Machine Learning Lab: Support Vector Regression
Nikhil Saraogi [21MCA1080]

Objective- My task is to predict whether a bank currency note is authentic or not based upon four attributes of the note i.e. skewness of the wavelet transformed image, variance of the image, entropy of the image, and kurtosis of the image. This is a binary classification problem and I'm using SVM algorithm to solve this problem

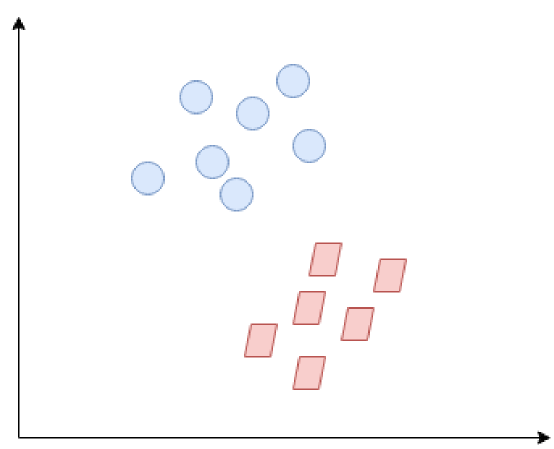
Definition & Working principle

SVR:

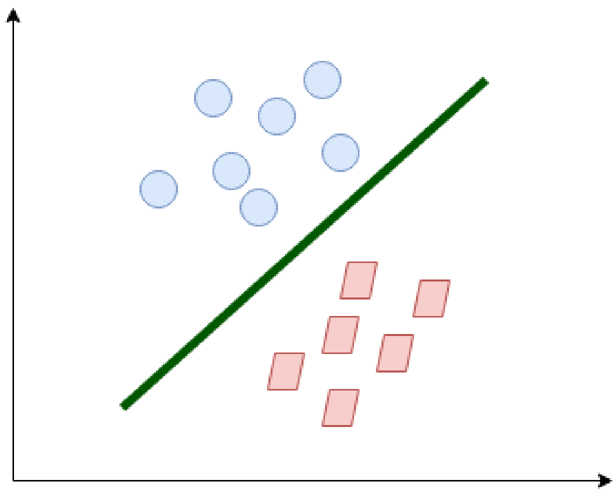
Support Vector Machines (SVMs in short) are machine learning algorithms that are used for classification and regression purposes. SVMs are one of the powerful machine learning algorithms for classification, regression and outlier detection purposes. An SVM classifier builds a model that assigns new data points to one of the given categories. Thus, it can be viewed as a non-probabilistic binary linear classifier.

Basic Explanation:

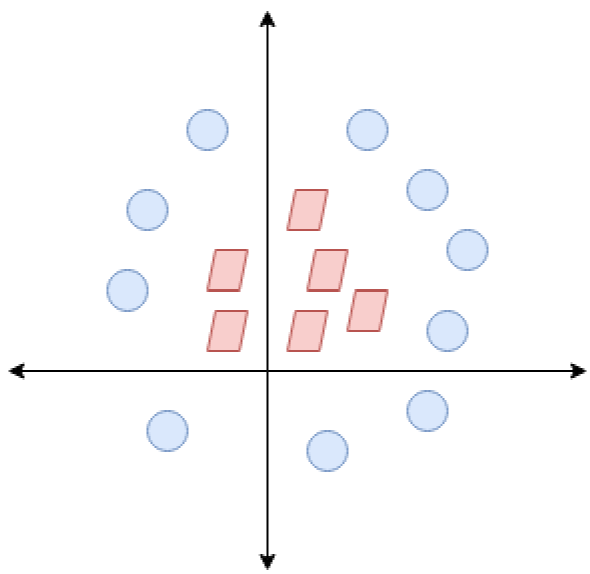
Let's say we have a plot of two label classes as shown in the figure below:



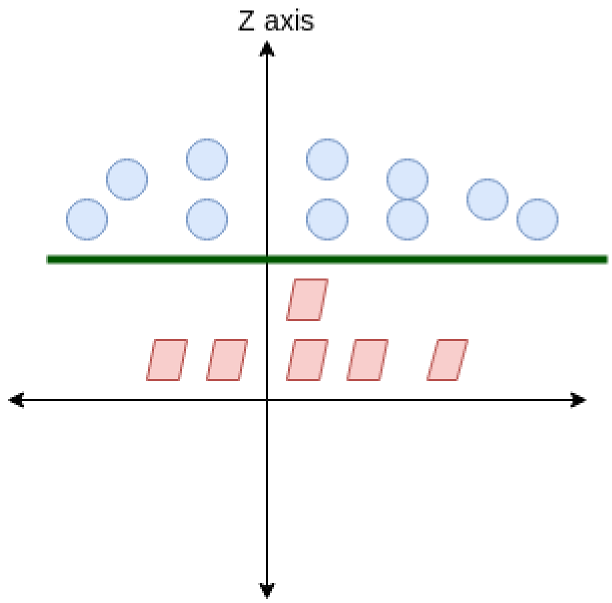
If we try to separate the classes , we might come up with this:



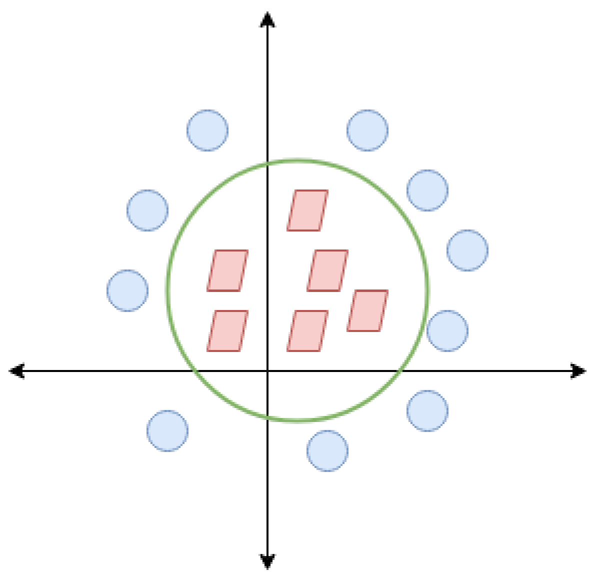
The line fairly separates the classes. This is what SVM essentially does – simple class separation. Now, what is the data was like this:



Here, we don't have a simple line separating these two classes. So we'll extend our dimension and introduce a new dimension along the z-axis. Wecan now separate these two classes:



When we transform this line back to the original plane, it maps to the circularboundary as I’ve shown here:

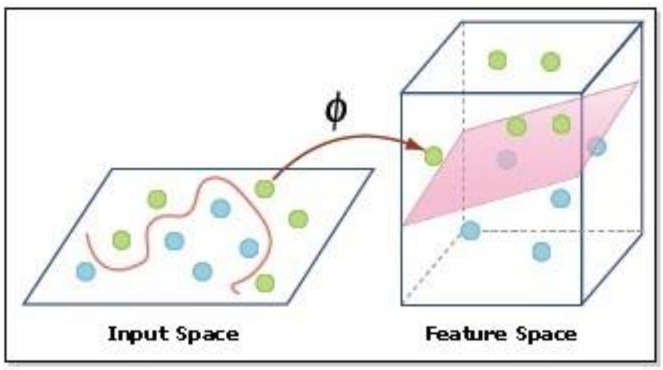


SVM performs just this! In multidimensional space, it looks for a line orhyperplane that divides these two classes. Then, using the classes to forecast, it assigns the new point a classification based on whether it islocated on the positive or negative side of the hyperplane.

Hyperparameters of the Support Vector Machine (SVM) Algorithm

There are a few important parameters of SVM that you should be aware ofbefore proceeding further:

- **Kernel:** Without raising the computing cost, a kernel aids in the discovery of a hyperplane in the higher dimensional space. Usually, asthe dimension of the data grows, the computing cost grows as well.
  When we can't go in a certain dimension because there isn't a dividinghyperplane there, we must move in a higher dimension instead.

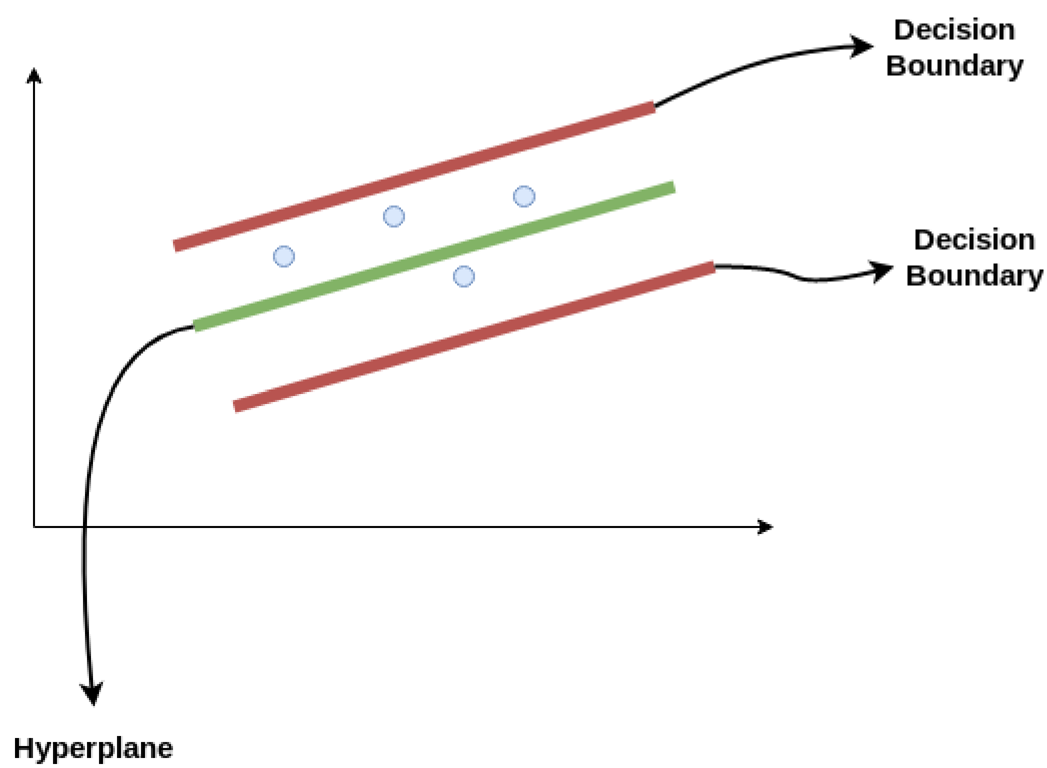


- **Hyperplane:** In SVM, this essentially acts as a boundary between twodata classes. However, this line will be utilised in Support Vector
  Regression to forecast the continuous output.

- **Decision Boundary:** A decision boundary can be conceptualised as ademarcation line (for the sake of simplification) where the positive examples are on one side and the negative examples are on the other.The instances can be categorised as either positive or negative along this exact line. Support Vector Regression will also use this same SVMconcept.

Support Vector Regression (SVR) uses the same principle as SVM, butfor regression problems.

The problem of regression is to find a function that approximates mapping from an input domain to real numbers on the basis of a training sample. Solet’s now dive deep and understand how SVR works actually.



Think of the green line as the hyperplane, and the two red lines as the decision boundaries. When using SVR, our goal is to essentially take intoaccount the points that are inside the decision boundary line. The hyperplane with the most points serves as our best fit line.

The decision boundary (the hazardous red line above!) is the first concept that we will comprehend. Think of these lines as being at any distance fromthe hyperplane, let's say 'a'. The lines that we draw at '+a' and '-a' distancesfrom the hyperplane are thus as follows. The text basically refers to this 'a' as epsilon.

Hyperplane

A hyperplane is a decision boundary which separates between given set of data points having different class labels. The SVM classifier separates data points using a hyperplane with the maximum amount of margin. This hyperplane is known as the maximum margin hyperplane and the linear classifier it defines is known as the maximum margin classifier.

Support Vectors

Support vectors are the sample data points, which are closest to the hyperplane. These data points will define the separating line or hyperplanebetter by calculating margins.

Margin

A margin is a separation gap between the two lines on the closest data points. It is calculated as the perpendicular distance from the line to supportvectors or closest data points. In SVMs, we try to maximize this separation gap so that we get maximum margin.

DATASET DESCRIPTION:

Dataset Name-**Bank Note Authentication UCI data**

Link- https://www.kaggle.com/datasets/ritesaluja/bank-note-authentication-uci-data

Data were extracted from images that were taken from genuine and forged banknote-like specimens. For digitization, an industrial camera usually used for print inspection was used. The final images have 400x 400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. Wavelet Transform tool were used to extract features from images.

1. variance of Wavelet Transformed image (continuous)
2. skewness of Wavelet Transformed image (continuous)
3. curtosis of Wavelet Transformed image (continuous)
4. entropy of image (continuous)
5. class (integer)

.

LIBRARIES USED:

NUMPY:

A general-purpose array processing package is called NumPy. It offers a high-performance multidimensional array object as well as utilities for interacting with these arrays. It is the core Python package for scientific computing. It is open-source software. It has a number of characteristics, including these crucial ones.An effective N-dimensional array object sophisticated (broadcasting) operations.Tools for combining C/C++ and Fortran code useful linear algebra, fourier transform, and random number capabilities.

NumPy can be used as a productive multi-dimensional container of generic data in addition to its apparent scientific applications. Numpy's ability to establish any data-types enables NumPy to quickly and easily interact with a wide range of databases.

PANDAS:

Pandas is an open-source library designed primarily for working quicklyand logically with relational or labeled data. It offers a range of data structures and procedures for working with time series and numerical data The NumPy library serves as the foundation for this library. Pandas is quick and offers its users exceptional performance & productivity.

MATPLOTLIB:

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

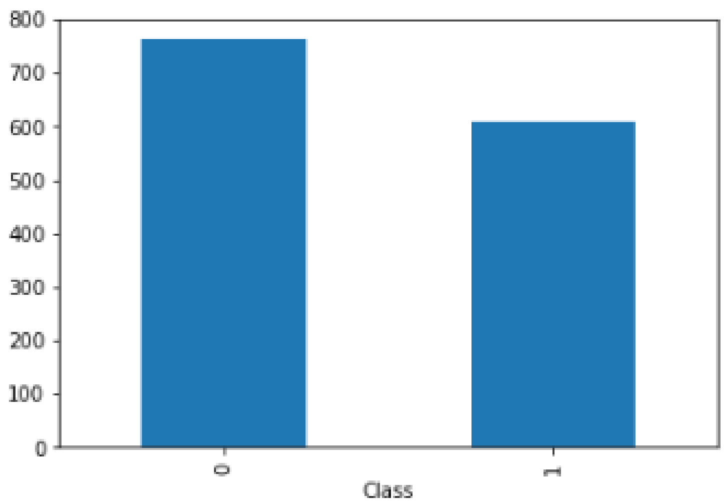
Implementation

```
dataset.head
```

```
<bound method NDFrame.head of       Variance  Skewness  Curtosis  Entropy  Class
0      3.62160   8.66610   -2.8073 -0.44699      0
1      4.54590   8.16740   -2.4586 -1.46210      0
2      3.86600  -2.63830    1.9242  0.10645      0
3      3.45660   9.52280   -4.0112 -3.59440      0
4      0.32924  -4.45520    4.5718 -0.98880      0
...        ...       ...       ...      ...    ...
1367   0.40614   1.34920   -1.4501 -0.55949      1
1368  -1.38870  -4.87730    6.4774  0.34179      1
1369  -3.75030 -13.45860   17.5932 -2.77710      1
1370  -3.56370  -8.38270   12.3930 -1.28230      1
1371  -2.54190  -0.65804    2.6842  1.19520      1

[1372 rows x 5 columns]>
```

```
dataset.groupby('Class').Class.count().plot.bar(ylim=0)
plt.show()
```



Linear SVR

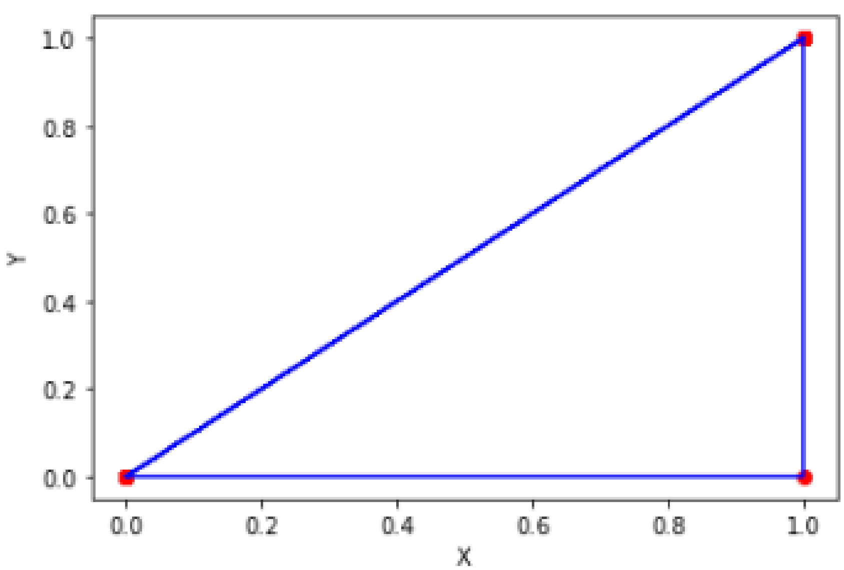
```
Accuracy :  0.9963636363636363
```

```
from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test,y_pred))
print(classification_report(y_test,y_pred))
```

```
[[153   0]
 [  1 121]]
              precision    recall  f1-score   support

           0       0.99      1.00      1.00       153
           1       1.00      0.99      1.00       122

    accuracy                           1.00       275
   macro avg       1.00      1.00      1.00       275
weighted avg       1.00      1.00      1.00       275
```

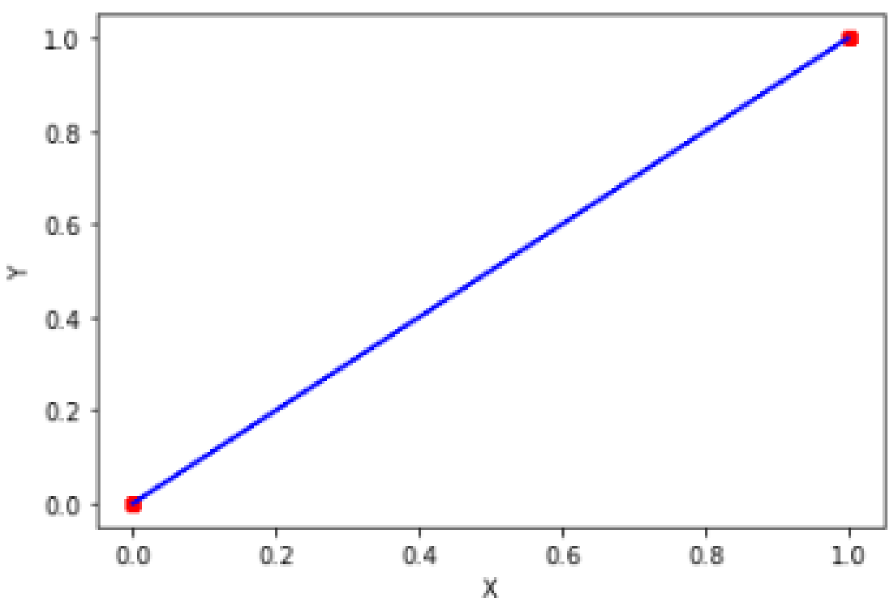


RBF SVR

```
Accuracy :  1.0
[[153   0]
 [  0 122]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       153
           1       1.00      1.00      1.00       122

    accuracy                           1.00       275
   macro avg       1.00      1.00      1.00       275
weighted avg       1.00      1.00      1.00       275
```

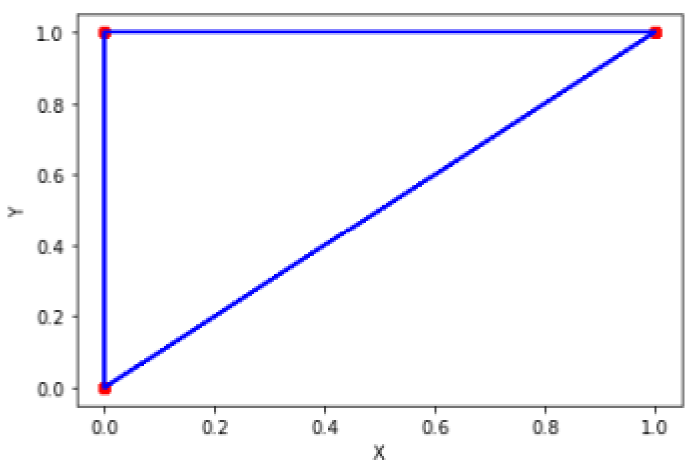


Polynomial SVR

```
Accuracy :  0.9818181818181818
[[148   5]
 [  0 122]]
              precision    recall  f1-score   support

           0       1.00      0.97      0.98       153
           1       0.96      1.00      0.98       122

    accuracy                           0.98       275
   macro avg       0.98      0.98      0.98       275
weighted avg       0.98      0.98      0.98       275
```



Result-

RBF SVR is provide accuracy Compare to Linear and non-linear.