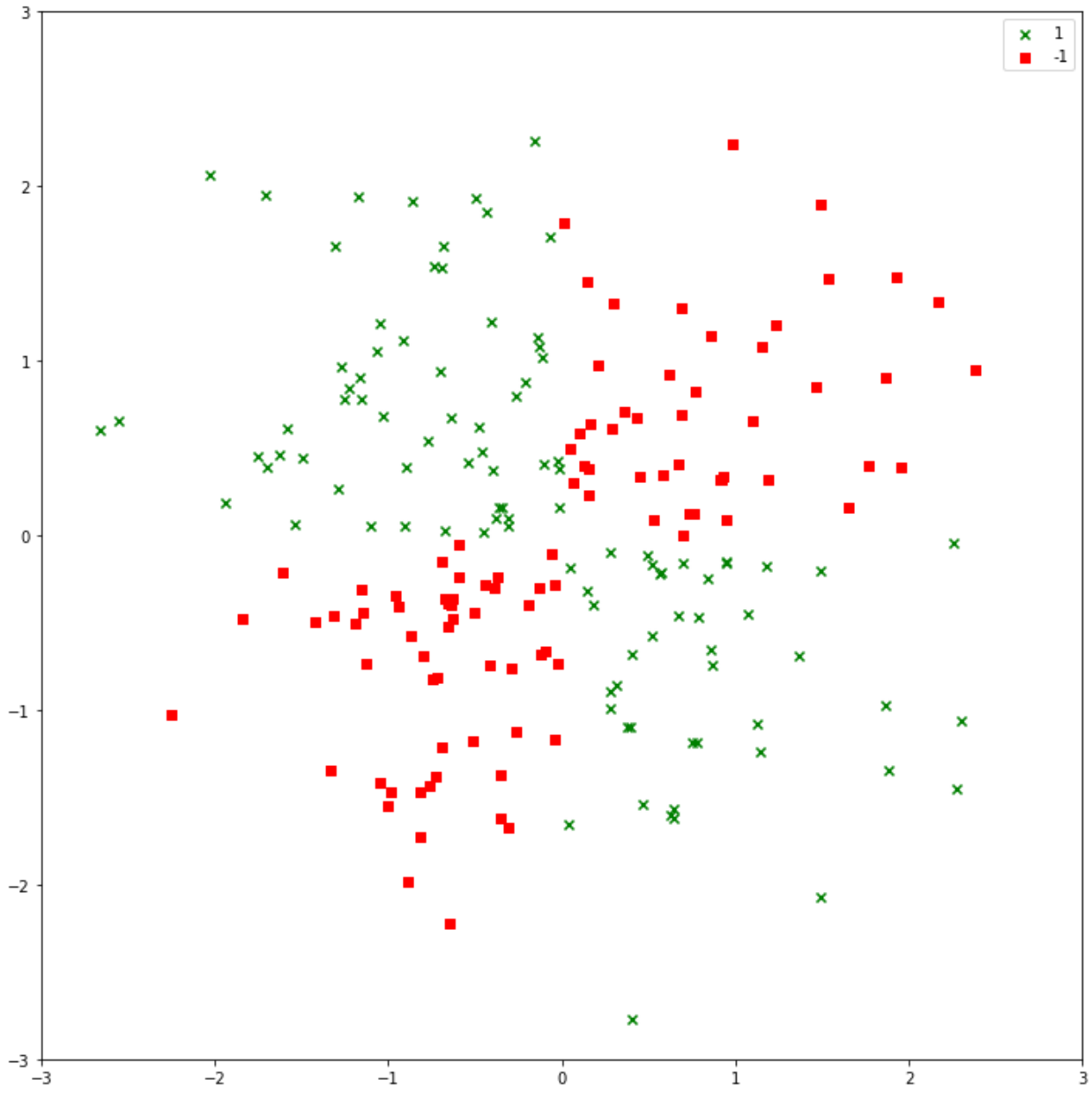
Mini - Project on SVM Kernels

Submitted By-

Jayant Choudhary (18NA30008)

Dataset

Following is the plot of the dataset-



We can see very clearly that the dataset is non-linearly separable.

Kernels

In practice, the SVM algorithm is implemented using a kernel. It uses a technique called the kernel trick. In simple words, a kernel is just a function that maps the data to a higher dimension where data is separable. A kernel transforms a low-dimensional input data space into a higher dimensional space. So, it converts non-linear separable problems to linear separable problems by adding more dimensions to it. Thus, the kernel trick helps us to build a more accurate classifier. Hence, it is useful in non-linear separation problems.

In the context of SVMs, there are 4 popular kernels – Linear kernel,Polynomial kernel,Radial Basis Function (RBF) kernel (also called Gaussian kernel) and Sigmoid kernel. These are described below -

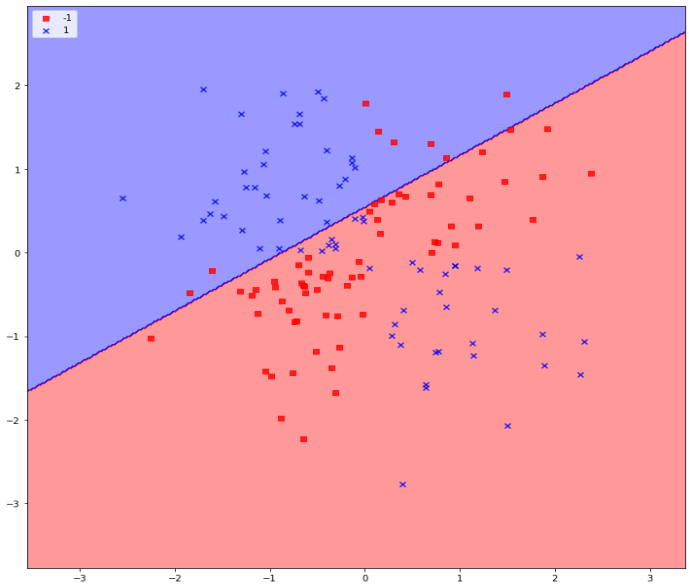
Linear Kernel

In linear kernel, the kernel function takes the form of a linear function as follows-

linear kernel : K(xi , xj) = xiT xj

Linear kernel is used when the data is linearly separable. It means that data can be separated using a single line. It is one of the most common kernels to be used. It is mostly used when there are large number of features in a dataset. Linear kernels are often used for text classification purposes.

Training with a linear kernel is usually faster, because we only need to optimize the C regularization parameter. When training with other kernels, we also need to optimize the γ parameter. So, performing a grid search will usually take more time. Following is the plot for the linear kernel -



Scores for Linear Kernel

Grid search was performed to obtain the best hyperparameters.
Train Accuracy is 0.678 and Test Accuracy is 0.683.

Confusion Matrix

	precision	recall	f1-score	support
-1	0.61	0.93	0.74	29
1	0.88	0.45	0.60	31
accuracy			0.68	60
macro avg	0.74	0.69	0.67	60
weighted avg	0.75	0.68	0.67	60

Conclusion: We can see very low test and train accuracy which is because of the obvious reason that the data is non-linear and cannot be separated through linear boundaries.

Polynomial Kernel

Polynomial kernel represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables. The polynomial kernel looks not only at the given features of input samples to determine their similarity, but also combinations of the input samples.

For degree-d polynomials, the polynomial kernel is defined as follows –

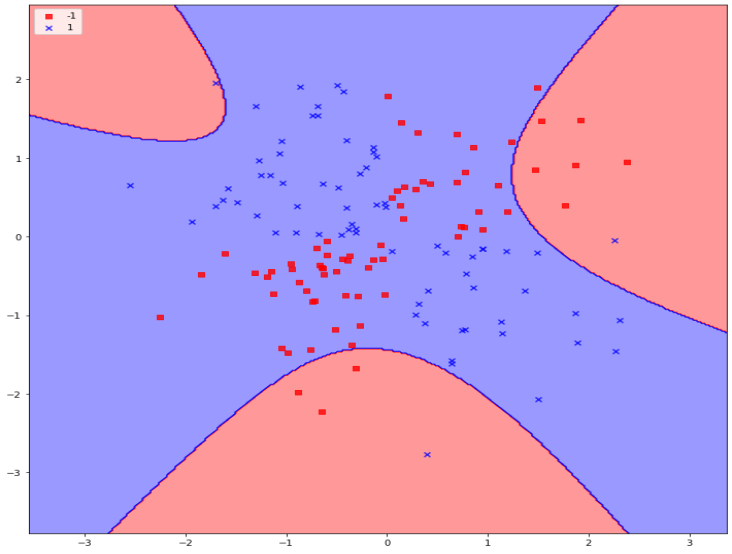
Polynomial kernel : K(xi , xj) = (γxiT xj + r)d , γ > 0

Polynomial kernel is very popular in Natural Language Processing. The most common degree is d = 2 (quadratic), since larger degrees tend to overfit on NLP problems.Following are the plots for the polynomial kernel -

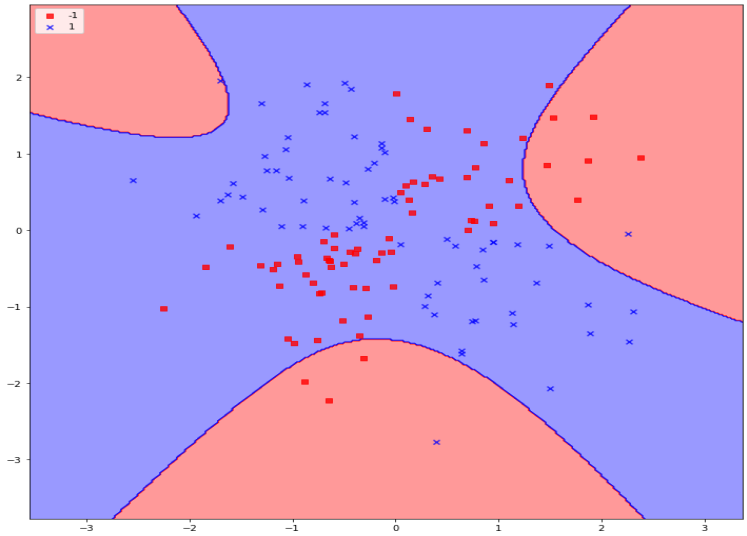
1) For gamma = 1 and C=1 :

Train Accuracy is 0.542 and Test Accuracy is 0.566.

	precision	recall	f1-score	support
-1	0.71	0.17	0.28	29
1	0.55	0.94	0.69	31
accuracy			0.57	60
macro avg	0.63	0.55	0.48	60
weighted avg	0.63	0.57	0.49	60



For `gamma=0.01, C=1000000000 :`



Train Accuracy is `0.542` and Test Accuracy is 0.566.

Confusion Matrix:

	precision	recall	f1-score	support
-1	0.71	0.17	0.28	29
1	0.55	0.94	0.69	31
accuracy			0.57	60
macro avg	0.63	0.55	0.48	60
weighted avg	0.63	0.57	0.49	60

Conclusion:

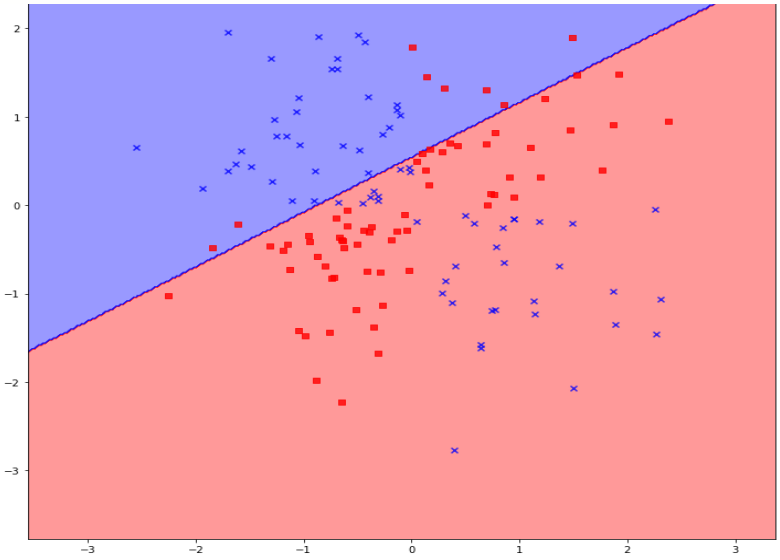
Grid search was performed to obtain the best hyperparameters but still we see that the polynomial kernel too gives a poor accuracy.

Sigmoid Kernel

Sigmoid kernel has its origin in neural networks. We can use it as the proxy for neural networks. Sigmoid kernel is given by the following equation –

sigmoid kernel : k (x, y) = tanh(αxTy + c)

For gamma=0.001, C=10000 :

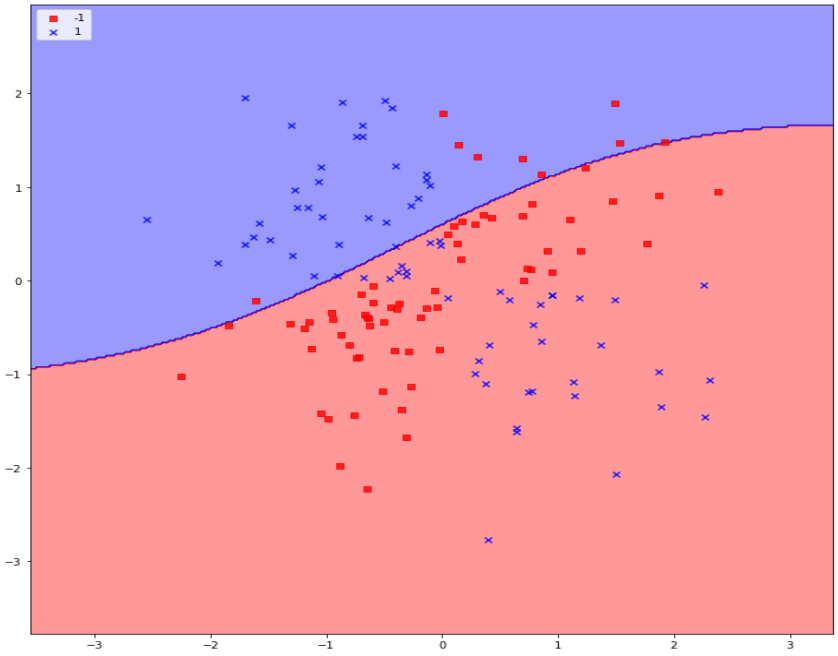


Scores:

Train Accuracy is 0.678 and Test Accuracy is 0.683.

	precision	recall	f1-score	support
-1	0.61	0.93	0.74	29
1	0.88	0.45	0.60	31
accuracy			0.68	60
macro avg	0.74	0.69	0.67	60
weighted avg	0.75	0.68	0.67	60

For gamma=0.01, C=1000 :



Scores:

Train Accuracy is `0.671` and Test Accuracy is 0.683.

	precision	recall	f1-score
-1	0.61	0.93	0.74
1	0.88	0.45	0.60
accuracy			0.68
macro avg	0.74	0.69	0.67
weighted avg	0.75	0.68	0.67

Conclusion:

Grid search was done to obtain the best hyperparameters.We saw that the Sigmoid kernel gave better accuracy then the linear and polynomial kernels.

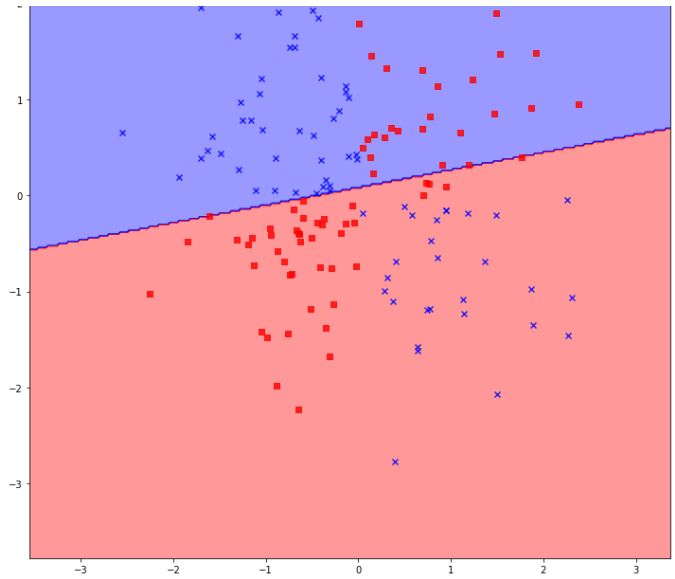
RADIAL BASIS FUNCTION(RBF) KERNEL

Radial basis function kernel is a general purpose kernel. It is used when we have no prior knowledge about the data. The RBF kernel on two samples x and y is defined by the following equation –

$$k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$$

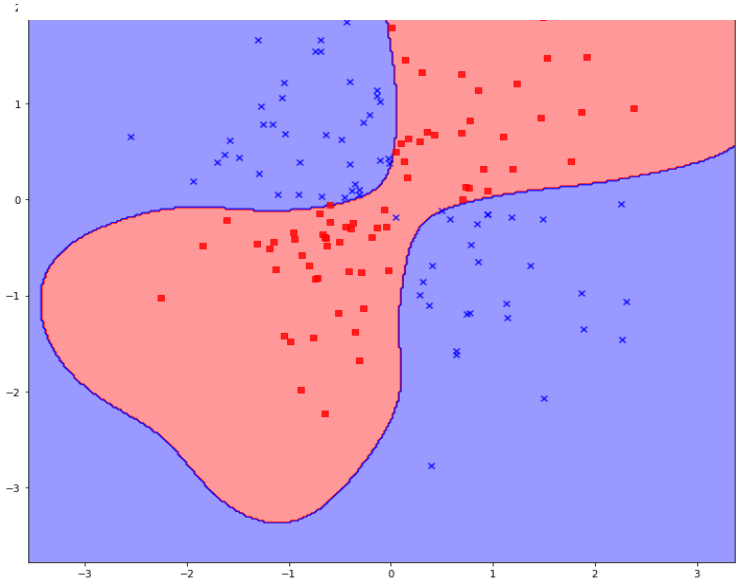
Effect of increasing gamma

For gamma=1/10000, C=1 :



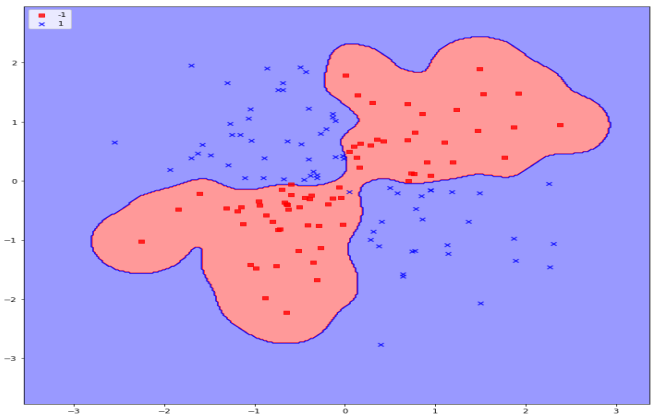
Train Accuracy is `0.614` and Test Accuracy is 0.55.

For gamma=1, C=1 :



Train Accuracy is `0.978` and Test Accuracy is 0.916

For gamma=10, C=1 :

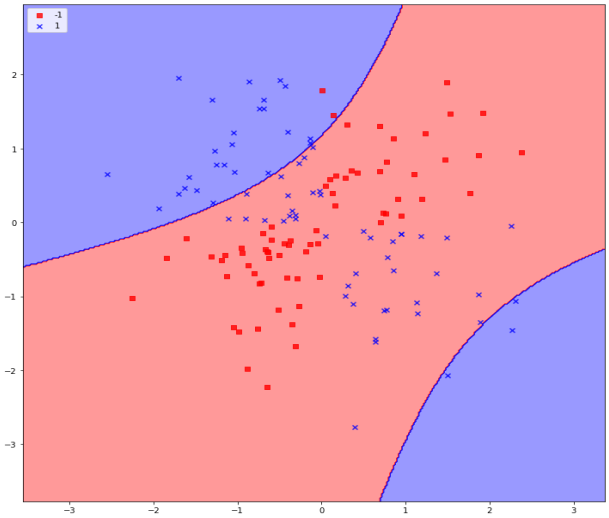


Train Accuracy is `0.985` and Test Accuracy is 0.916

The gamma parameter defines how far the influence of a single training example reaches, with low values meaning ‘far’ and high values meaning ‘close’. In other words, with low gamma, points far away from the plausible separation line are considered in calculation for the separation line. Whereas high gamma means the points close to the plausible line are considered in calculation. This is very clearly visible in the graphs.

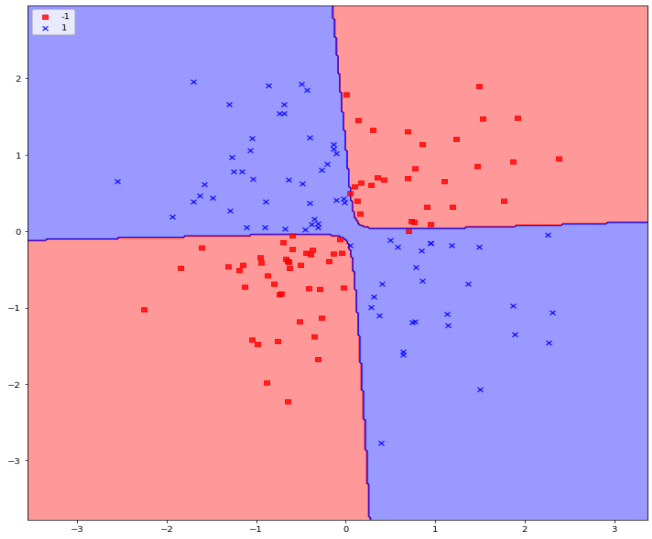
Effect of increasing C

For gamma=0.01, C=10 :



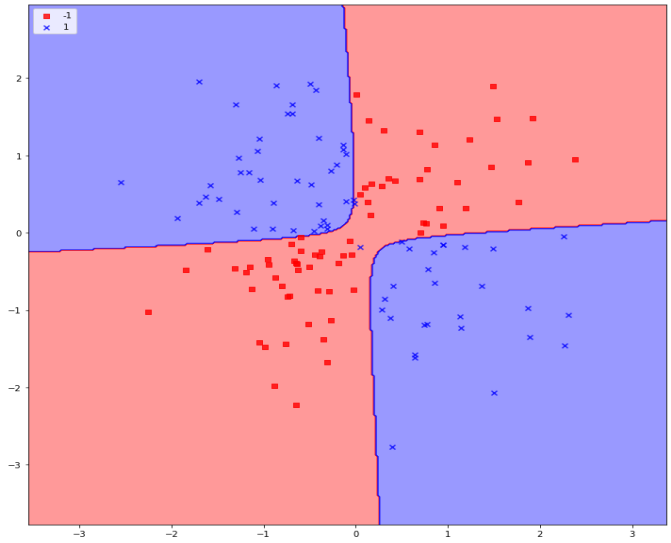
Train Accuracy is `0.685` and Test Accuracy is 0.666

For gamma=0.01, C=2000 :



Train Accuracy is `0.985` and Test Accuracy is 0.966

For gamma=0.01, C=10000 :



Train Accuracy is `0.985` and Test Accuracy is 0.916

We saw that after further increasing C the test accuracy reduced. The reason being that C is a penalty parameter the larger C means we want to classify a larger number of train points correctly and hence we

will have a smaller margin which leads to misclassification of newly seen data since the margin was low. Which can be basically called as overfitting. We saw that the test accuracy was reduced for very high C.

Conclusion: We can see that the RBF kernel worked best among all the kernels.The best hyperparameters were chosen using the grid search. For gamma = 0.01 and C = 2000 we get the best train and test accuracy. The accuracies and confusion matrix are as follows -

Train Accuracy is `0.985` and Test Accuracy is 0.966

Confusion matrix:

	precision	recall	f1-score
-1	0.97	0.97	0.97
1	0.97	0.97	0.97
accuracy			0.97
macro avg	0.97	0.97	0.97
weighted avg	0.97	0.97	0.97