SPR HW 4 - Report

Mohammad Ahmadi 9531202
Amir Hossein Ansari 9531888

Contents

1 Part A: Model Selection

1.1 In this part you should implement one of the famous approaches of cross validation: Ten-Time-Ten-Fold.

No report.

1.2 Explain how you choose the best parameters using cross validation?

By cross validation different test samples are randomly selected from train dataset and by that we can train and test our algorithm with using only train dataset. So we can test accuracy of algorithm using different parameters or configurations and select the best one, and because test datas are selected randomly we can trust the result.

2 Part B: Generative Classification

2.1 GMM Classifier

2.1.1 a- Construct a GMM classifier with K Gaussian components.

No report.

2.1.2 b- Use ten-time-ten-fold cross validation to determine the best K. You should test the values of K = 1; 3; 5; 7.

No report.

2.1.3 c- Plot the accuracies and also the variances of ten-time-ten-fold cross validation for different values of K and explain how you determine the best K?

Here we have different values for C and Sigma. All of the combinations of these two parameters are tested using ten time ten folds and the parameter with the best mean accuracy is selected.

Shapes for heart dataset:

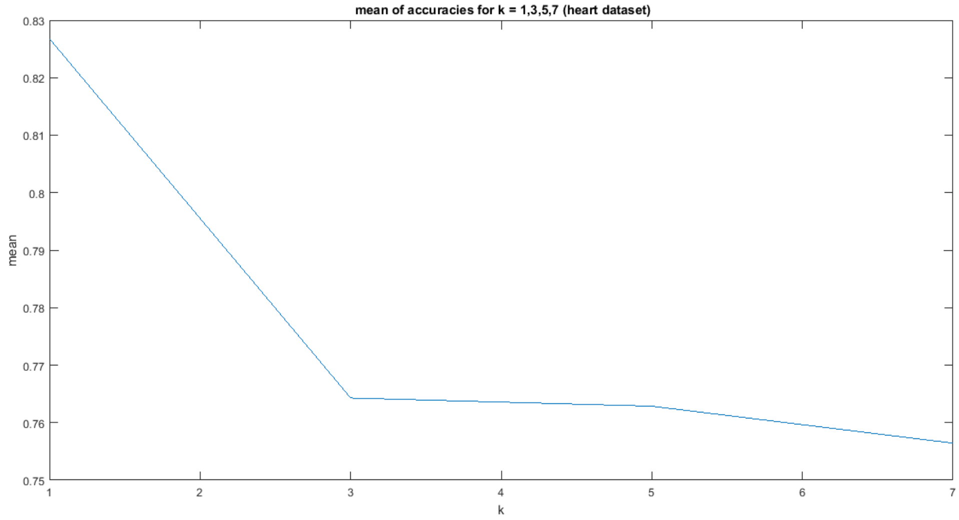

Figure 1: Mean of accuracy using tentimes ten fold for heart dataset

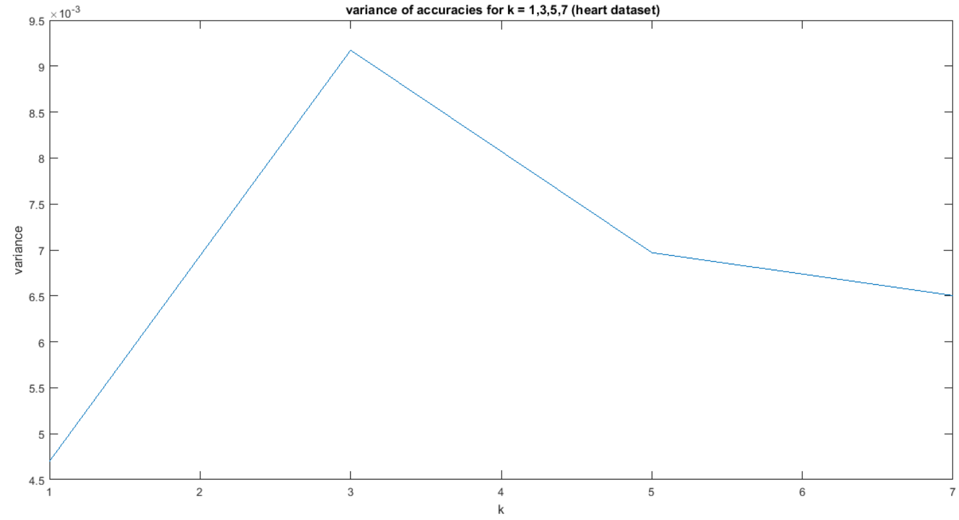

Figure 2: Variance of accuracy using tentimes ten fold for heart dataset

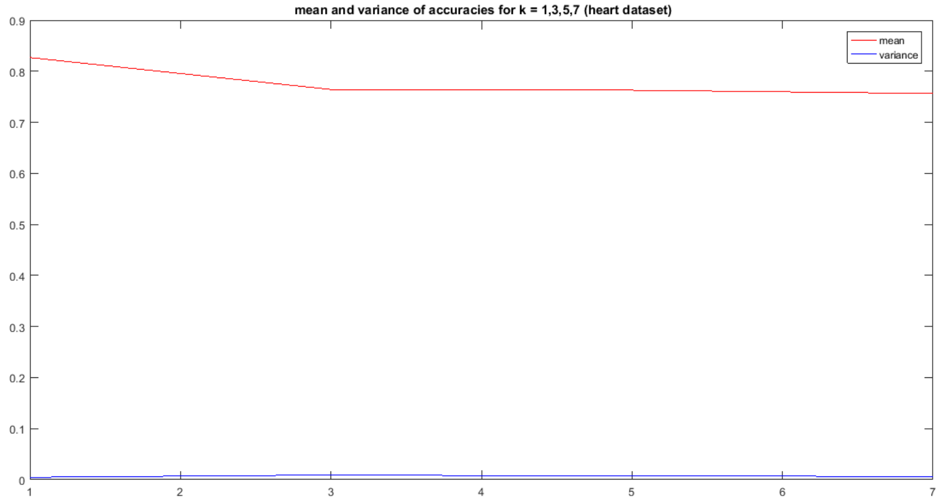

Figure 3: Mean and variance of accuracy using tentimes ten fold for heart dataset

Shapes for vehicle(xaa) dataset:

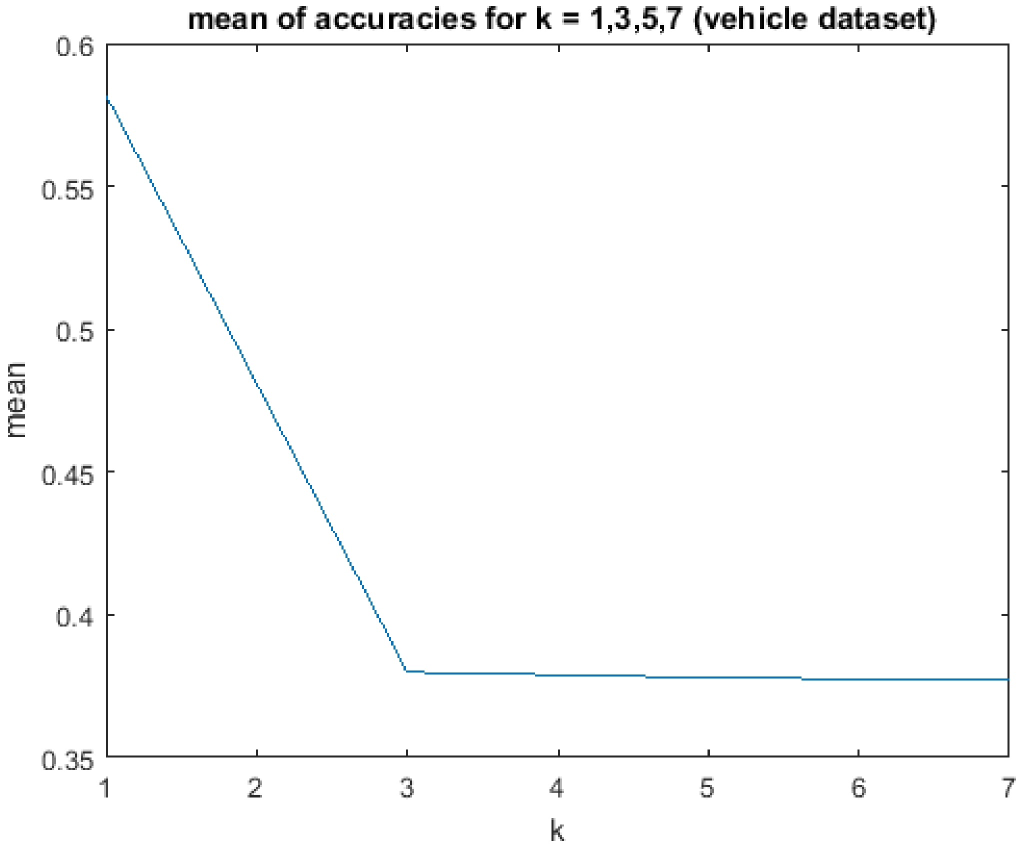

Figure 4: Mean of accuracy using tentimes ten fold for vehicle dataset

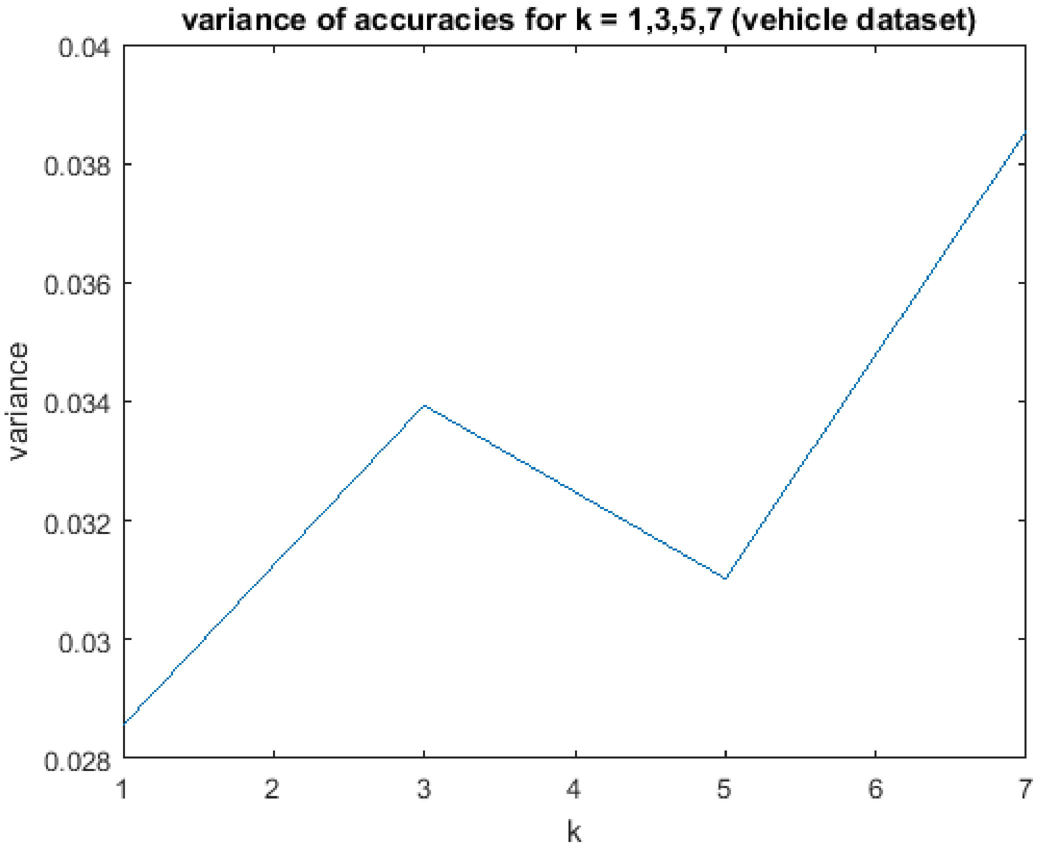

Figure 5: Variance of accuracy using tentimes ten fold for vehicle dataset

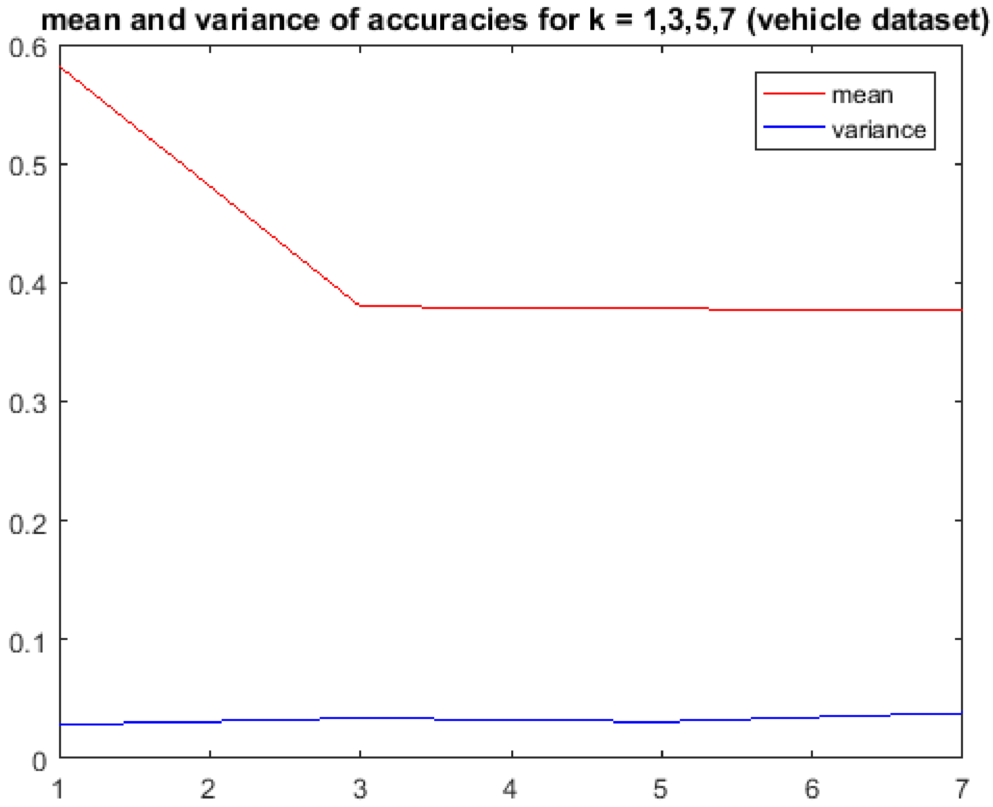

Figure 6: Mean and variance of accuracy using tentimes ten fold for vehicle dataset

2.1.4 d- Report the test accuracy using the selected model (best K).

For heart dataset best accuracy was 0.8268 (82.68%) for K = 1 .
For vehicle dataset best accuracy was 0.5820 (58.20%) for K = 1 .

3 Part C: Discriminative Classification

3.1 Linear SVM

3.1.1 (a) Train the SVM using two different values of the penalty parameter, i.e., C=1 and C=100.

No report.

3.1.2 (b) Plot the data and the decision boundary

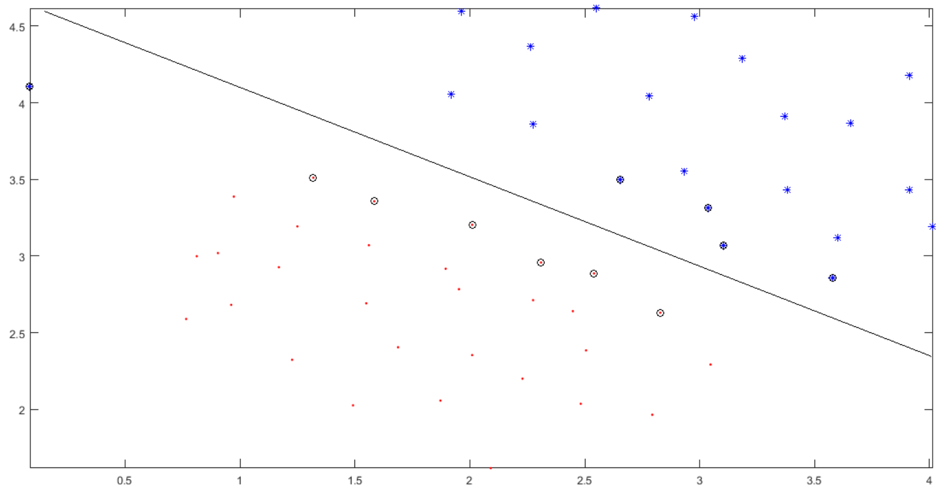

Figure 7: data and decision boundary for c = 1

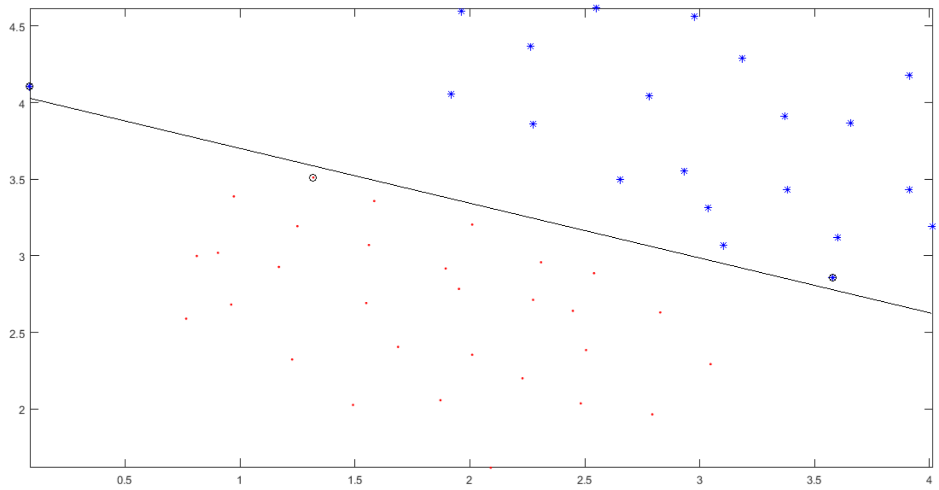

Figure 8: data and decision boundary for c = 100

3.1.3 (c) Report the train accuracy for both C=1 and C=100.

For c = 1 accuracy is 98.0392% .
For c = 100 accuracy is 100% .

3.2 Kernel SVM for two-class problem

3.2.1 (a) Train SVM with the penalty parameter C and the standard deviation for RBF kernel. Determine the best value C by ten-time-ten-fold cross validation.

3.2.2 (b) Plot train and test accuracies and their corresponding variances of ten-time-tenfold cross validation for different values of C and

Heart dataset:

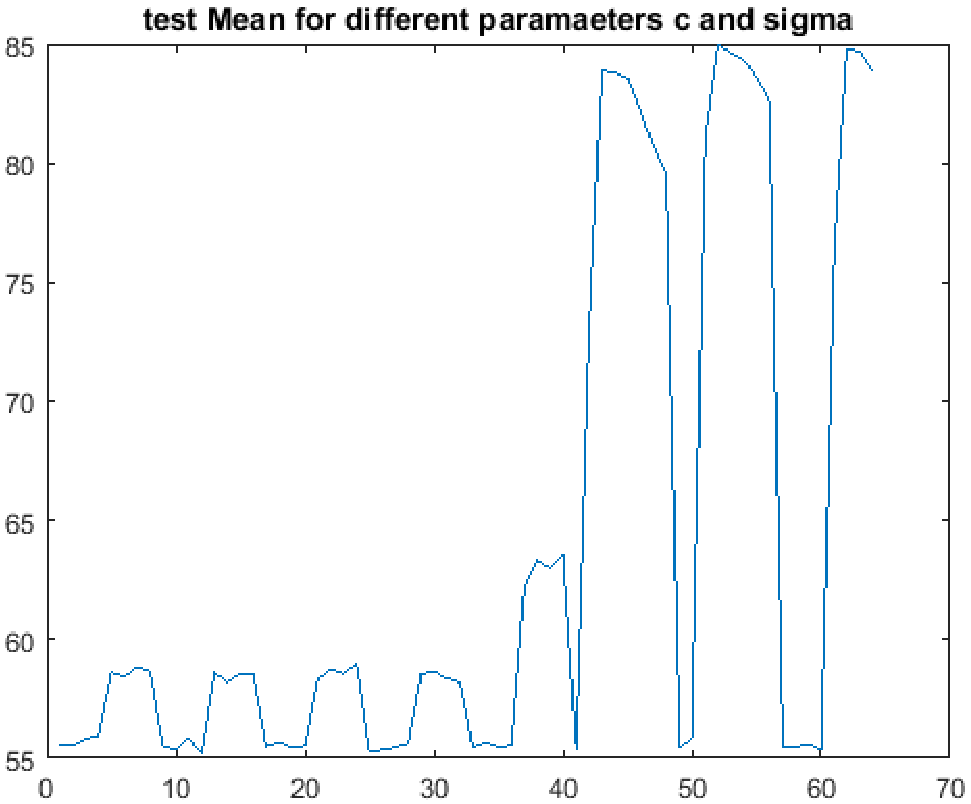

Figure 9: mean of accuracy for test data (heart dataset)

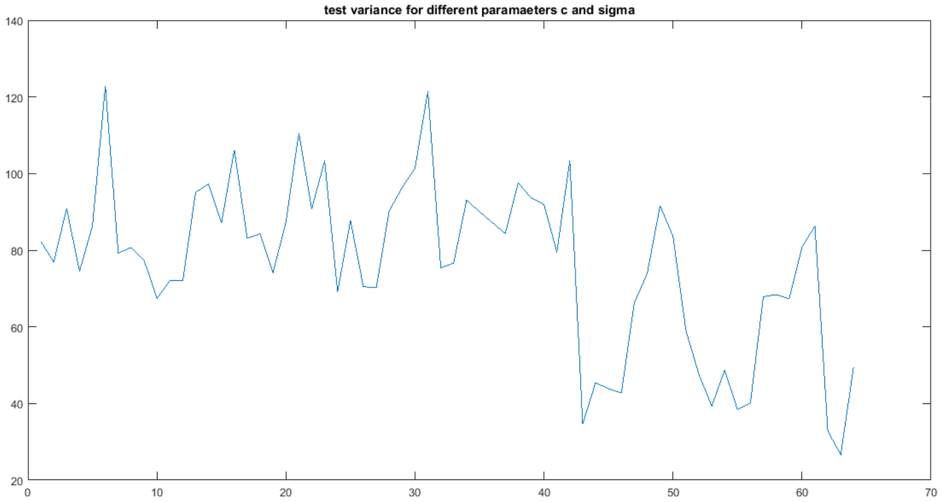

Figure 10: variance of accuracy for test data (heart dataset)

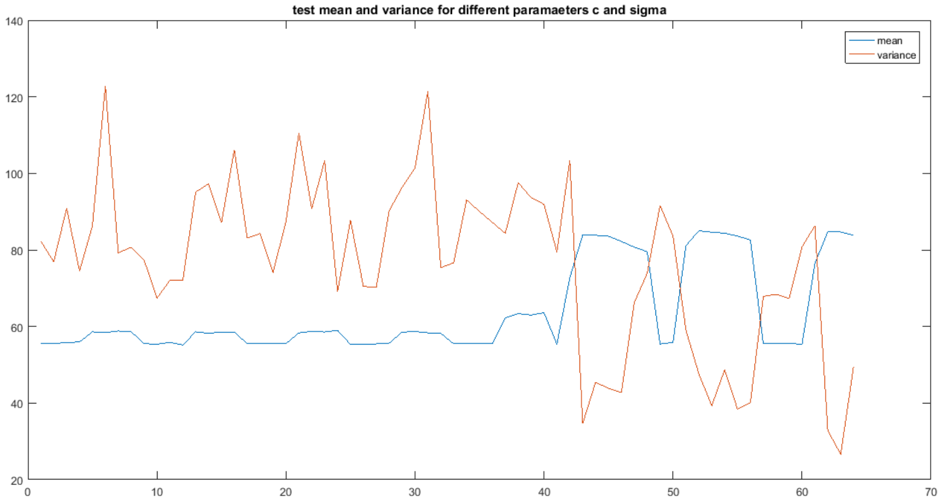

Figure 11: mean and variance of accuracy for test data (heart dataset)

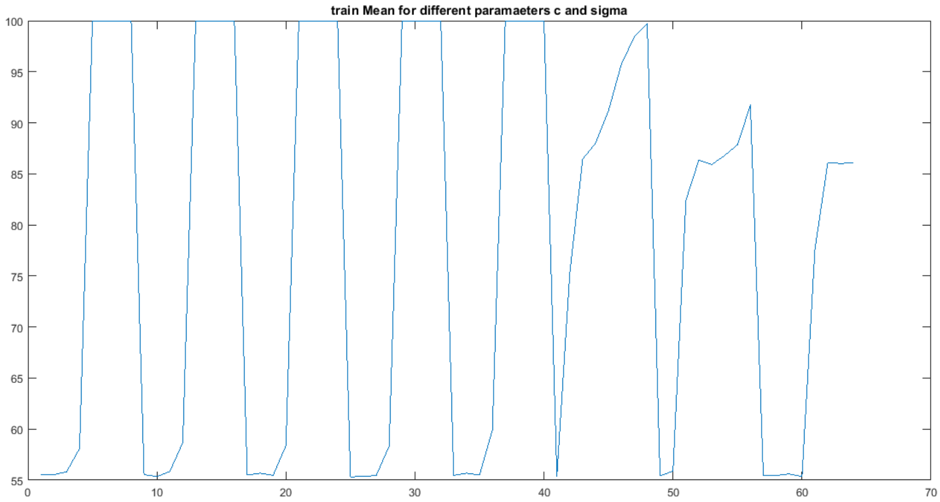

Figure 12: mean of accuracy for train data (heart dataset)

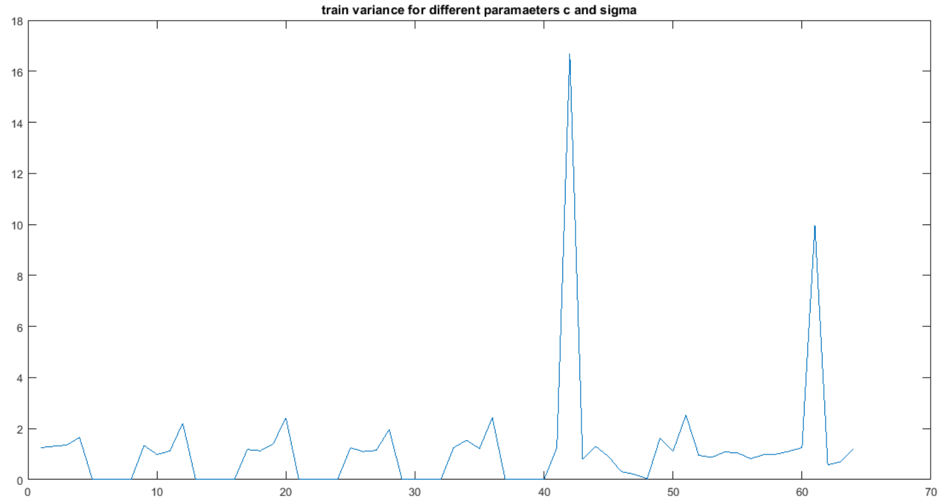

Figure 13: variance of accuracy for train data (heart dataset)

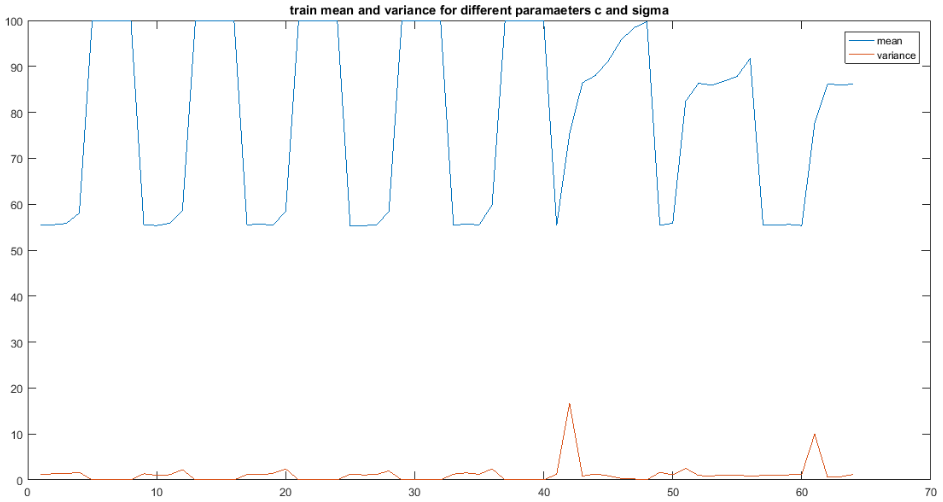

Figure 14: mean and variance of accuracy for train data (heart dataset)

hwDataset2:

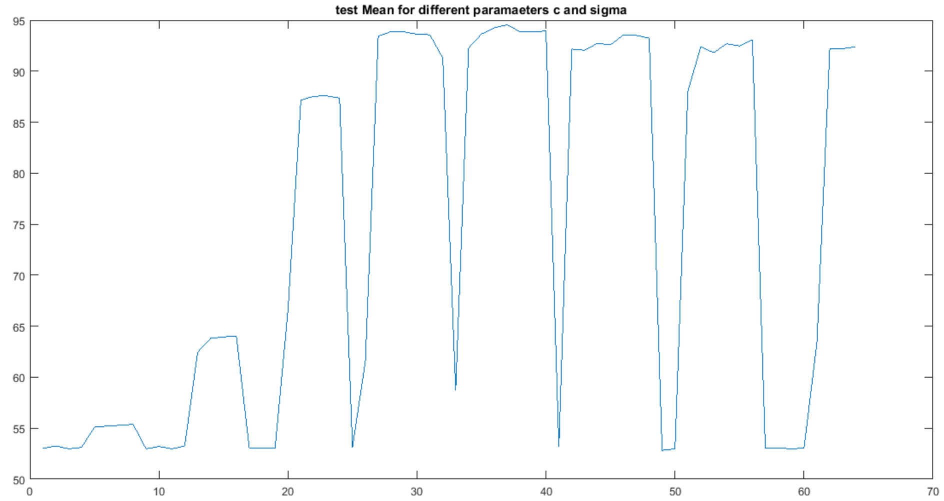

Figure 15: mean of accuracy for test data (hwDataset2 dataset)

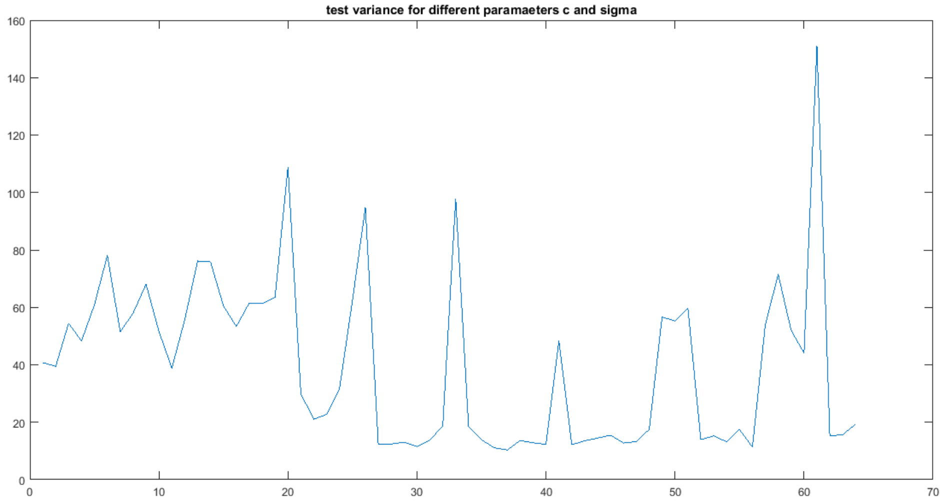

Figure 16: variance of accuracy for test data (hwDataset2 dataset)

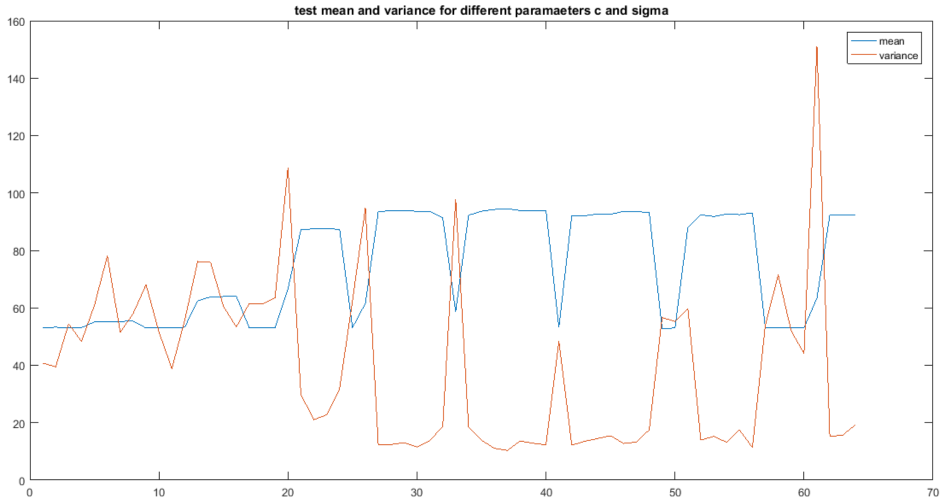

Figure 17: mean and variance of accuracy for test data (hwDataset2 dataset)

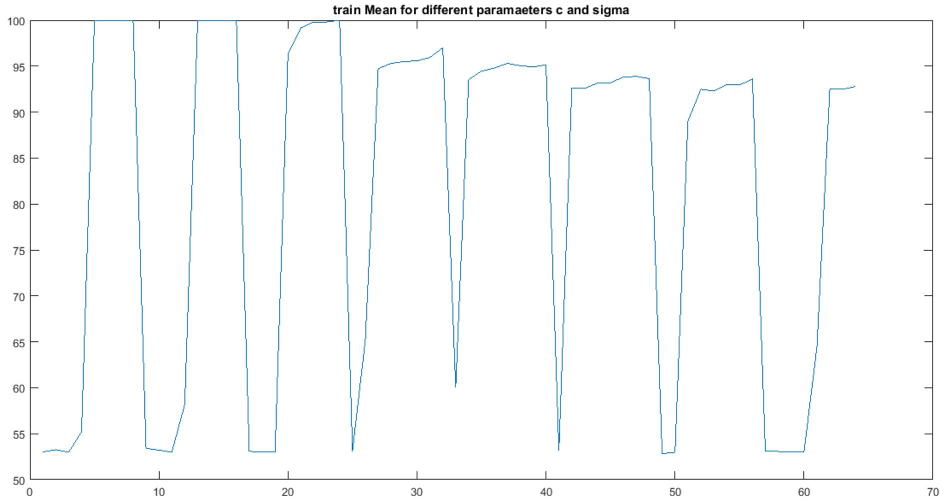

Figure 18: mean of accuracy for train data (hwDataset2 dataset)

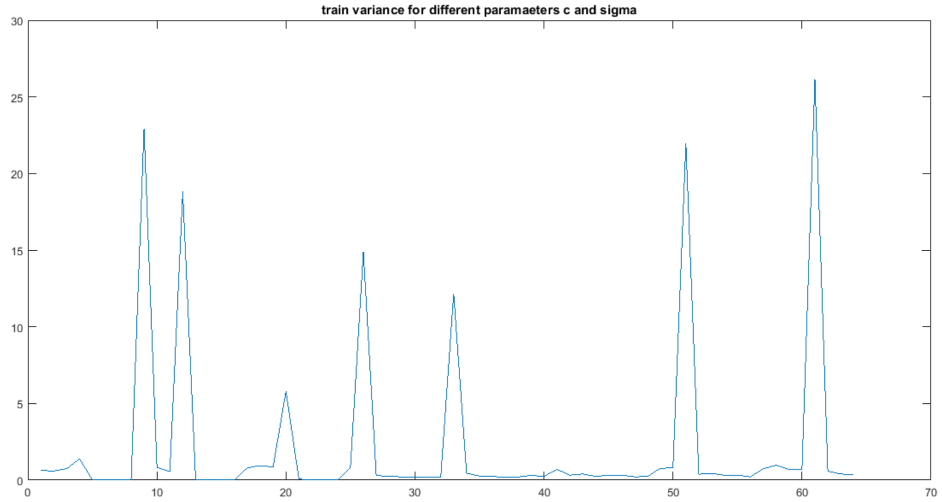

Figure 19: variance of accuracy for train data (hwDataset2 dataset)

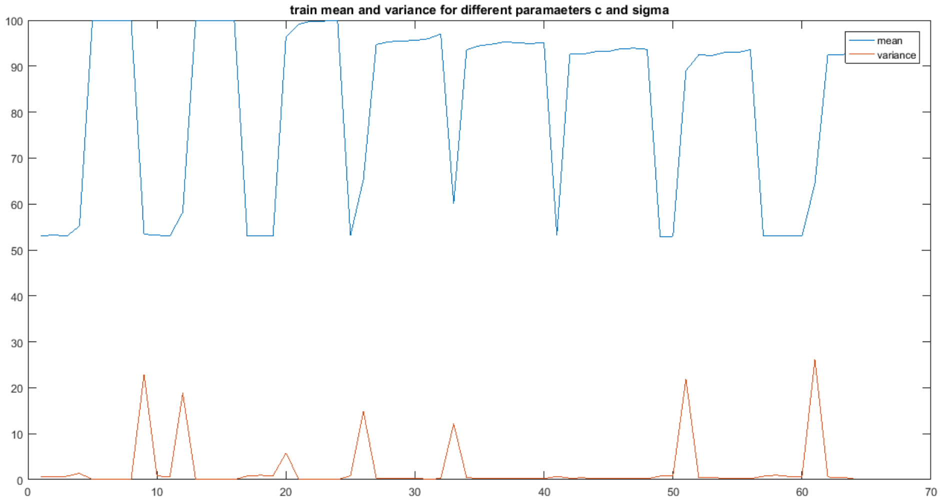

Figure 20: mean and variance of accuracy for train data (hwDataset2 dataset)

3.2.3 (c) Plot the data and the decision boundary for hwDdataset2 (for best model)

Best is C = 1 and Sigma = 1

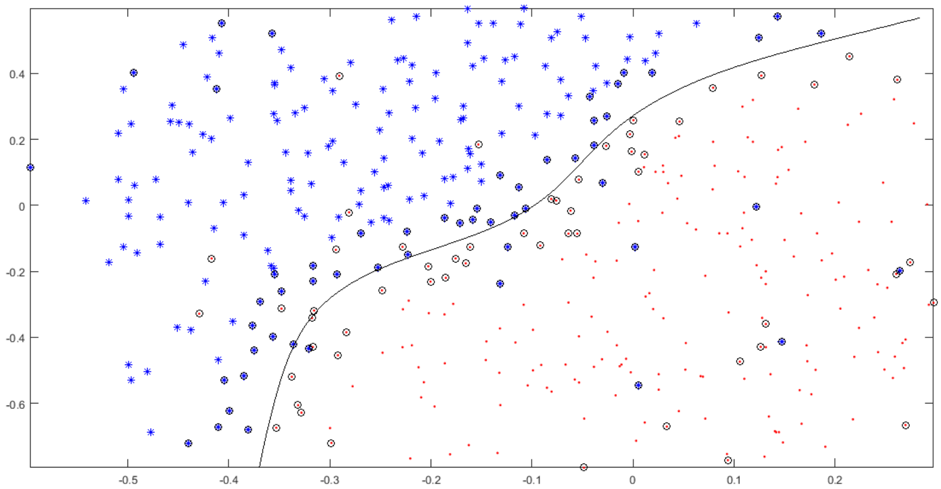

Figure 21: data and decision boundary and support vectors

3.2.4 (d) Report the test accuracy using the selected model (best C and sigma)

hwDataset2 : C = 1 Sigma = 1 , best test accuracy = 94.5476 .
heart : C = 0.4 Sigma = 10 , best test accuracy = 85.0000 .

3.3 Kernel SVM for multi-class problem

3.3.1 (a) Design a multi-class SVM classifier with one-against-all method

No report.

3.3.2 (b) Determine the best value of C by cross validation

C = 40 , sigma = 10

3.3.3 (c) Plot the train and test accuracies and their corresponding variances of ten-time-tenfold cross validation for different value of C and sigma

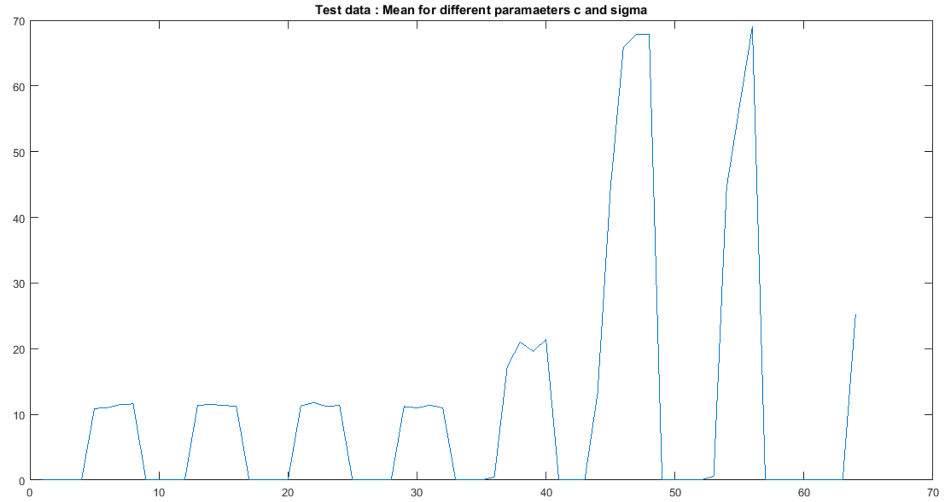

Figure 22: mean of accuracy for test data (vehicle dataset)

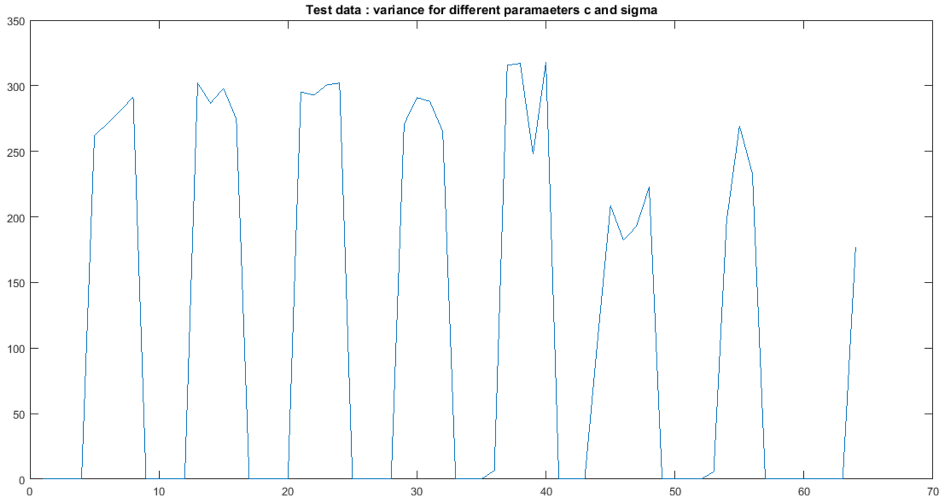

Figure 23: variance of accuracy for test data (vehicle dataset)

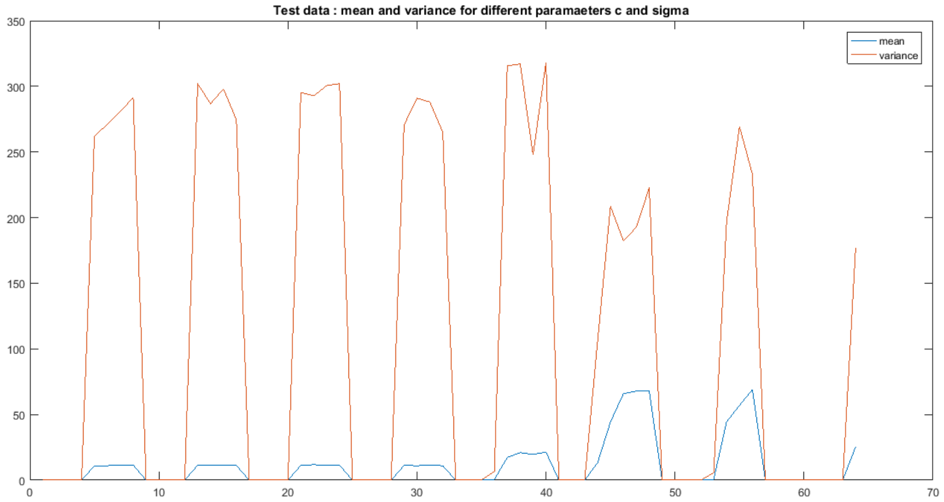

Figure 24: mean and variance of accuracy for test data (vehicle dataset)

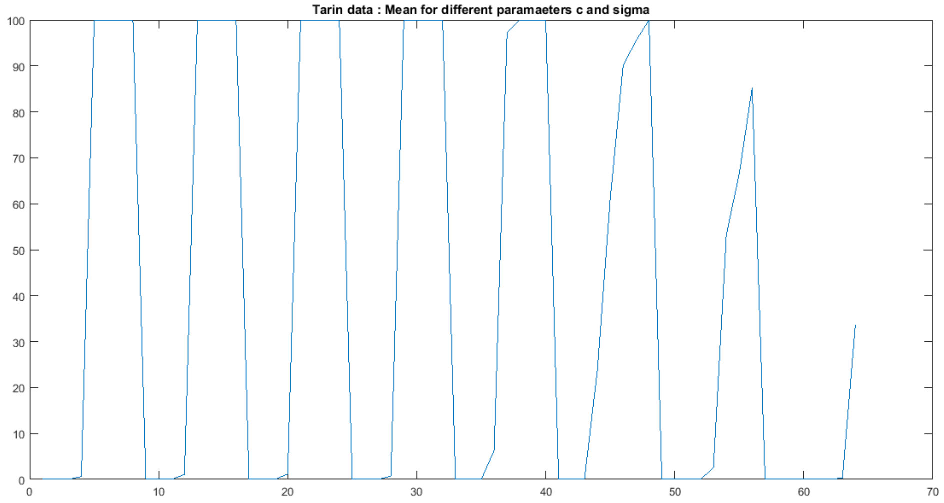

Figure 25: mean of accuracy for train data (vehicle dataset)

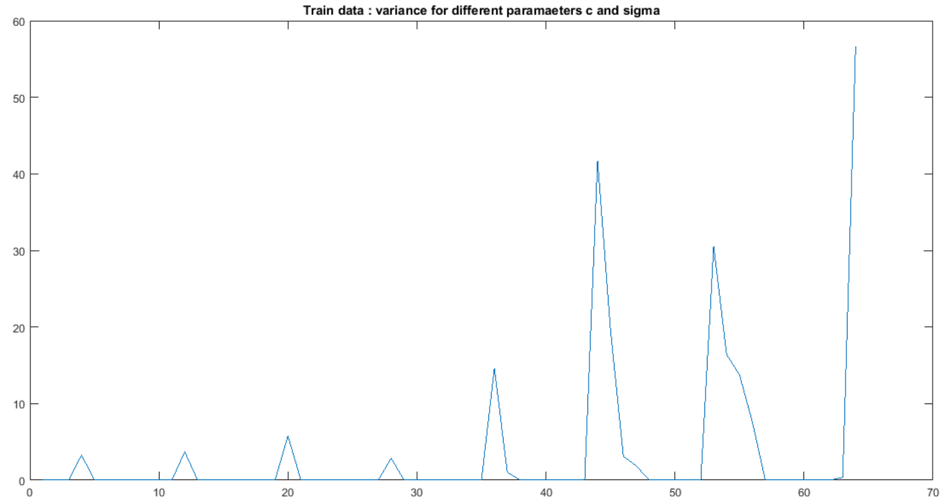

Figure 26: variance of accuracy for train data (vehicle dataset)

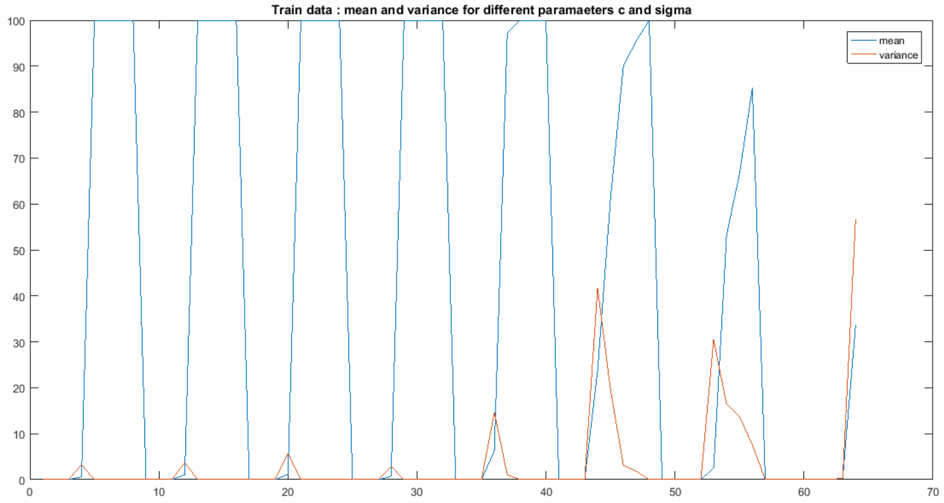

Figure 27: mean and variance of accuracy for train data (vehicle dataset)

3.3.4 (d) Report your test accuracy using the selected model.(best C and sigma)

Best C = 40, Sigma = 10.
best accuracy = 69 .

3.3.5 (e) Compare the results of part 2.I with 3.II & 3 according to the accuracies

For heart dataset:
GMM accuracy is 82.68% .
SVM accuracy is 85%.
* accuracy of svm is more than gmm.
* for gmm only one parameter needs to be tuned but for svm there are two parameters that must be tuned and it is easier to tune gmm's parameter than tuning svm's parameters. * Also if the data distribution is not a mixture of guassians(samples are not from guassian distribution) GMM will definitly fail but SVM is free from data distribution and can be more widely applied.

For vehicle dataset:
GMM classifier best accuracy is 58.20% .
SVM(one vs all) accuracy of 69% was achieved.
* accuracy of SVM is more than gmm.
* Because of small amount of samples compared to number of features we had problem with GMM but there was no problem in using SVM.
* Other reasons from above is true here too.

4 Part E: Extra

4.1 Many email services today provide spam filters that are able to classify emails into spam and non-spam email with high accuracy.

The best accuracy achieved is 98.8 %.
With parameters c = 51 and sigma = 351.

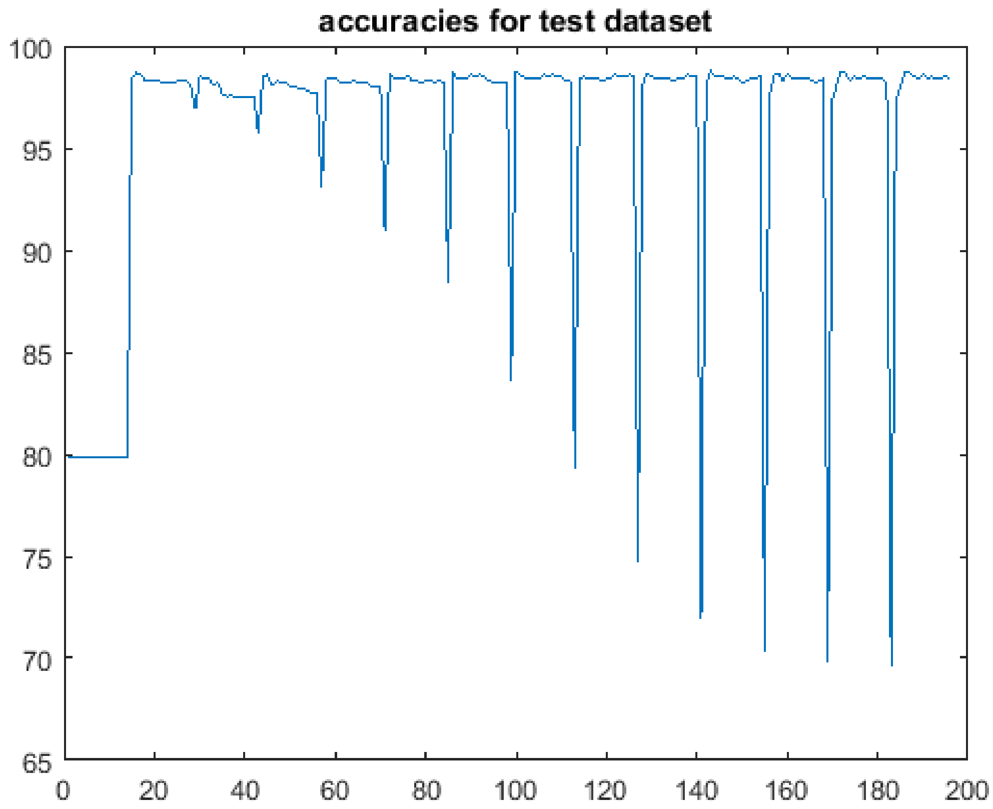

Figure 28: accuracy of test data for different values of c and sigma

4.2 Compare this approach to your implementation of assignment #2 spam classifier? What is your intuition and which one is better? why?

Using naive bayes the accuracy of 94.80% is achieved.
* SVM achieved better accuracy compared to naive bayes. * But in naive bayes there was no need to tune any parameter and because of that the overall cost time of naive bayes is less than SVM.
*Also we know that for naive bayes it's better to have a large number of samples.