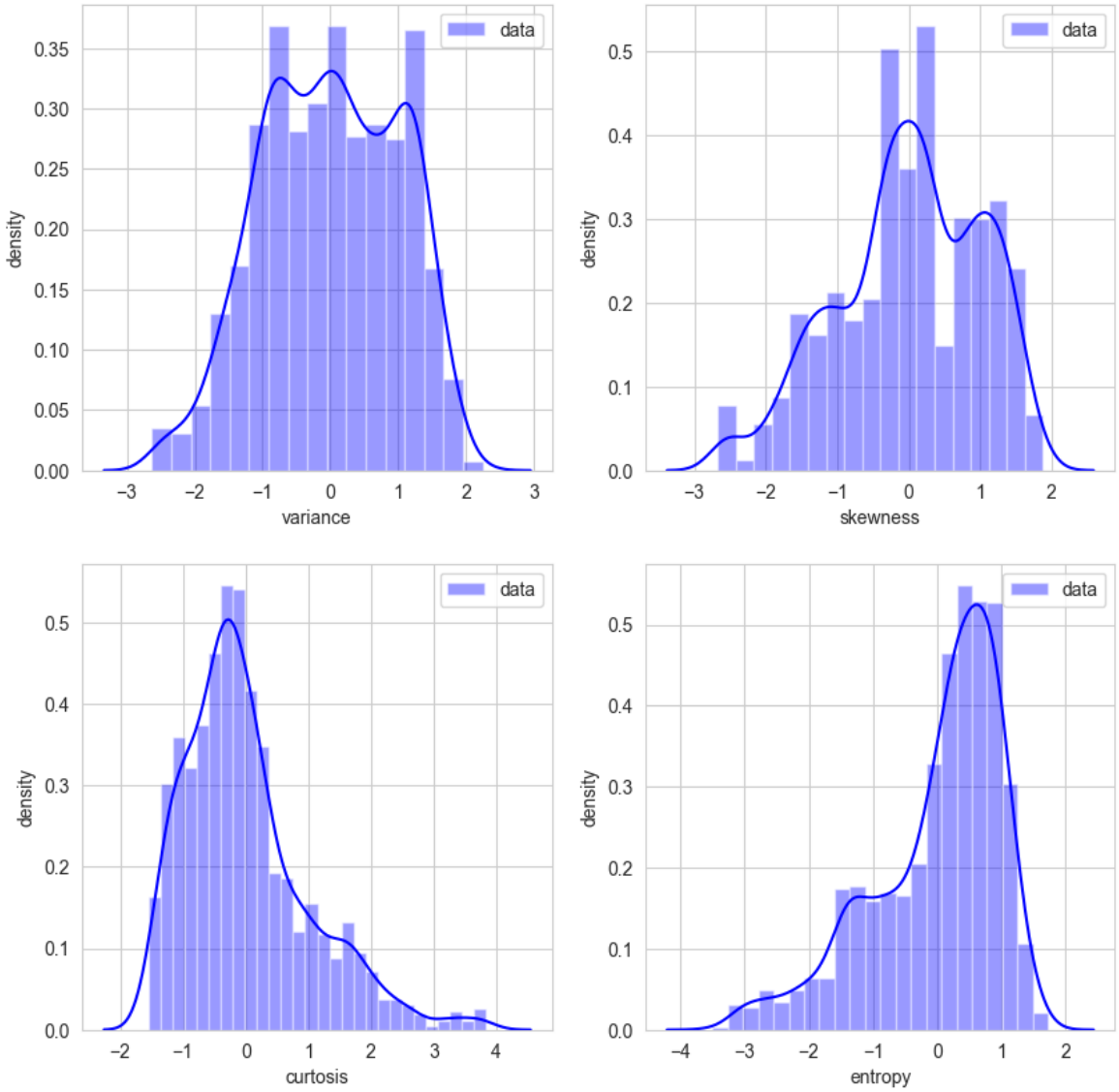
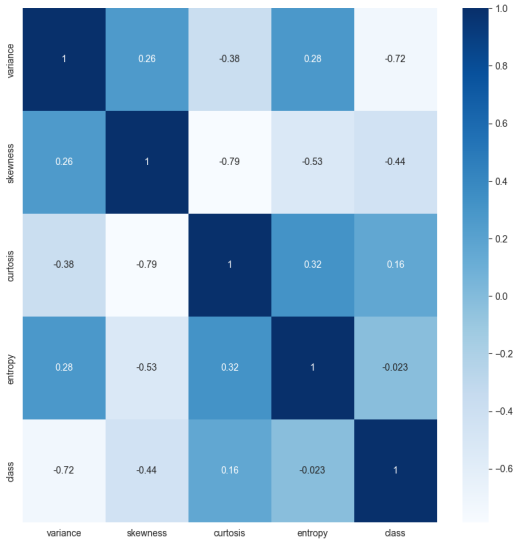
PRML LAB REPORT 11

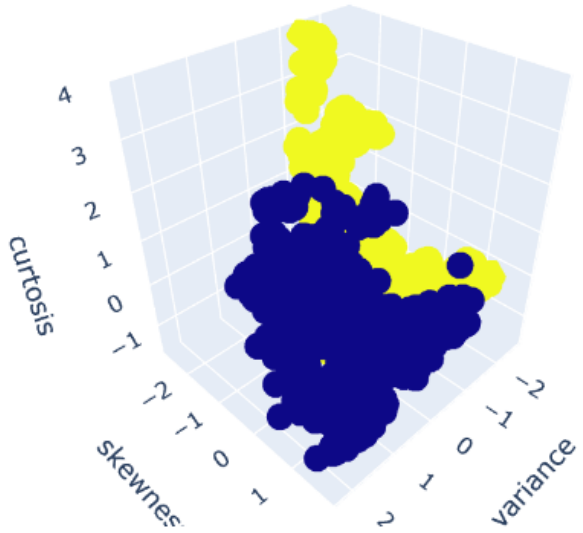
MANISH(B21CS044)

Q.1 Pre-Process the dataset by handling missing values and normalizing the data. Split in the ratio 70:20:10 for train-test-validation. [10 Marks]

In this part I have preprocessed the data and normalized it using standard scalar.
Here are some plots for the visualization of the dataset :

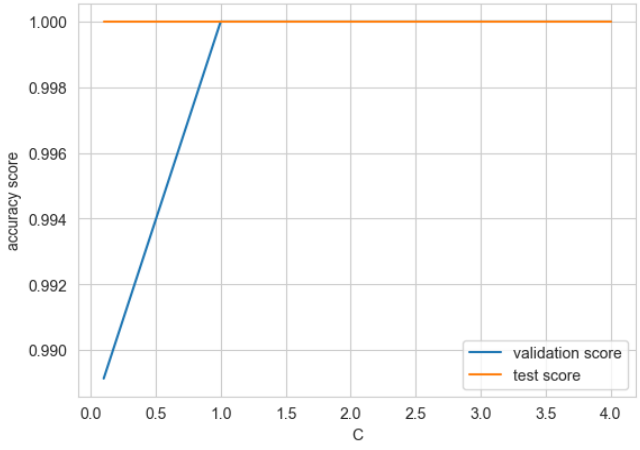






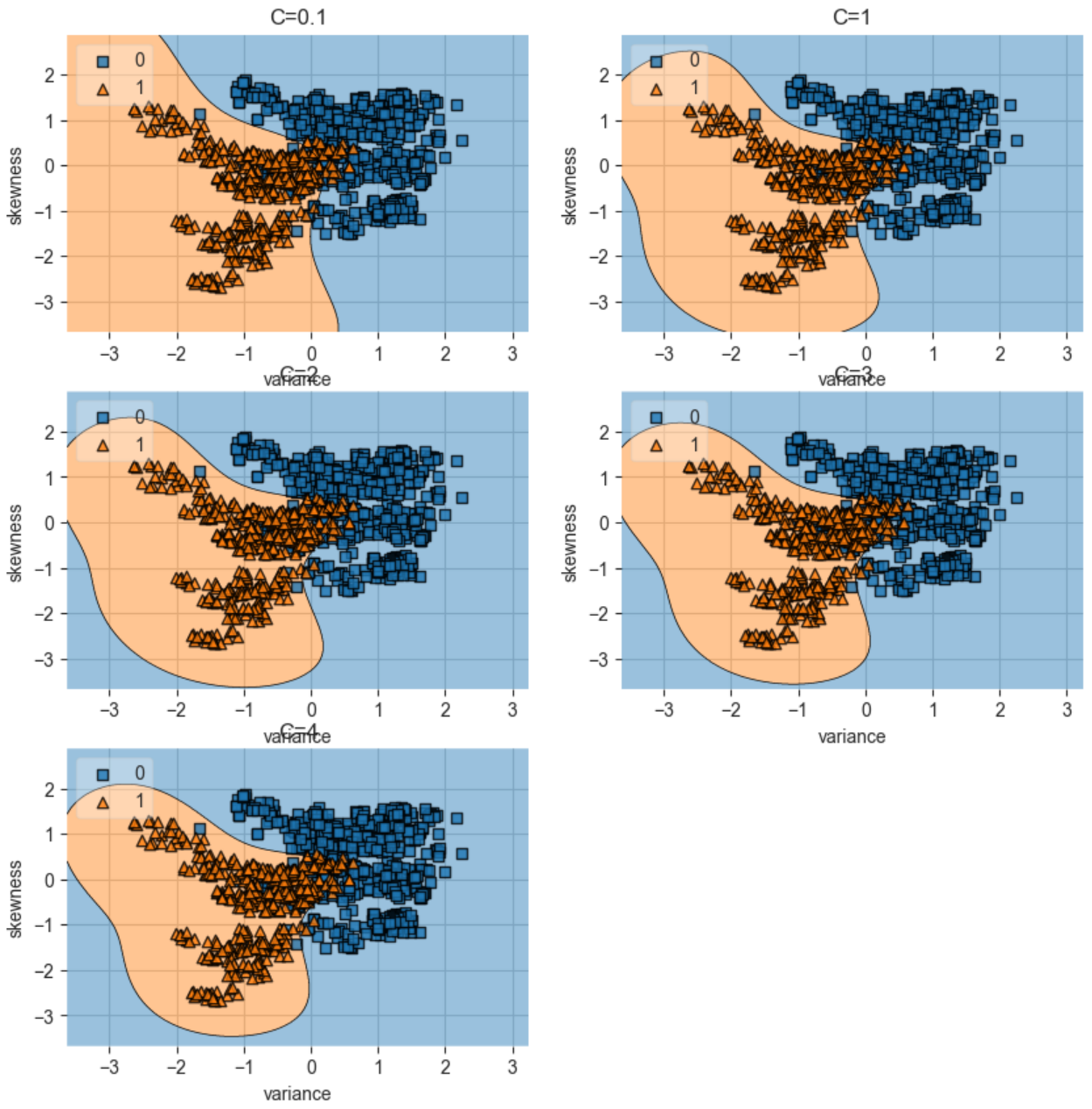
Q2. Train an SVM classifier using Sklearn library. Analyze the performance (classification accuracy) for different values of 'C'. (Choose at least 5 different values of C) [25 marks]

In this part I have selected 5 different values of C as [0.1,1,2,3,4]. Here are the results for the same:




```
 FROM ABOVE GRAPH WE CAN SEE THAT THE BEST VALUE OF C IS NEAR 1 WHERE THE
TEST AND VALIDATION ACCURACY IS SATURATED AS WELL AS THE DIFFERENCE
BETWEEN THEM IS MINIMUM.
```

Here are the decision Boundaries for the same:



As we can see, as we increase the value of C, the number of misclassified points decreases and the decision boundaries more precisely classify the points, but the model also becomes more complex and overfits the data.

Q3. Use various types of kernels(RBF, Linear, Quadratic etc) and train the SVM model using the Sklearn library. Plot the decision boundary for different svm models trained. [25 marks]

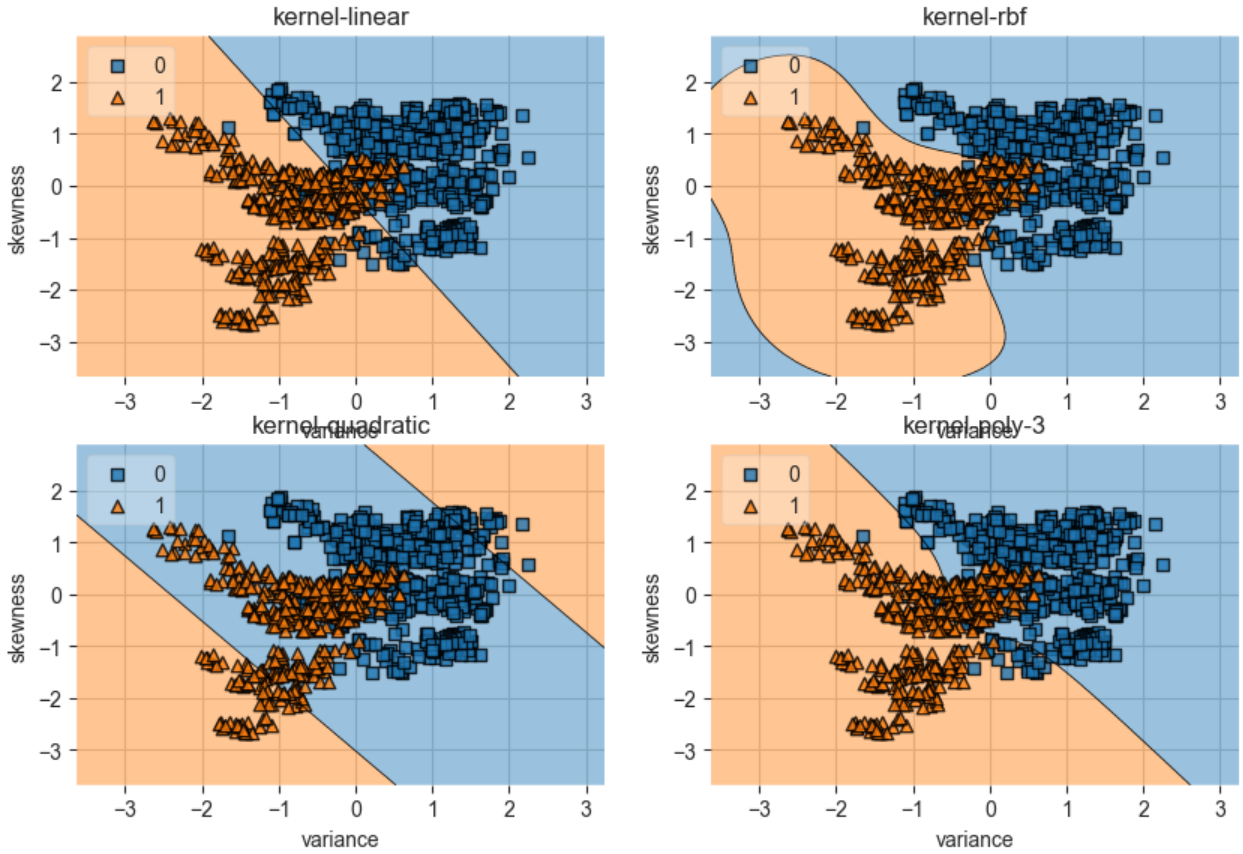
Note: While solving questions involving plotting the decision boundary, it is recommended to first perform dimension reduction of the dataset into two/three dimensions using LDA. And then plot the decision boundaries on the dataset, to gain a better visual understanding.

Instead of this, you can also choose to plot the decision boundary using the 2 features from the dataset, who have the highest correlation with the target.

In this part of applying LDA, it gives only 1 principal component. So, we can't use LDA in this part as we will not be able to Plot decision boundaries on this data.

As we have also visualized the dataset, Variance and Skewness have high correlation with the target column. So, we will be using these two features to plot decision boundaries.

Here we have used 4 kernel : linear , quadratic , 3-degree poly and rbf.



```
As we see from the decision boundary plots, the dataset is not linear. So,
the linear kernel is not performing well.

whereas the RBF kernel is performing well as it is able to classify the
points properly. because it is tightly fitting the dataset.

Also, the quadratic kernel is not performing well as it is underfitting
the dataset.

But the 3 degree polynomial kernel is performing well as it is able to
classify the points properly. It is able to go through the part where
points from different classes are mixed together.
```

Q4) Use the online tool: Link. It is an interactive SVM visualizer. Initially, one needs to manually add data points to the graph and then choose various values of hyperparameters (C, gamma, kernel, etc.). After that, the tool outputs the decision boundary and also highlights the support vectors.

For this question, create 2 types of datasets, according to your preferences, and analyze how each hyperparameter affects the SVM model.

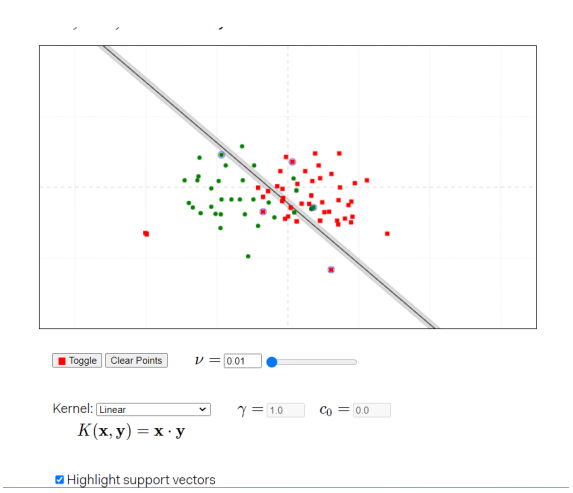
To do so, try changing the kernel, C, gamma, etc. and add screenshots of the outputs in the report. Try to perform hyperparameter tuning to generate the best possible decision boundary for each dataset created.

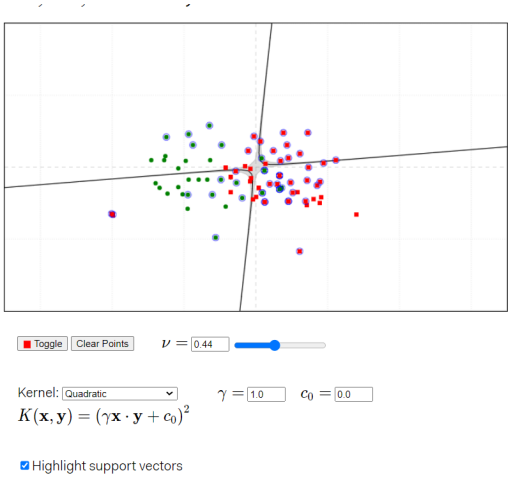
Also, mention your observations in the report along with potential reasons for your observations. [20+20 marks]

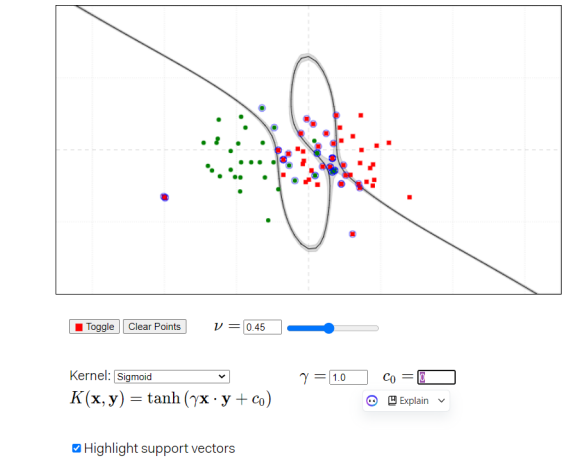
For this part, I have created two types of datasets. The first dataset is linearly separable, and the second type of dataset is a type of Xor dataset, which is not linearly separable. Here are the screenshots for the SVM models using different hyperparameters.

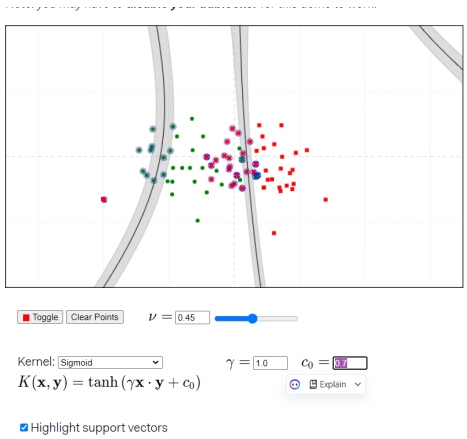
Here are the results for the first type of dataset:

For this part I have used kernel functions as Linear , quadratic and sigmoid and varied C and nu for the dataset and results are that since this dataset is linear and easily separable,So almost all types of kernels are able to correctly separate the boundaries of the dataset.









Here are the results of the second type of dataset.

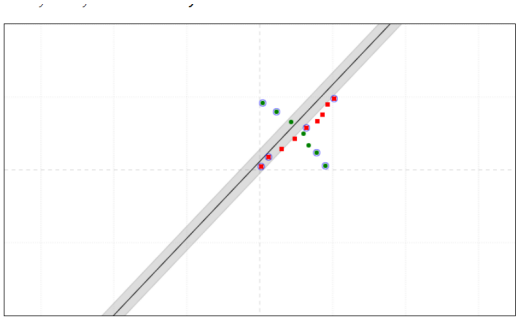
Here also, I have used linear, quadratic, and sigmoid as kernel functions and varied the hyperparameters.

As the dataset is not linear. So, the linear kernel is not able to identify boundaries between classes.

whereas quadratic and sigmoid are able to separate the boundaries easily.

Also , as we saw the effect of C in the above and this dataset, as we increase the value of C, the model is able to classify the boundaries with more precision as expected.

And also on decreasing the nu it performs well.

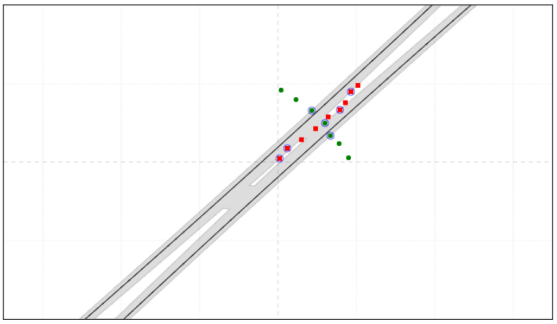


Toggle Clear Points $\nu =$ 0.27

Kernel: Linear $\gamma =$ 1.0 $c_0 =$ 0.0

$$K(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$$

Highlight support vectors

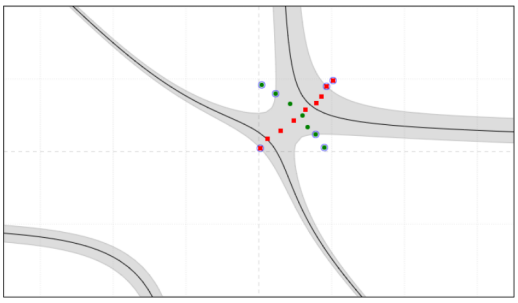


Toggle Clear Points $\nu =$ 0.27

Kernel: Quadratic $\gamma =$ 0.9 $c_0 =$ 0.0

$$K(\mathbf{x}, \mathbf{y}) = (\gamma \mathbf{x} \cdot \mathbf{y} + c_0)^2$$

Highlight support vectors



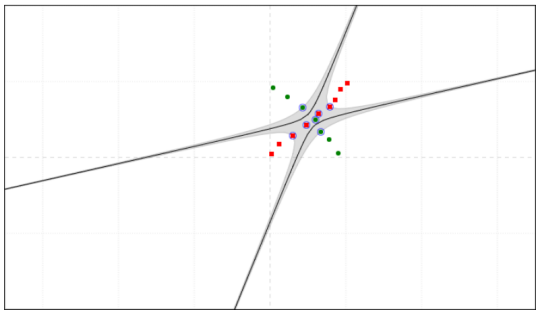
Toggle Clear Points $\nu =$ 0.27

Kernel: Sigmoid $\gamma =$ 0.9 $c_0 =$ 0.5

$$K(\mathbf{x}, \mathbf{y}) = \tanh(\gamma \mathbf{x} \cdot \mathbf{y} + c_0)$$

Explain

Highlight support vectors



Toggle Clear Points $\nu =$ 0.27

Kernel: Quadratic $\gamma =$ 0.9 $c_0 =$ 0.5

$$K(\mathbf{x}, \mathbf{y}) = (\gamma \mathbf{x} \cdot \mathbf{y} + c_0)^2$$

Explain

Highlight support vectors