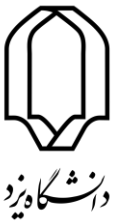
In the name of Allah



Computer Engineering Faculty of Yazd University
Artificial Intelligence

Machine Learning Course

Homework 3

(Support Vector Machine)

Instructor: Dr. Yazdian

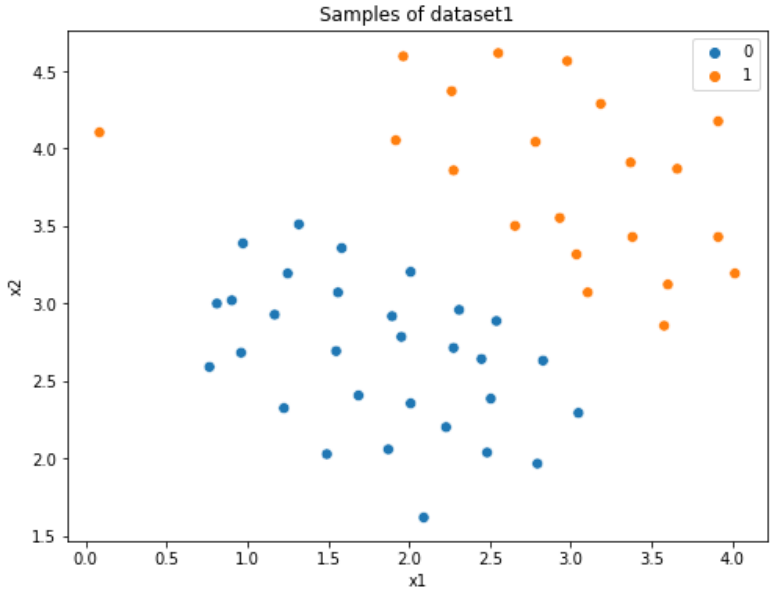
Shiva Zeymaran

Student ID: 40109434

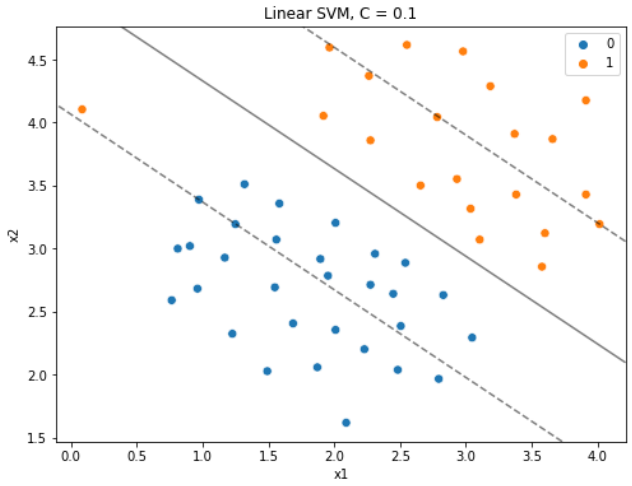
Fall 2022

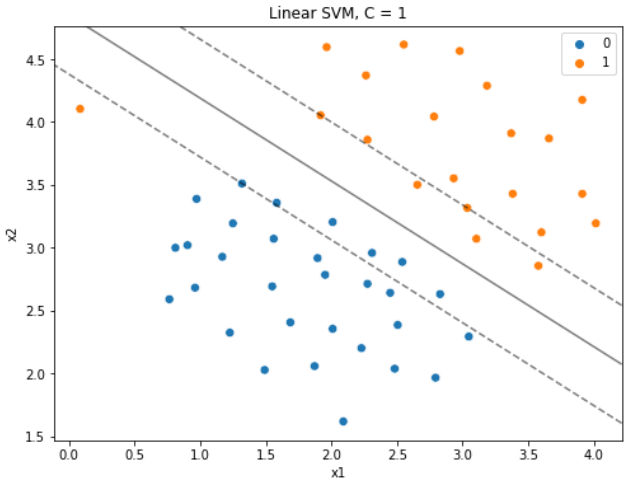
Part A – Linear SVM for Two-class Problem

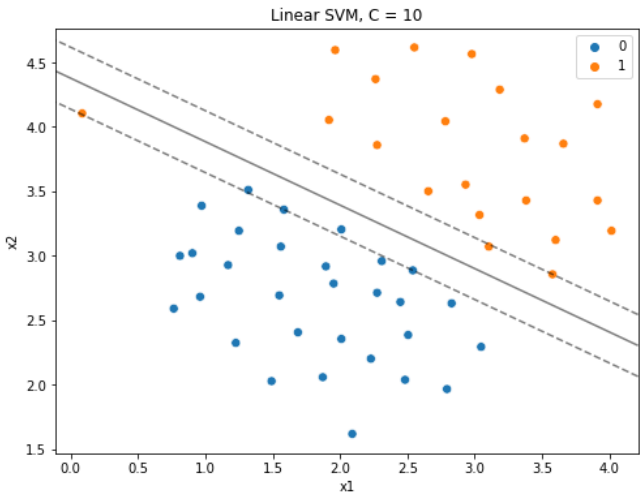
1. Dataset visualization

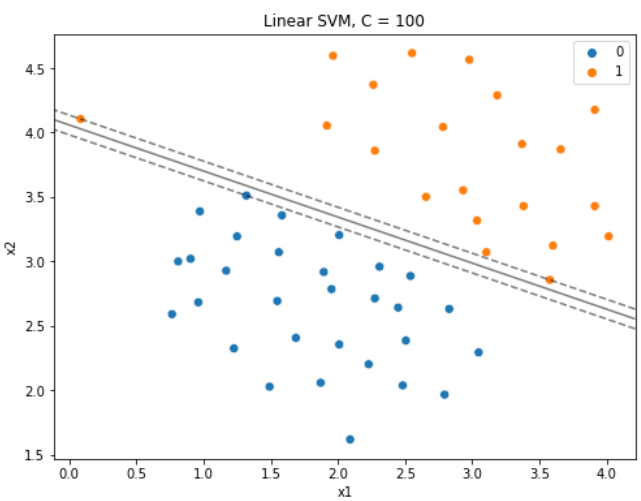


2. Plot decision boundaries of linear SVM models using different C values

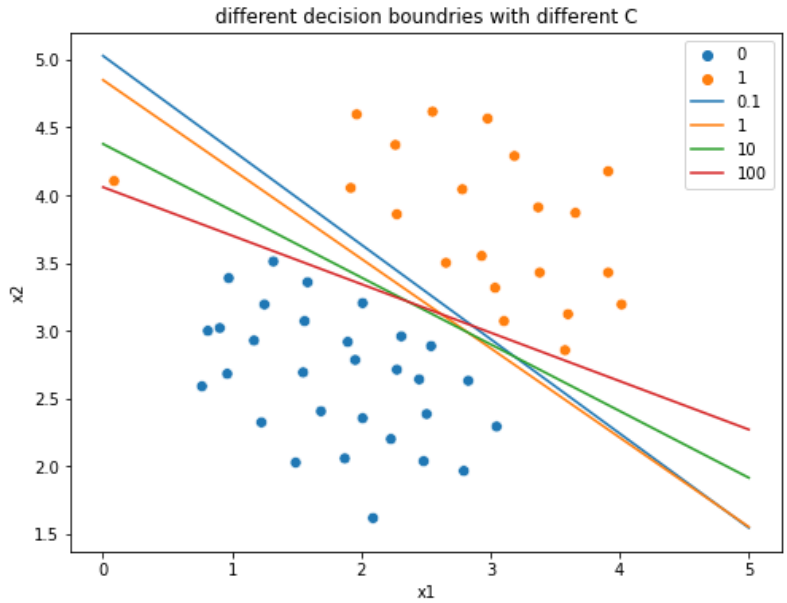






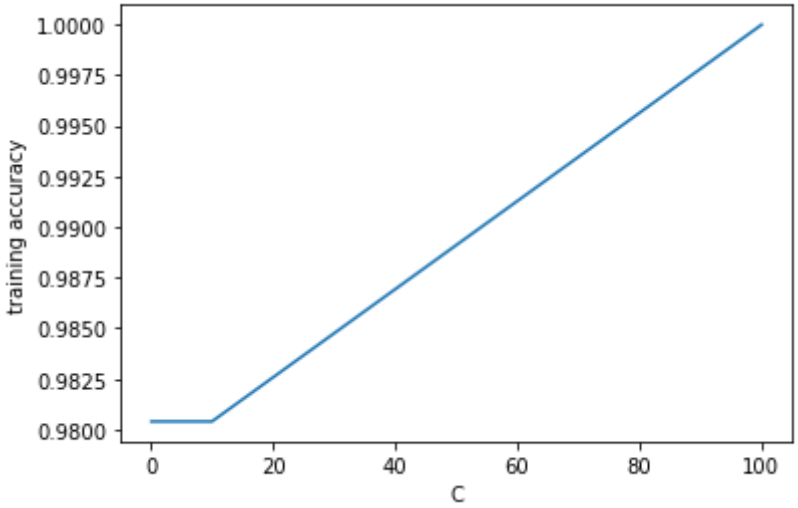


3. Plot different decision boundaries with different C on one figure



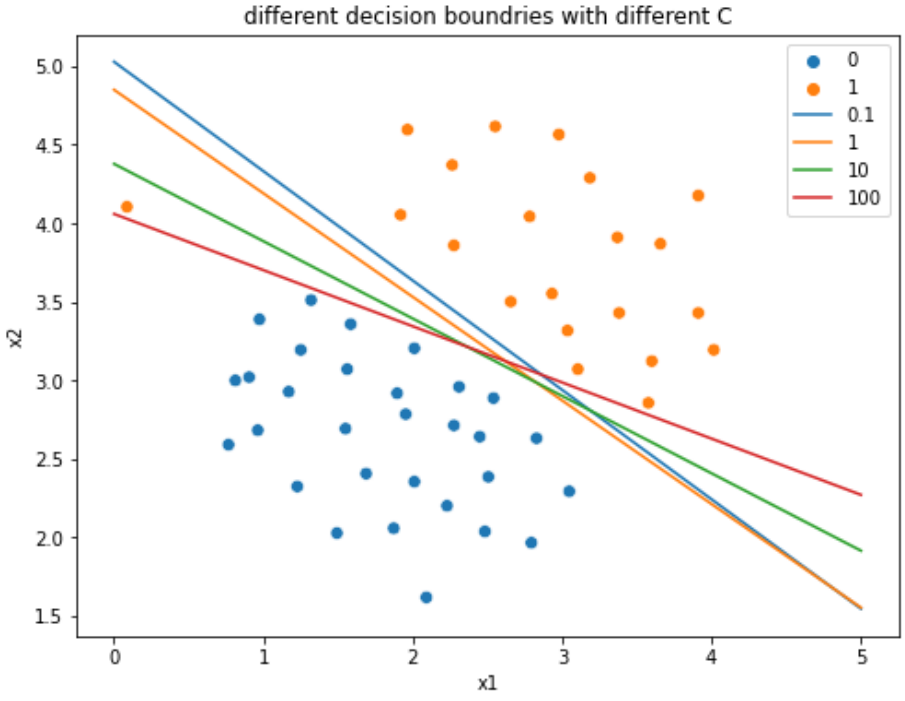
Increasing the value of C causes the SVM to classify more accurate and the large value of C tells the SVM to try to classify all the samples correctly.

4. Plot training accuracy in terms of C



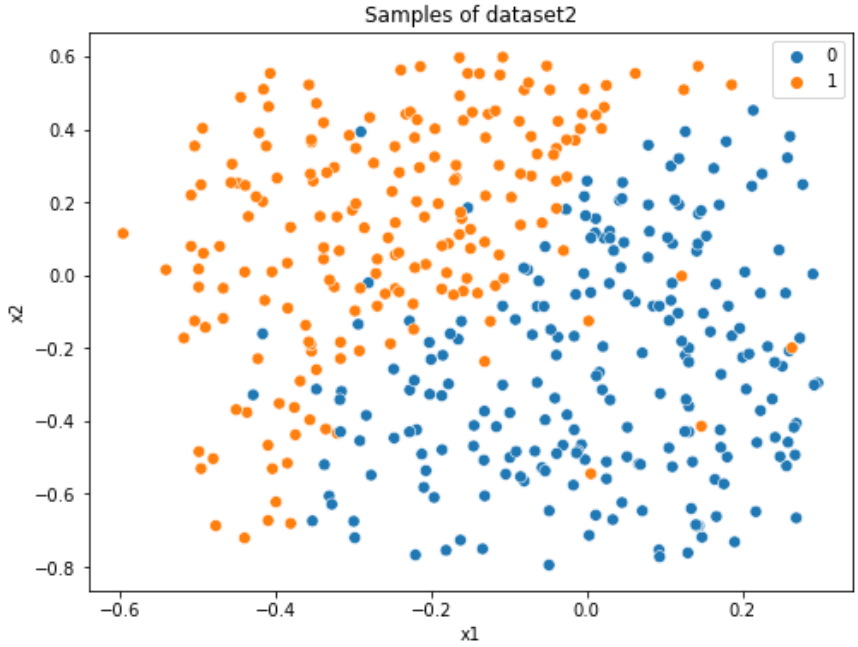
5. Training accuracy of the SVM models are equal in 3 first cases (C = {0.1, 1, 10}) and it is equal to 0.98; but this value is equal to 1.0 when C = 100 which means the classifier is over-fitted on train data and it can classify all training samples truly. Therefore, by increasing the value of C parameter, the training accuracy increases and the model will be over-fitted.
   In the other hand, increasing the value of C will increase the complexity of the model. As a result, the best value for C is 0.1 in this case; because its corresponding model has high training accuracy and low complexity. Increasing the C parameter after this value, has no effect on accuracy (before over-fitting) but increases the complexity.

6. By increasing the value of C, the SVM model tries to classify all the examples correctly. As it is clear from the following plot, the model changes in each step trying to classify the outlier truly. Therefore, the movement of the decision boundary is because of the outlier point.

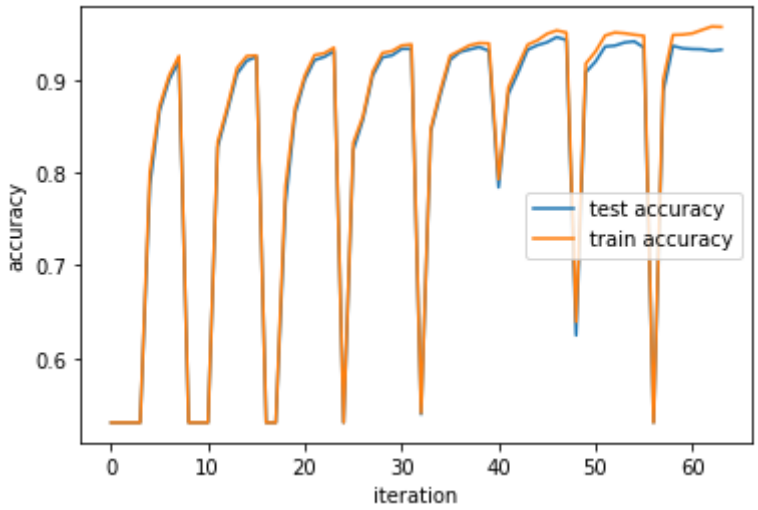


Part B - Kernel SVM for Two-class Problem

1. Dataset visualization



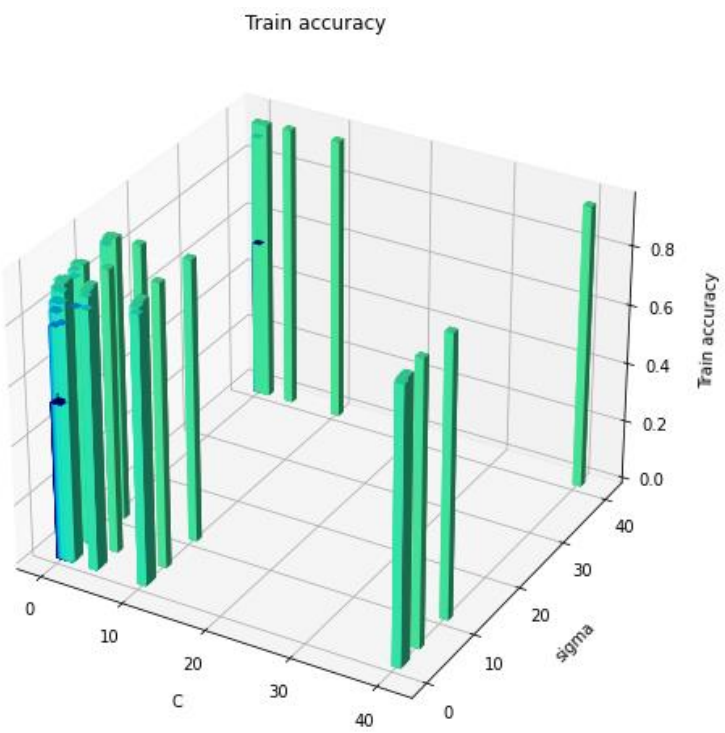
2. After training the SVM model on this dataset using 64 different pairs of sigma and C, we reached 64 values for test and train accuracies. We plot both of these values in one figure which shows the different values of accuracies in each of these 64 iterations:

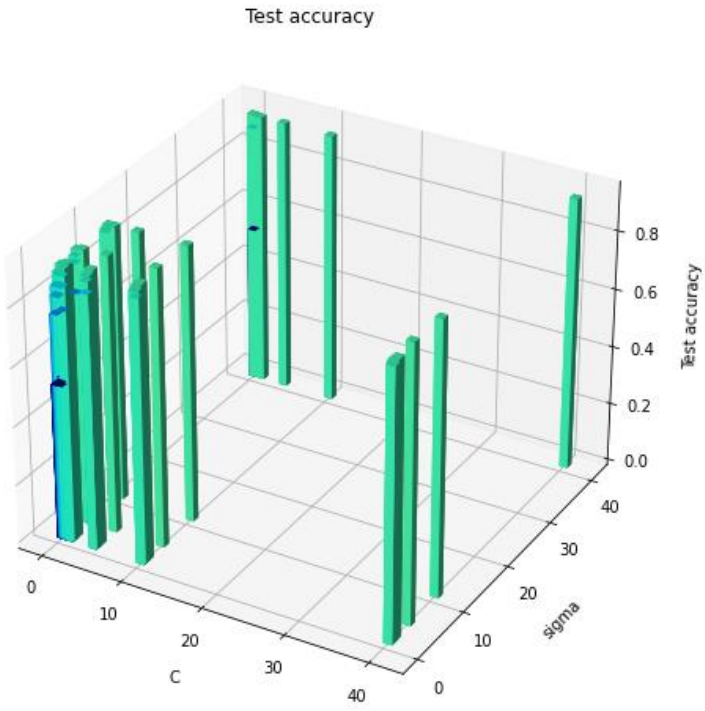


We can understand that the maximum test accuracy is in 47th iteration. In this iteration we also reached a good training accuracy. After this peak, in other peaks we also have high training accuracy but the test accuracy

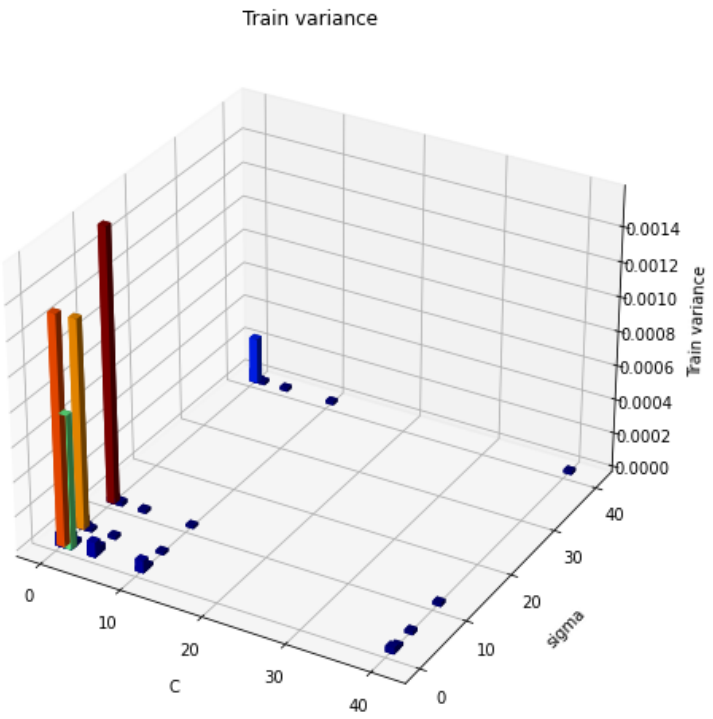
decreases which means the over-fitting occurred. Therefore, the best model is when we choose C = 10 and sigma = 4.

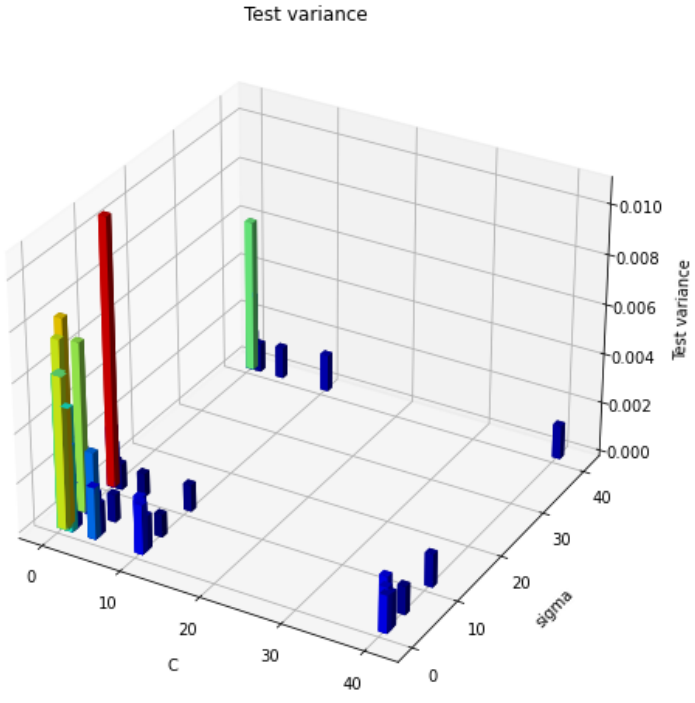
3. Plot train and test accuracies for different values of C and sigma



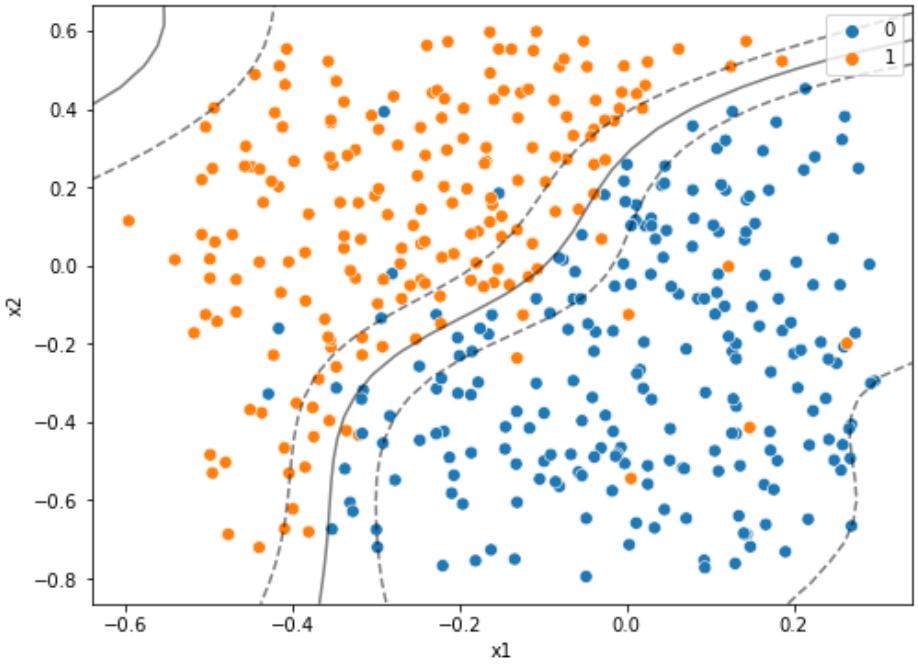


Plot the train and test corresponding variances





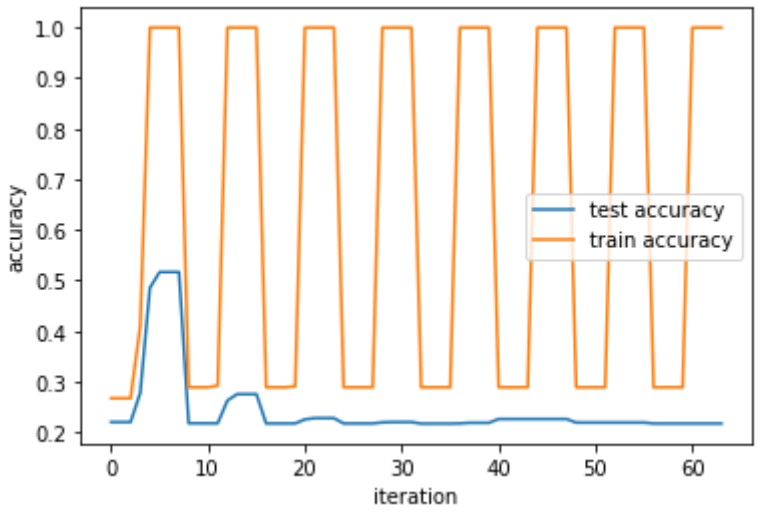
4. Plot the data and the decision boundary for the best C and sigma (which we discussed it is 10 for C and 4 for sigma)



5. The best selected model as we shown in Q2 is when we have C = 10 and sigma = 4. In this case, which happened in 47th iteration (case), the test accuracy was 0.94.
   This model has the highest possible test accuracy among the examined cases. After that, by increasing the value of C or sigma the train accuracy increases but the test accuracy decreased. Therefore, the higher values for C and sigma, leads us to over-fitting and having more complex model.

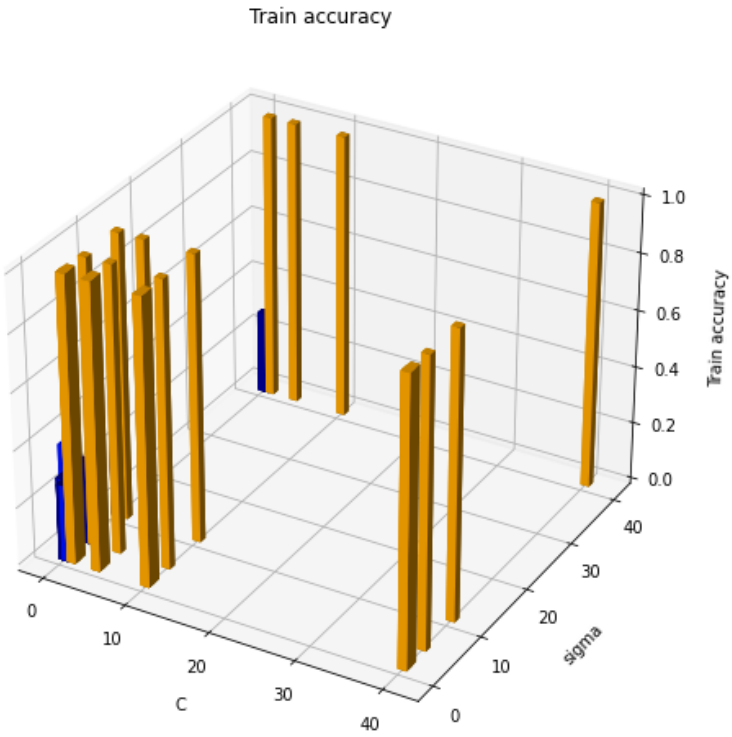
Part C - Kernel SVM for Multi-class Problem

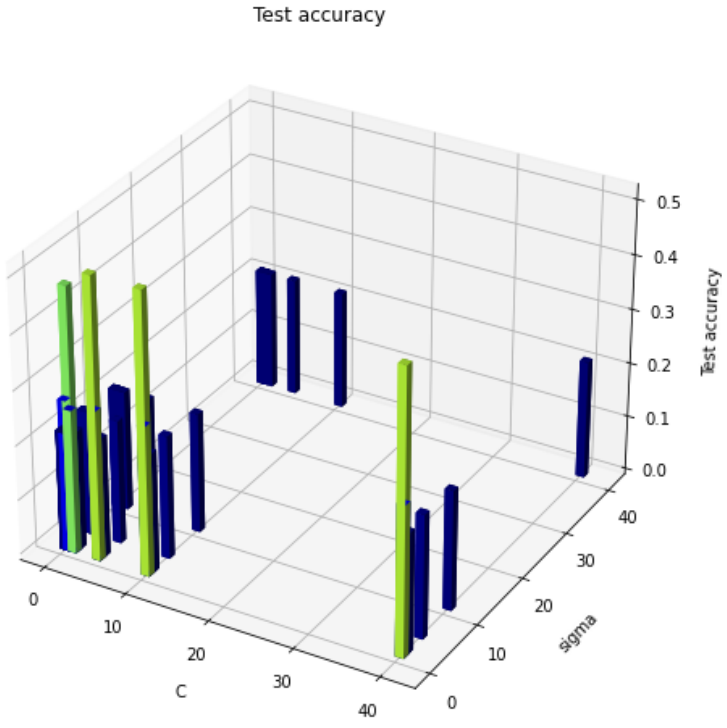
1. After training models with different values for C and sigma, we have 64 values of test and train accuracies which are as follows:



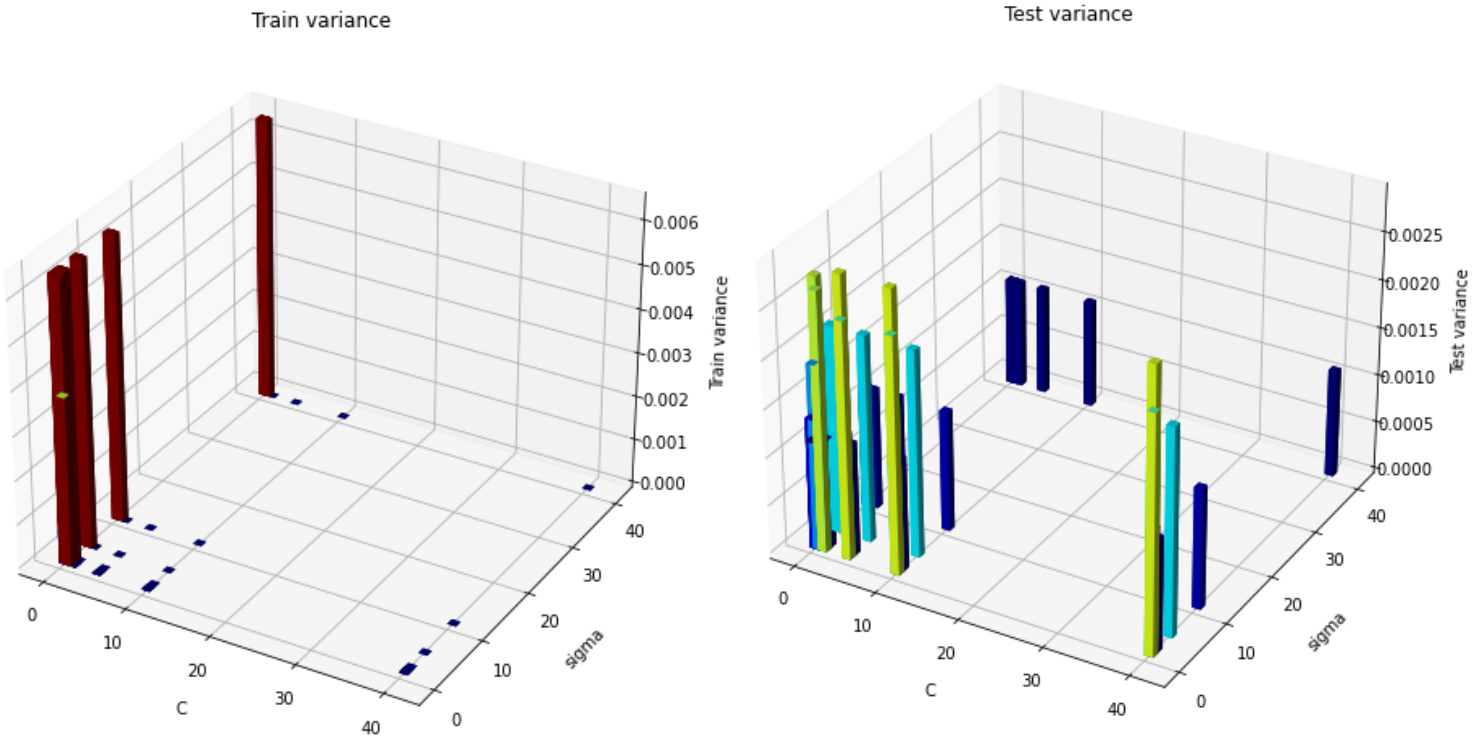
We understand that in 6th iteration, we have the maximum test accuracy and very good train accuracy. After that, the test accuracy decreases with high training accuracy; so we have over-fitting.
In 6th iteration the C is 4 and sigma is 0.01 which are the best values for these two parameters.

Plot the train and test accuracies for different values of C and sigma

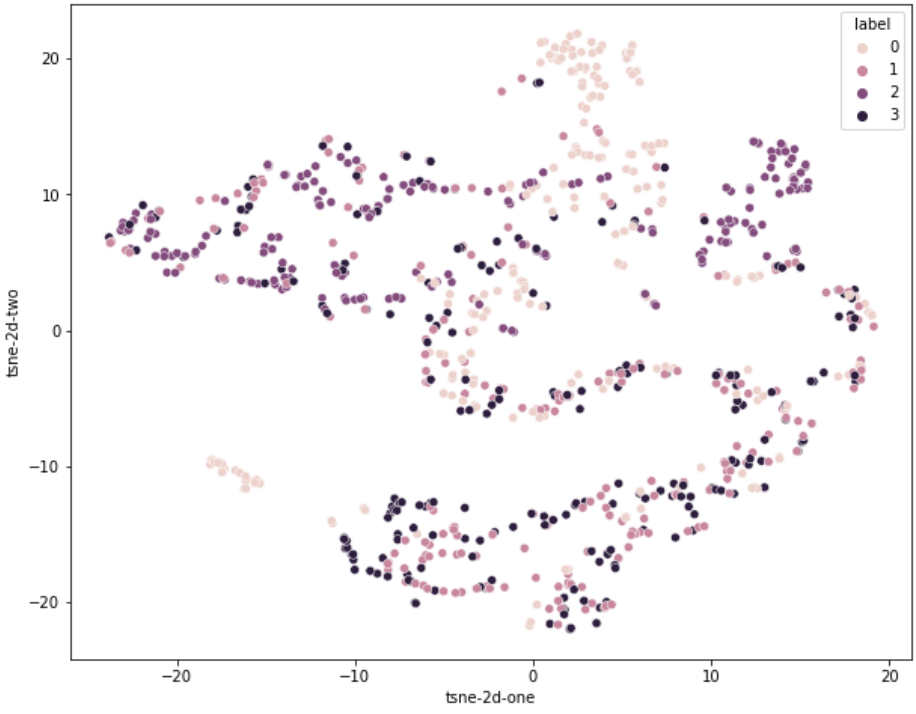




Plot train and test corresponding variances



2. As we discussed in last part, the best values for C and sigma are: C = 4, sigma = 0.01.
Plot the multi-dimensional data in 2D space using TSNE



3. The best selected model as we shown in Q1 is when we have C = 4 and sigma = 0.01. In this case, which happened in 6^{th} iteration (case), the test accuracy was 0.52.
   This model has the highest possible test accuracy among the examined cases. After that, by increasing the value of C or sigma the train accuracy remains high but the test accuracy decreased. Therefore, the higher values for C and sigma, leads us to over-fitting and having more complex model.

Part E – Conclusion

- When we have two-class problem which the data can be classified linearly, we should use linear SVM. This model only needs one parameter to be set which is C. As we increase the value of C, the SVM classifies training data more accurate. We should be aware of over-fitting and complexity of model to choose the best value for C.
- When our given dataset is a two-class problem but it can not be classified linearly, we should use kernel SVM. In this case we have two parameters of C and sigma. We can use cross validation to find the best values for these two parameters.
- For multi-class problem, we should use kernel SVM, too (because it is non-linear). Also, we need to use one the approaches of OVO or OVA. Here again we should find the best values for parameters C and sigma that can be done using cross validation.
- It seems that in two-class problem, C and sigma can be larger than when we have multi-class problem; because by increasing the value of these two parameters, the complexity of the model increases much faster in multi-class problem.
- Running kernel SVM for multi-class problem takes more time than using it for two-class problem.