CS464 Introduction to Machine Learning

Homework 3 Report

İrem Ecem Yelkanat
21702624
Section 2

1 PCA

In this question, support vector machine (SVM) classifier is trained to classify given samples as malignant or tumor. The given data is split into two sets as training set and test set, which training set contains the first 500 samples of the given data and test set contains the remaining samples. K-fold cross validation with k value 10 is used to tune hyper parameters C and gamma of SVM. SVM model is obtained from sklearn library. The folds that are going to be used in cross validations are generated beforehand the parts of the question.

Question 1.1

In this part of the question, linear SVM model without any kernel is trained with different C values including 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}. For each C value, SVM model is trained with 10-fold cross validation, and in each iteration the accuracy values obtained by testing the model with left-out fold are stored. Taking the mean of the accuracy values, mean cross validation accuracies are calculated for each C value and visualized on bar plot. The mean cross validation values can be seen below and visualization of the values can be seen in *Figure 1*.

- For C = 10^{-3} : 0.882,
- For C = 10^{-2} : 0.9359999999999999
- For C = 10^{-1} : 0.9399999999999998
- For C = 10^{0} : 0.9400000000000001
- For C = 10^{1} : 0.9400000000000001
- For C = 10^{2} : 0.9400000000000001

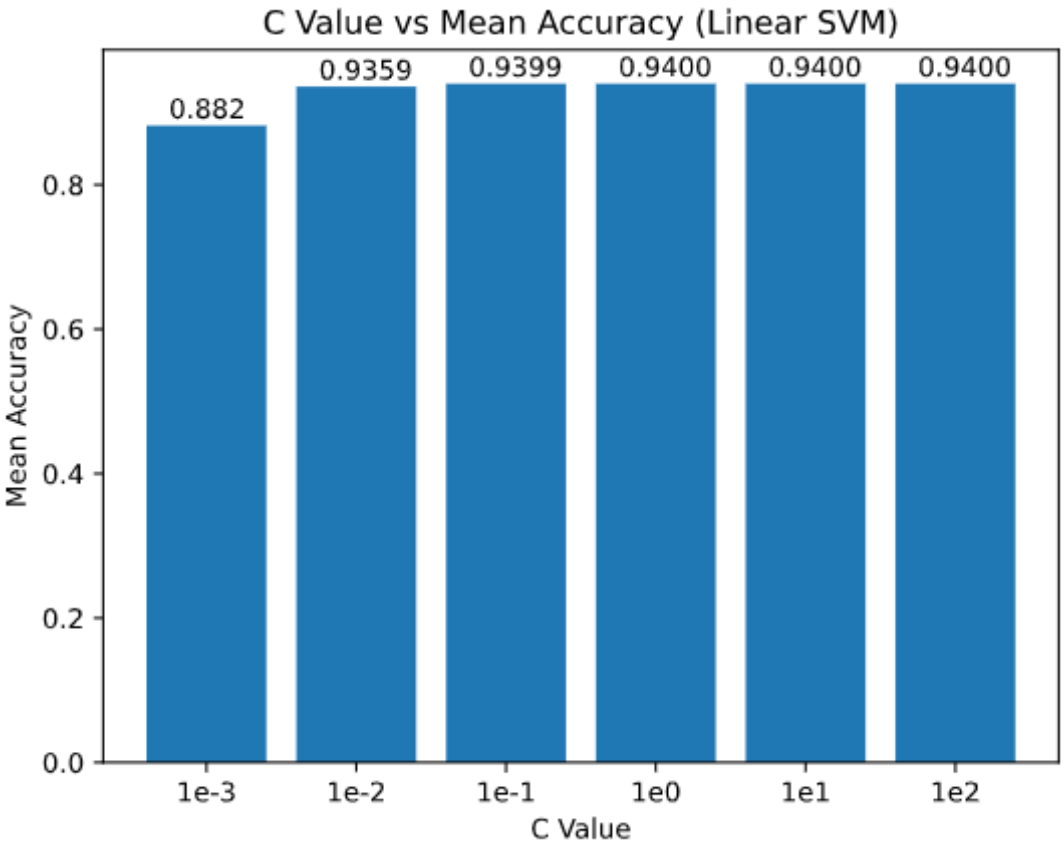

Figure 1: C Value vs Mean Cross Validation Accuracy (Linear SVM)

As it can be seen in *Figure 1*, the mean cross validation accuracy increases as the C value increases, and for the values 10^{-1}, 10^{0}, 10^{1}, 10^{2}, the values of mean accuracy can be considered as the same as being 0.94 since the accuracy value for C = 10^{-1} differs from the others by less than 0.0001. Considering these values 1 is selected as the optimum C value, as it is the middle of these values that is not too small or too large, and it is considered as default value for SVM.

After optimum C value is decided as 10^{-1}, SVM model is trained with training set using this C value hyper parameter. Then, the model is run on test set. The metrics calculated can be seen below and the resulting confusion matrix can be seen in *Figure 2*.

- Accuracy is: 0.9899497487437185
- Precision is: 0.9565217391304348
- Recall is: 1.0
- F1 Score is: 0.977777777777777
- F2 Score is: 0.990990990990991

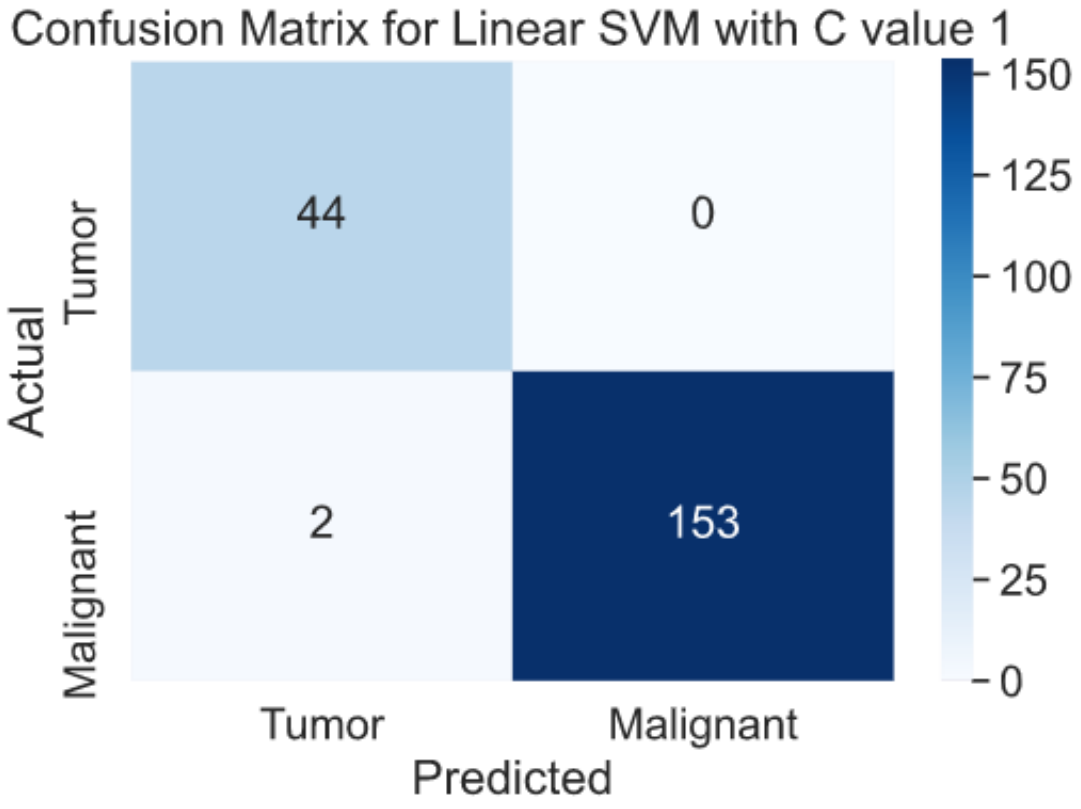

Figure 2: Confusion matrix for Linear SVM with C value 1

Question 1.2

In this part of the question, SVM model with radial basis function (RBF) kernel is trained with the C value determined in Question 1.1 and different gamma values including 2^{-4}, 2^{-3}, 2^{-2}, 2^{0}, 2^{1}. For each gamma value, SVM model is trained with 10-fold cross validation, and in each iteration the accuracy values obtained by testing the model with left-out fold are stored. Taking the mean of the accuracy values, mean cross validation accuracies are calculated for each gamma value and visualized on bar plot. The mean cross validation values can be seen below and visualization of the values can be seen in *Figure 3*.

- For C = 2^{-4} : 0.9400000000000002
- For C = 2^{-3} : 0.9440000000000002
- For C = 2^{-2} : 0.932
- For C = 2^{0} : 0.858
- For C = 2^{1} : 0.7940000000000002

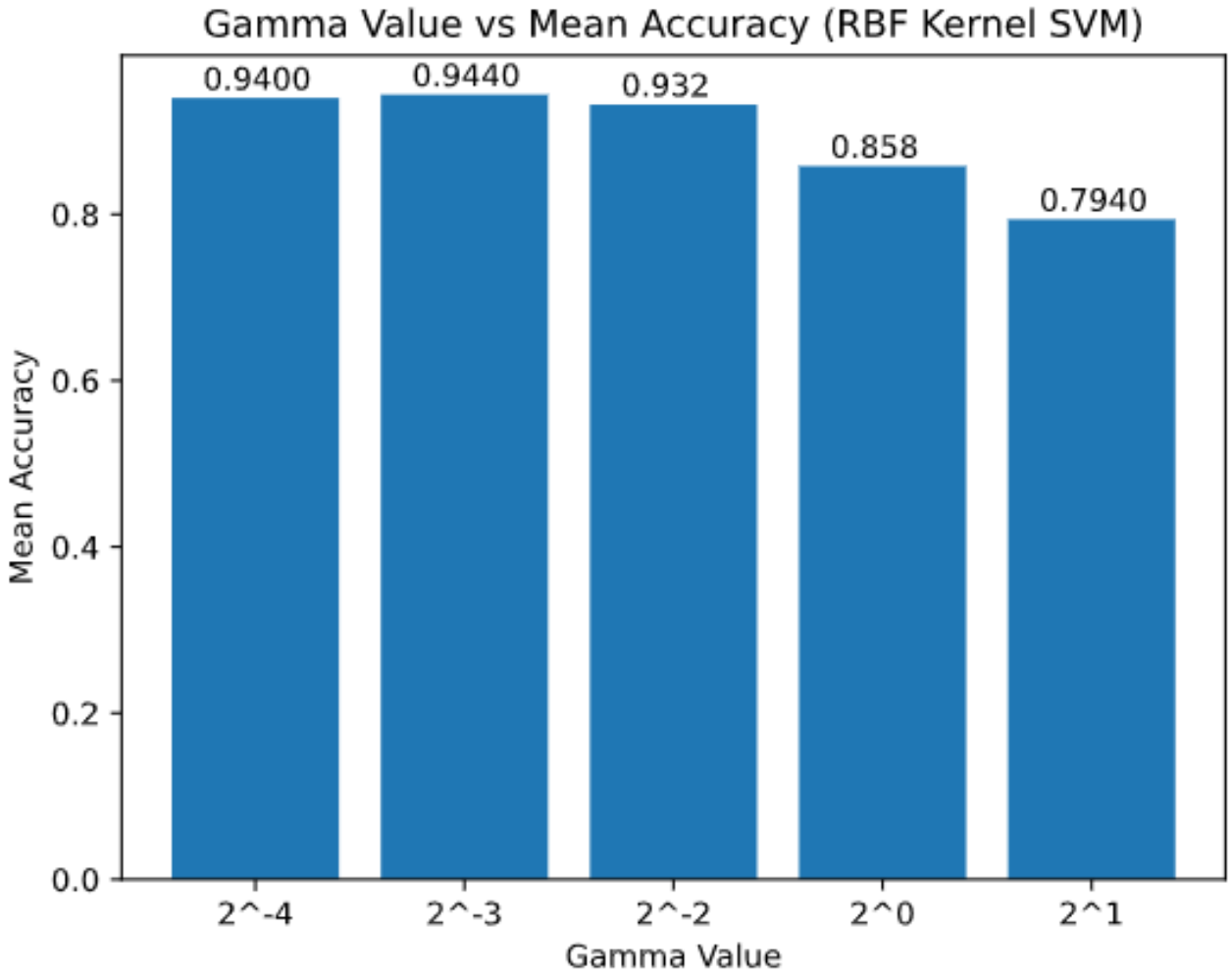

Figure 3: Gamma Value vs Mean Cross Validation Accuracy (RBF Kernel SVM)

As it can be seen in *Figure 3*, the mean cross validation accuracy is maximum when gamma value is 2^{-3}. Considering this fact values 2^{-3} is selected as the optimum gamma value.

After optimum gamma value is decided as 2^{-3}, SVM model is trained with training set using this gamma value hyper parameter. Then, the model is run on test set. The metrics calculated can be seen below and the resulting confusion matrix can be seen in *Figure 4*.

- Accuracy is: 0.9849246231155779
- Precision is: 0.9361702127659575
- Recall is: 1.0
- F1 Score is: 0.967032967032967
- F2 Score is: 0.9865470852017937

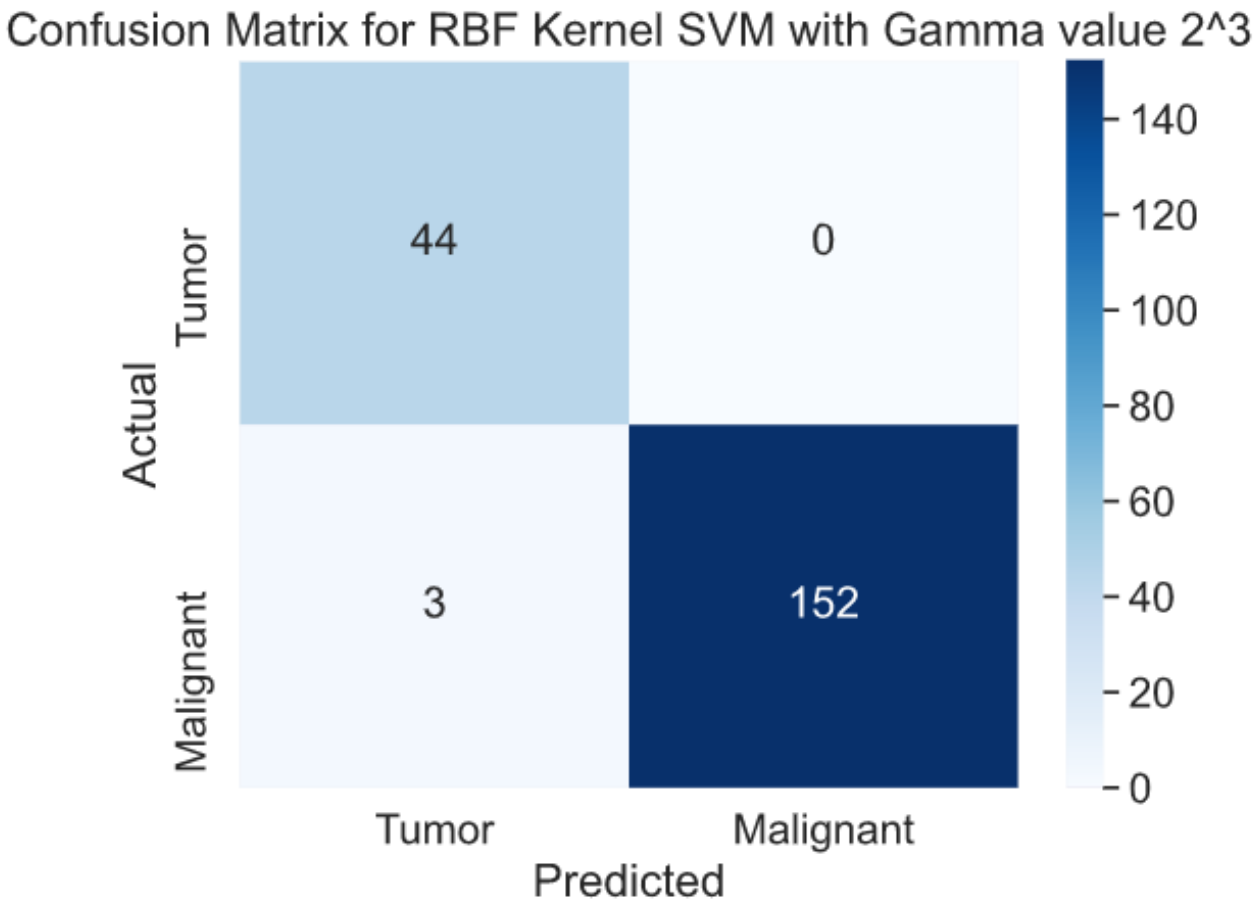

Figure 4: Confusion matrix for RBF Kernel SVM with Gamma value 2^{-3}