Stage One – SVMs

Linear SVM:

I trained the ML model on two variants of datasets.

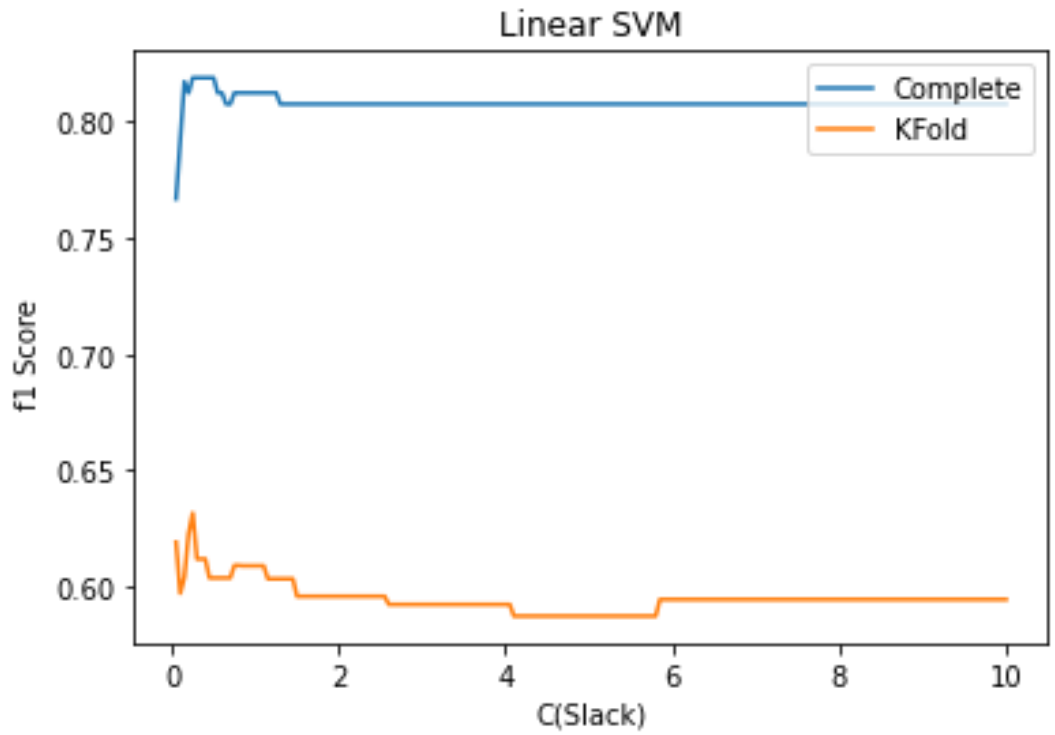
1. Complete dataset
2. K fold (Number of folds = 5)

I have run a loop for slack variable. I iterated the slack variable C from 0 to 10 with an increment of 0.05.

The maximum accuracy obtained was

```
Accuracy on entire training Data : 0.8306878306878307
Accuracy on  KFold Data : 0.6667140825035561
```

Here is the plot of F1 scores and C values:



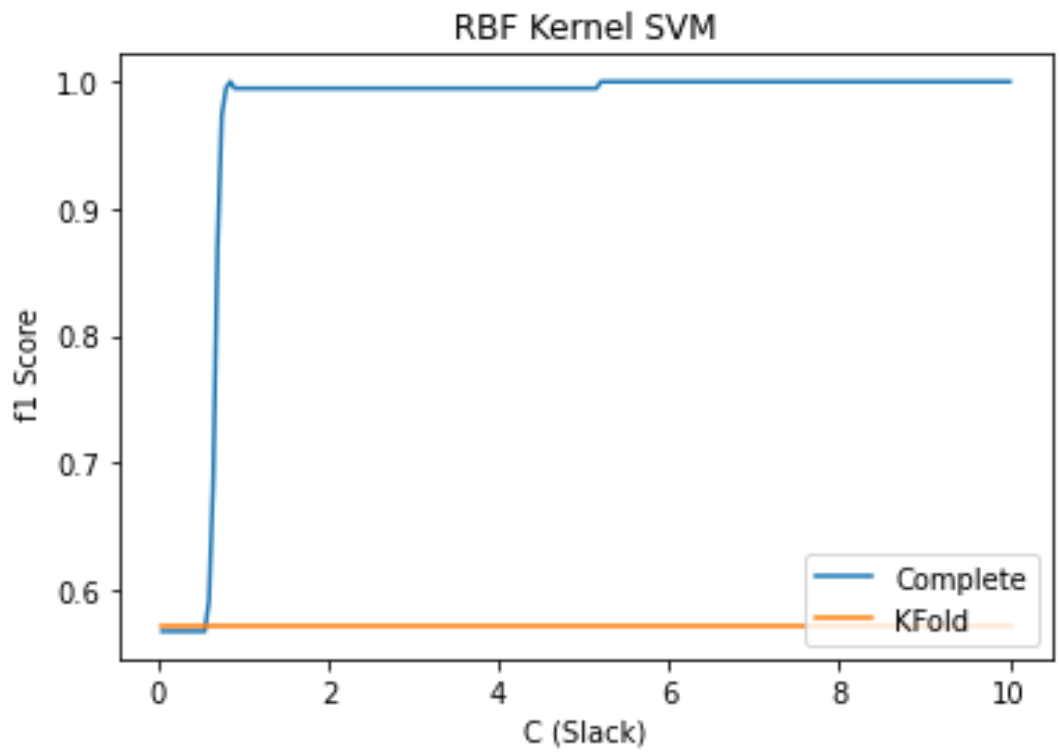
Kernel SVM with Changing C:

Like the first question, I used the similar approach.
I ran the loop for slack from 0 to 10 with an increment of 0.05.

The maximum accuracy obtained was

```
Accuracy on entire training Data : 1.0
Accuracy on  KFold Data : 0.6928876244665718
```

Here is the plot of F1 scores and C values:



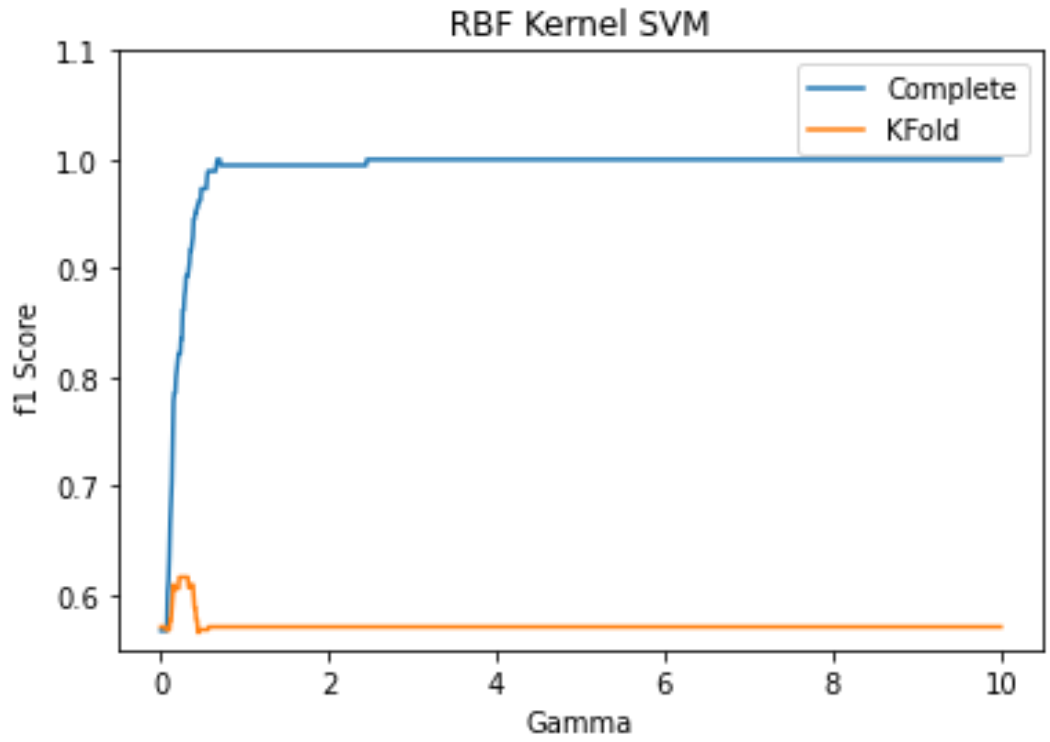
RBF Kernel SVM with Changing Gamma:

I ran the gamma values from 0.01 to 10.00 using an increment of 0.01

The maximum accuracy obtained was

```
Accuracy on entire training Data : 1.0
Accuracy on KFold Data : 0.7036984352773826
```

Here is the plot of F1 scores and C values:

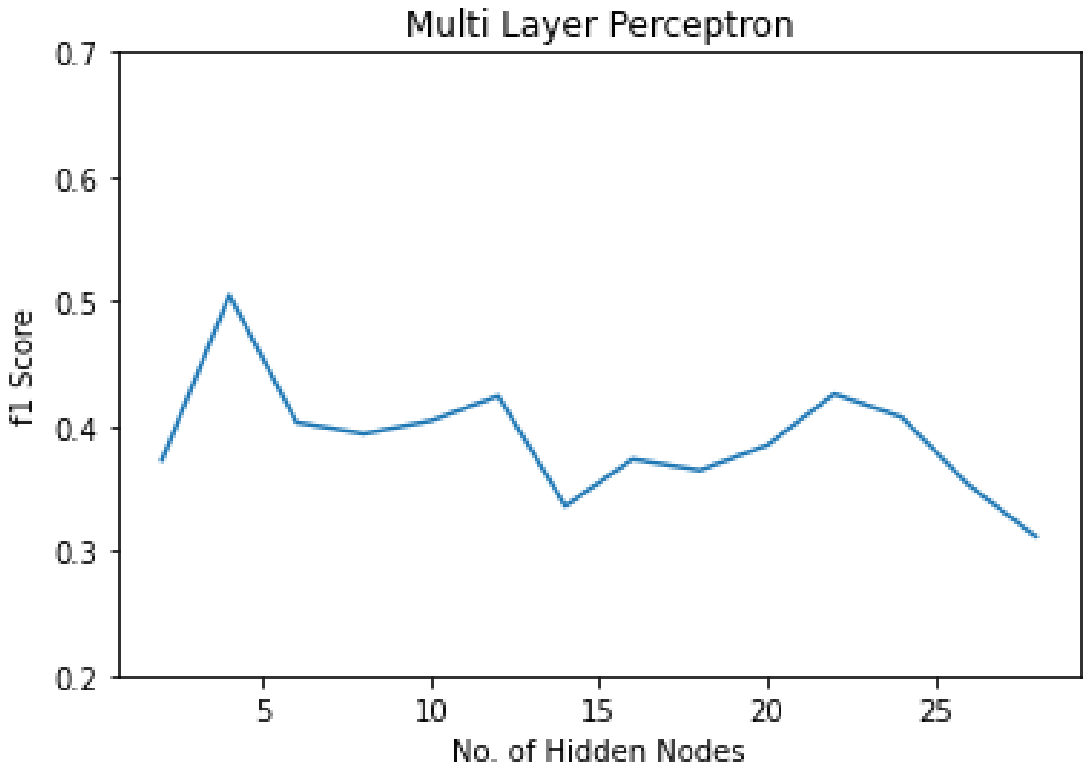


4. Observations:

- I wasn't able to reach perfect memorization on Linear SVM.
- I could reach perfect memorization on RBF kernel SVM.
- F1 score for KFold peaked to a value greater than 0.6 when I changed the gamma value but the F1 for Kfold remained the same for the entire time when I changed C

Stage 2

- Best performance for 10 hidden units is 1.0 for scaled data.
- Best performance achieved for unscaled data was around 0.7 with the hidden units as 25. Due to hardware limitations, I couldn't test with higher number of units.
- I have tried the number of hidden units from 2 to 30 and got an f1 score averaging around 0.4 for k fold data where k=5.



Rbf kernel had the best accuracy of 70 percent for Kfold validation.

```
Accuracy on entire training Data : 1.0
Accuracy on KFold Data : 0.7036984352773826
```

How to run the code :

Import the ipynb file and restart and run the kernel. This gives the graphs and accuracies for corresponding cells.

Note : The training time for the last part where we sweep the number of hidden units was around 3-4 hours.