CENG499 HW3 Report

Onat Özdemir

January 15, 2022

1 DT

1.1 Info Gain without Prepruning

The visualization of the tree can be found in **info_gain_without_prepruning.pdf** file under visualizations folder. Test set accuracy of this model was found as 0.9333333333333333.

1.2 Info Gain with Prepruning

The visualization of the tree can be found in **info_gain_with_prepruning.pdf** file under visualizations folder. Test set accuracy of this model was found as 0.9333333333333333.

1.3 Avg Gini Index without Prepruning

The visualization of the tree can be found in **avg_gini_without_prepruning.pdf** file under visualizations folder. Test set accuracy of this model was found as 0.9.

1.4 Avg Gini Index with Prepruning

The visualization of the tree can be found in **avg_gini_with_prepruning.pdf** file under visualizations folder. Test set accuracy of this model was found as 0.9.

2 SVM

2.1 Task 1

As can be seen in the plots given below, as C becomes larger, margin becomes smaller. That means, by selecting larger C values, we can achieve higher classification accuracy. But as a drawback, we would have smaller margins.

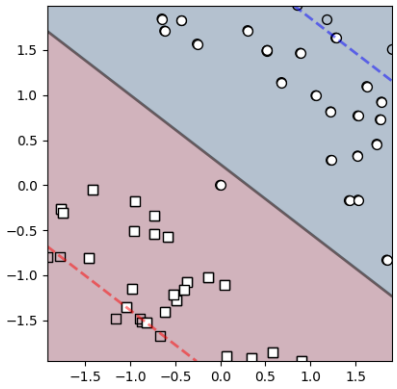

Figure 1: Classifier with C = 0.01

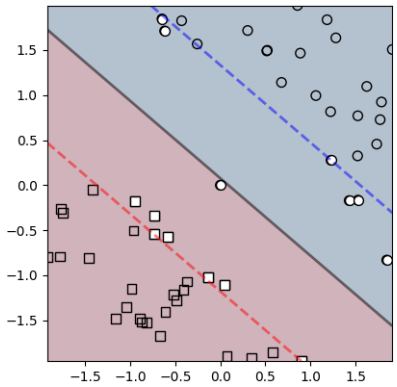

Figure 2: Classifier with C = 0.1

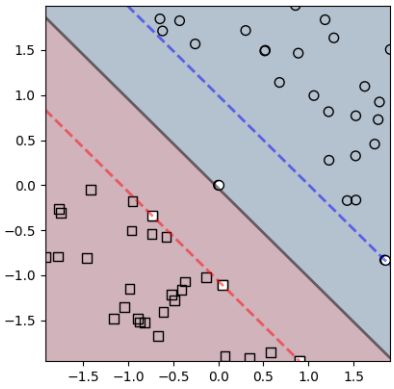

Figure 3: Classifier with C = 1

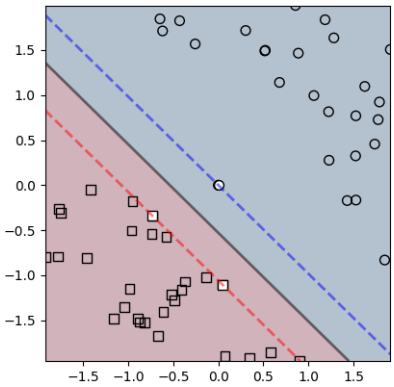

Figure 4: Classifier with C = 10

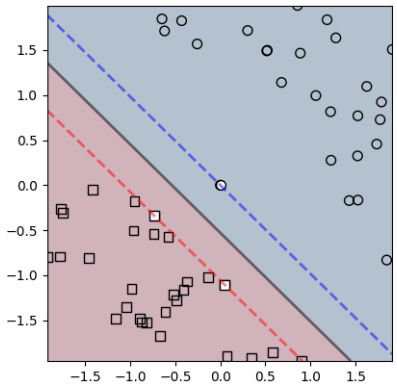

Figure 5: Classifier with C = 100

2.2 Task 2

By analyzing the plots given below, we can make some observations. First of all, change in kernel can affect the boundary lines and the classification accuracy, dramatically. For instance, in this dataset, while rbf works quite well, choosing linear and polynomial kernels would not be a good option.

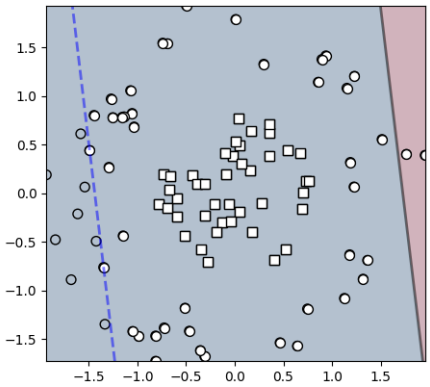

Figure 6: Classifier with kernel = linear

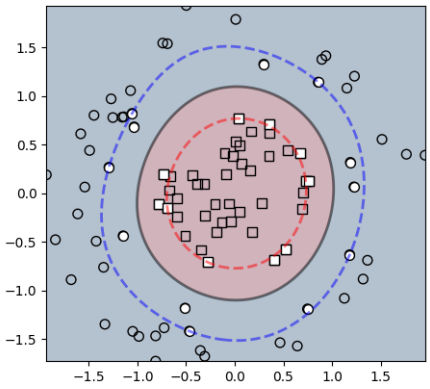

Figure 7: Classifier with kernel = rbf

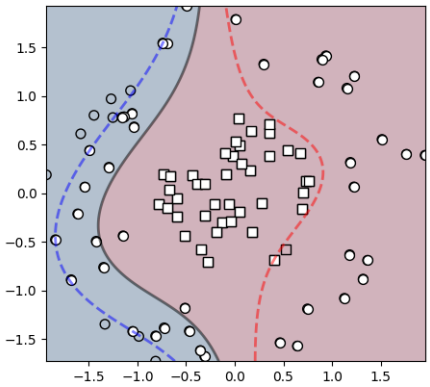

Figure 8: Classifier with kernel = polynomial

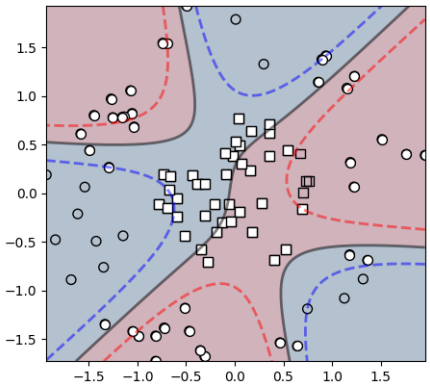

Figure 9: Classifier with kernel = sigmoid

2.3 Task 3

For this part, I used 5-fold Cross-validation and tried 18 different configurations. Below results are the average of the validation accuracy results obtained from these 5 splits.

2.3.1 Gamma = 0.001

	Kernel		
C	Linear	Rbf	Polynomial
0.1	0.734	0.706	0.572
1	0.692	0.761	0.689
10	0.664	0.794	0.736

Table 1: Mean Validation Accuracy Results

2.3.2 Gamma = 0.01

	Kernel		
C	Linear	Rbf	Polynomial
0.1	0.734	0.774	0.770
1	0.692	0.813	0.787
10	0.664	0.831	0.766

Table 2: Mean Validation Accuracy Results

2.3.3 The Best Configuration

The best validation accuracy was obtained when C = 10, gamma = 0.01, kernel = rbf. The test set accuracy of the best configuration was calculated as **0.815**.

2.4 Task 4

2.4.1 Imbalanced

Test Set Accuracy: **0.95**

Actual Class / Predicted Class	0	1
0	0	50
1	0	950

Table 3: Confusion Matrix

In this case, accuracy is not a good metric. As can be seen in the confusion matrix, although the model classifies every instance as 1, its accuracy is quite high. In imbalanced datasets, different metrics for instance f1 score can be more benefitial than accuracy.

2.4.2 Oversampled

Test Set Accuracy: **0.951**

Actual Class / Predicted Class	0	1
0	12	38
1	11	939

Table 4: Confusion Matrix

The highest accuracy is achieved in the oversampled case.

2.4.3 Undersampled

Test Set Accuracy: **0.798**

Actual Class / Predicted Class	0	1
0	33	17
1	185	765

Table 5: Confusion Matrix

Among all the cases, the lowest accuracy belongs to undersampled case. In undersampling to balance the dataset, we lose some information by deleting samples from dataset. This information loss might cause this low accuracy score.

2.4.4 Class Weight Balanced

Test Set Accuracy: **0.936**

Actual Class / Predicted Class	0	1
0	16	34
1	30	920

Table 6: Confusion Matrix