CENG499 HW3 Report

Yusuf Eren Tunç

January 9, 2021

1 Part 1: Decision Tree

1.1 Information Gain

Test accuracy is 0.93
Referring diagram is Information Gain Decision Tree pdf.

1.2 Gain Ratio

Test accuracy is 0.94
Referring diagram is Gain Ratio Decision Tree pdf.

1.3 Average Gini Index

Test accuracy is 0.89
Referring diagram is Average Gini Index Decision Tree pdf.

1.4 Gain Ratio with Chi-squared Pre-pruning

Test accuracy is 0.94
Referring diagran is Gain Ratio Decision Chi Squarred Tree pdf.

1.5 Gain Ratio with Reduced Error Post-pruning

Due to my fatal error, I couldn't manage to create decision tree. When I started to homework, I kept decision tree as a normal python array. Because of that, I get stack overflow errors almost every time at gain ratio with reduced error post-pruning decision tree implementation.

2 Part 2: Support Vector Machine

2.1 First Part

C is mainly act at avoiding misclassifying at each training example. For very small C values, we get wrong classified predictions. For very large C values, we don't neglect outliers but that time margin get too small. We can see that effects on figures.

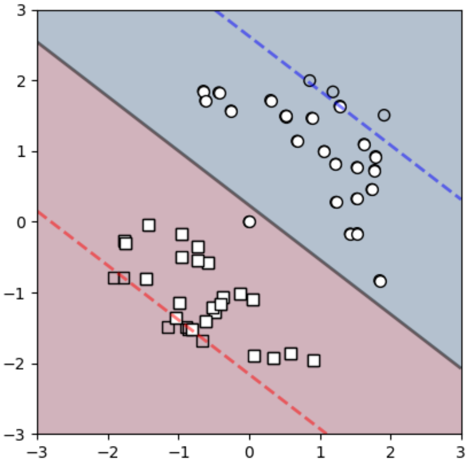

Figure 1: C = 0.01

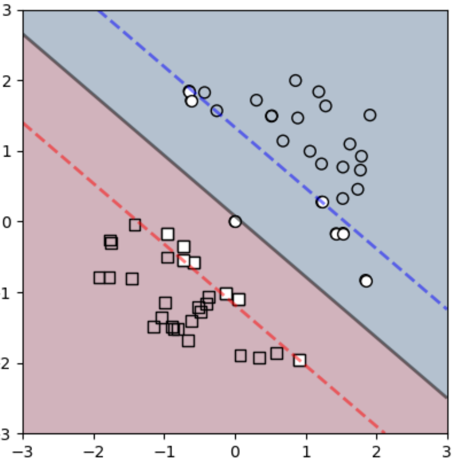

Figure 2: C = 0.1

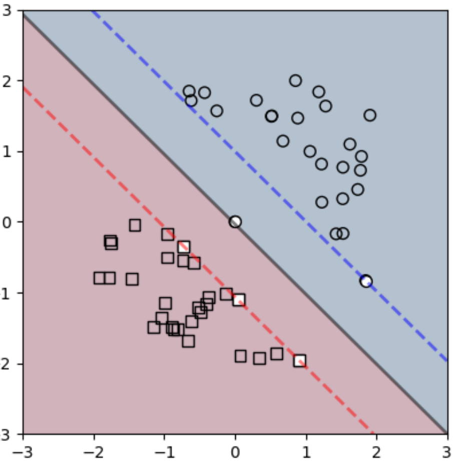

Figure 3: C = 1

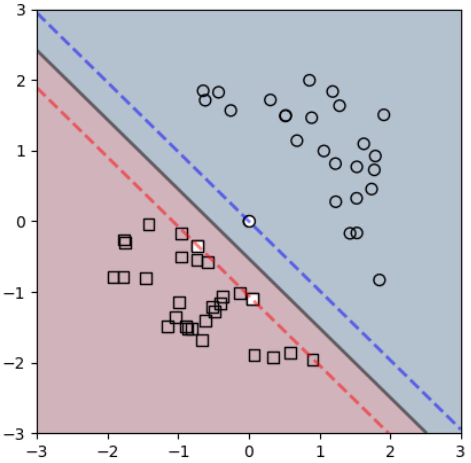

Figure 4: C = 10

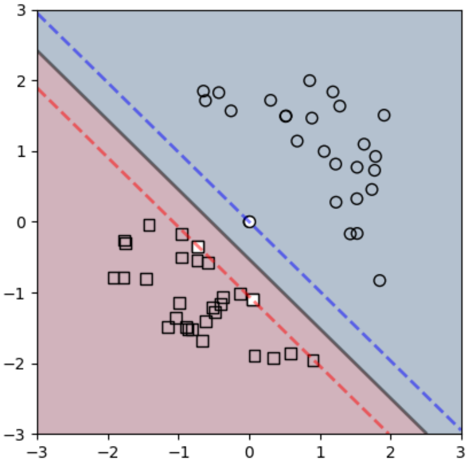

Figure 5: C = 100

2.2 Second Part

Since we had nonlinear seperation, only rbf kernel successfully classified test data because rbf kernel is much better at nonlinear data classification.

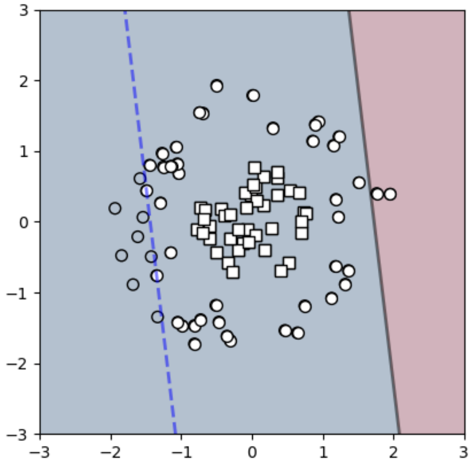

Figure 6: Linear Kernel

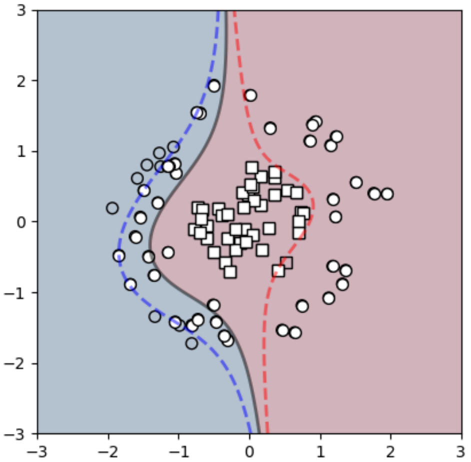

Figure 7: Polynomial Kernel

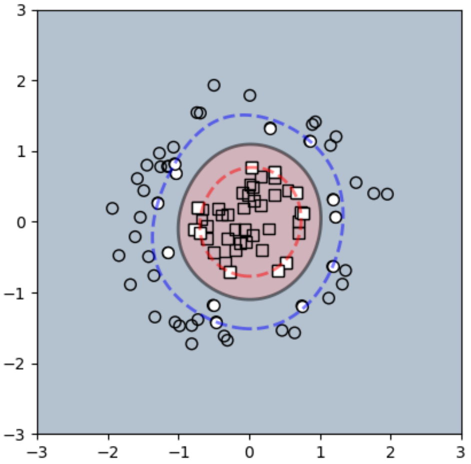

Figure 8: RBF Kernel

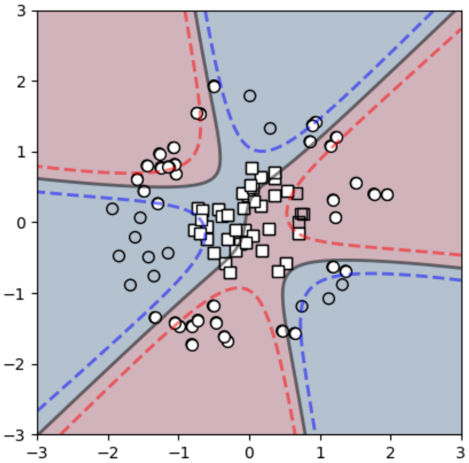

Figure 9: Sigmoid Kernel

2.3 Third Part

gamma	C				
	0.01	0.1	1	10	100
-	0.63	0.65	0.70	0.71	0.71

Table 1: Linear kernel

Selected hyperparameters are RBF kernel with 1 C and 1 Gamma. Test accuracy is 0.75.

gamma	C				
	0.01	0.1	1	10	100
0.00001	0.54	0.54	0.54	0.54	0.54
0.0001	0.54	0.54	0.54	0.55	0.54
0.001	0.54	0.54	0.54	0.56	0.62
0.01	0.73	0.71	0.71	0.56	0.62
0.1	0.68	0.74	0.71	0.71	0.62
1	0.65	0.73	0.74	0.71	0.71

Table 2: RBF kernel

gamma	C				
	0.01	0.1	1	10	100
0.00001	0.54	0.54	0.54	0.68	0.73
0.0001	0.73	0.54	0.54	0.56	0.73
0.001	0.73	0.73	0.54	0.54	0.61
0.01	0.73	0.73	0.73	0.54	0.54
0.1	0.68	0.73	0.73	0.73	0.54
1	0.56	0.73	0.73	0.73	0.73

Table 3: Polynomial kernel

.

gamma	C				
	0.01	0.1	1	10	100
0.00001	0.54	0.54	0.54	0.54	0.54
0.0001	0.54	0.54	0.54	0.54	0.54
0.001	0.54	0.54	0.54	0.54	0.56
0.01	0.50	0.54	0.54	0.54	0.61
0.1	0.54	0.48	0.54	0.54	0.61
1	0.63	0.54	0.48	0.54	0.54

Table 4: Sigmoid kernel

2.4 Fourth part

2.4.1 Without handling the imbalance problem

Test accuracy is 0.84.

Accuracy can't be only performance metric. It is relies only test data. If test data has some imbalances or such problems, accuracy wouldn't reflect to performance of model. We can use confusion matrix for better performance metrics.//

Confusion matrix is // True positive is 1 and it means that model predict positive and it is true.

1	190
0	949

Table 5: Confusion matrix

True negative is 949 and it means that model predict negative and it is true.

False Positive is 190 and it means that model predict positive. However, it is negative.

False Positive is 0 and it means that model predict negative. However, it is positive.

2.4.2 Oversampling the minority class

Test accuracy is 0.81.
Confusion matrix is

64	127
92	857

Table 6: Confusion matrix

True positive is 64 and it means that model predict positive and it is true.
True negative is 857 and it means that model predict negative and it is true.
False Positive is 127 and it means that model predict positive. However, it is negative.
False Positive is 92 and it means that model predict negative. However, it is positive.
Oversampling doesn't affect test accuracy much. However, It causes improvement at confusion matrix. While, our test accuracy stay same, our model was improved.

2.4.3 Undersampling the majority class

Test accuracy is 0.83.
Confusion matrix is

24	167
26	923

Table 7: Confusion matrix

True positive is 24 and it means that model predict positive and it is true.
True negative is 923 and it means that model predict negative and it is true.
False Positive is 167 and it means that model predict positive. However, it is negative.
False Positive is 26 and it means that model predict negative. However, it is positive.
Under sampling doesn't affect test accuracy much. However, It causes improvement at confusion matrix. Moreover, It seems that, it is a bit much ineffective than Oversampling. While, our test accuracy stay same, our model was improved.

2.4.4 Setting the class_weight to balanced

Test accuracy is 0.75.
Confusion matrix is

87	104
173	776

Table 8: Confusion matrix

True positive is 87 and it means that model predict positive and it is true.
True negative is 776 and it means that model predict negative and it is true.
False Positive is 104 and it means that model predict positive. However, it is negative.
False Positive is 173 and it means that model predict negative. However, it is positive.
With balanced class weight, confusion matrix values get much closer than without balance. However, there is a significant decrease at test accuracy. Therefore, it improve confusion matrix but effects to test accuracy badly. It looks like, it is a loss trade-off.