Pattern Recognition and Machine Learning

Indian Institute of Technology, Jodhpur

LAB 4

Report

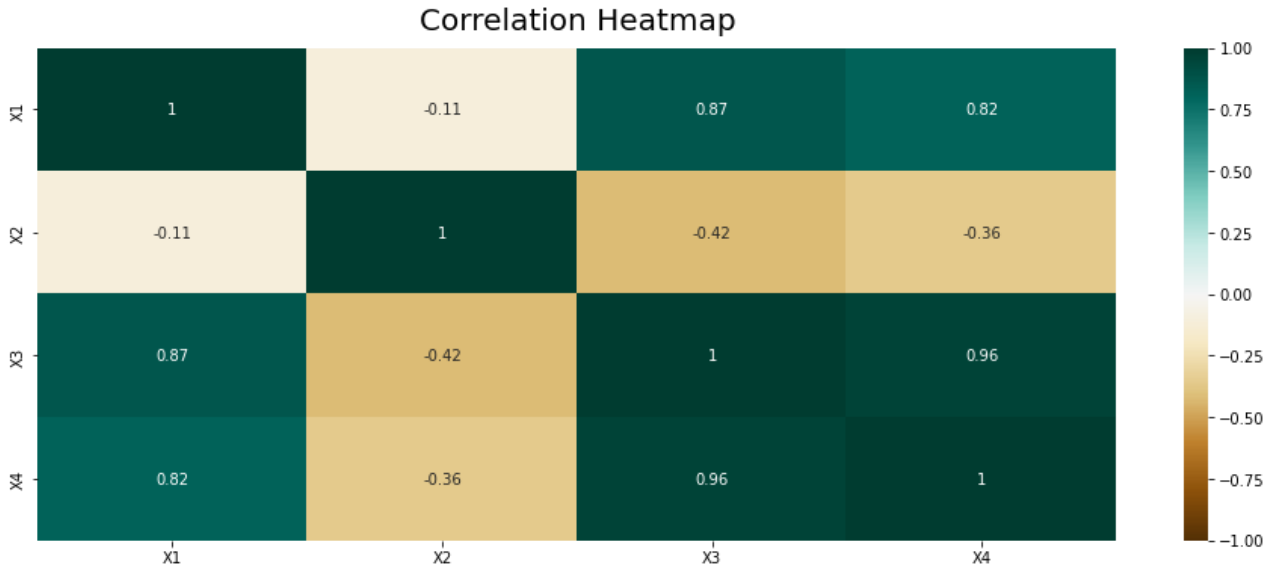
Harshil Kaneria
Department of Computer Science, IIT Jodhpur

Feb 14, 2023

1)

Pre-processing and Exploratory Analysis:

Plotted a **Correlation Heatmap**.



We can infer form the heatmap that the 'X1' and 'X3', 'X1' and 'X4', and 'X3' and 'X4' columns are very highly correlated. High correlation induces error in the prediction so we will remove 'X3' and 'X4' columns.

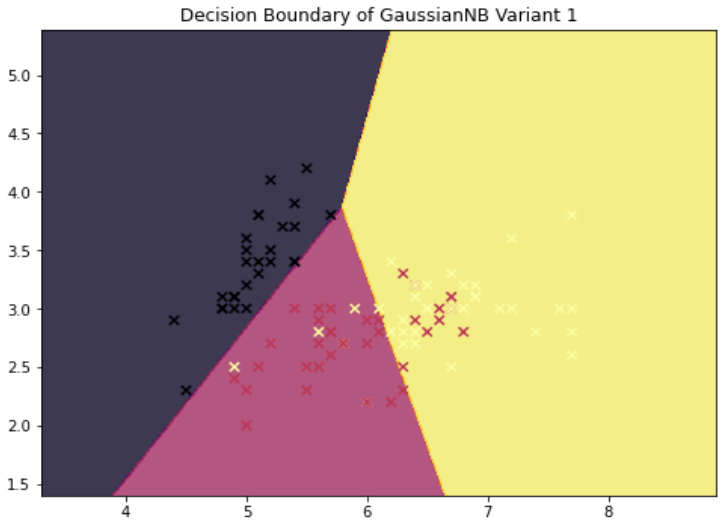
Splitting Into Training and Testing data: Split data into **training set (70%)** and **testing set (30%).**

1) Implemented a Gaussian Naïve Bayes Classifier using class. The three variants are named **1**, **2** and **3** respectively for three cases.

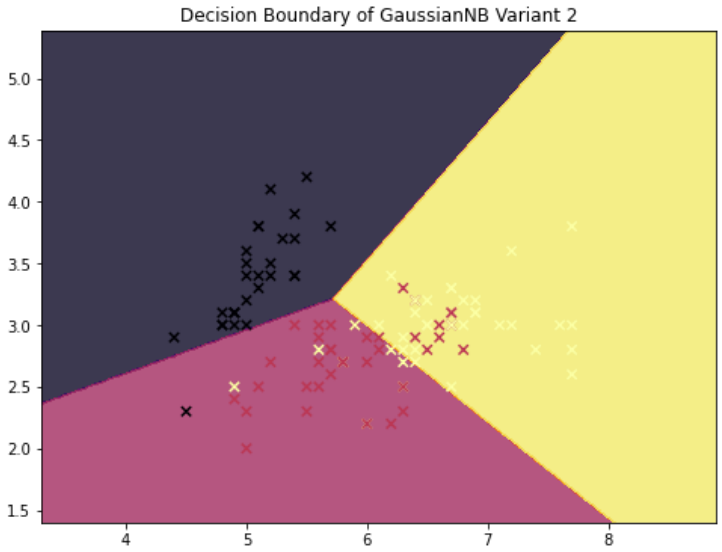
2) The 'Gaussian_NBC' is the name of class. The following are the methods of the class:
   a) Train
   b) Test
   c) Predict – returns predicted array to test function
   d) g(x) (discriminant function) – return the g(x) value for each class
   e) plot_decision_boundary – plots decision boundary
   f) predict array - returns the array of predicted values

3)

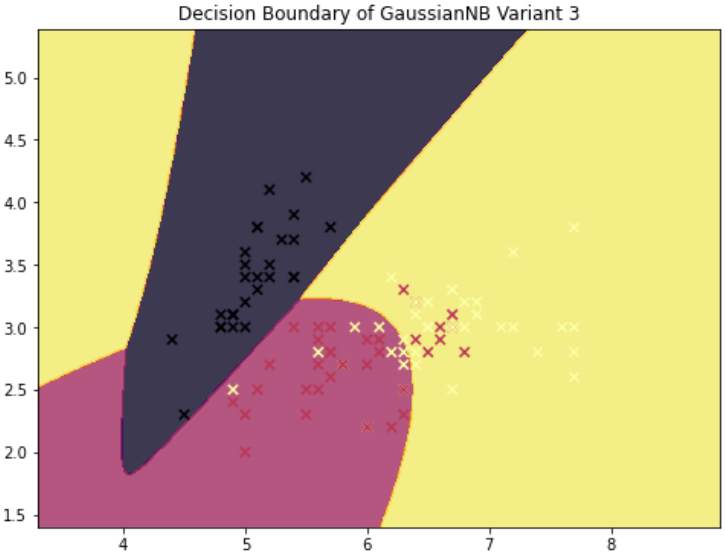
a) Decision Boundary for Variant 1:



b) Decision Boundary for Variant 2:



c) Decision Boundary for Variant 3:

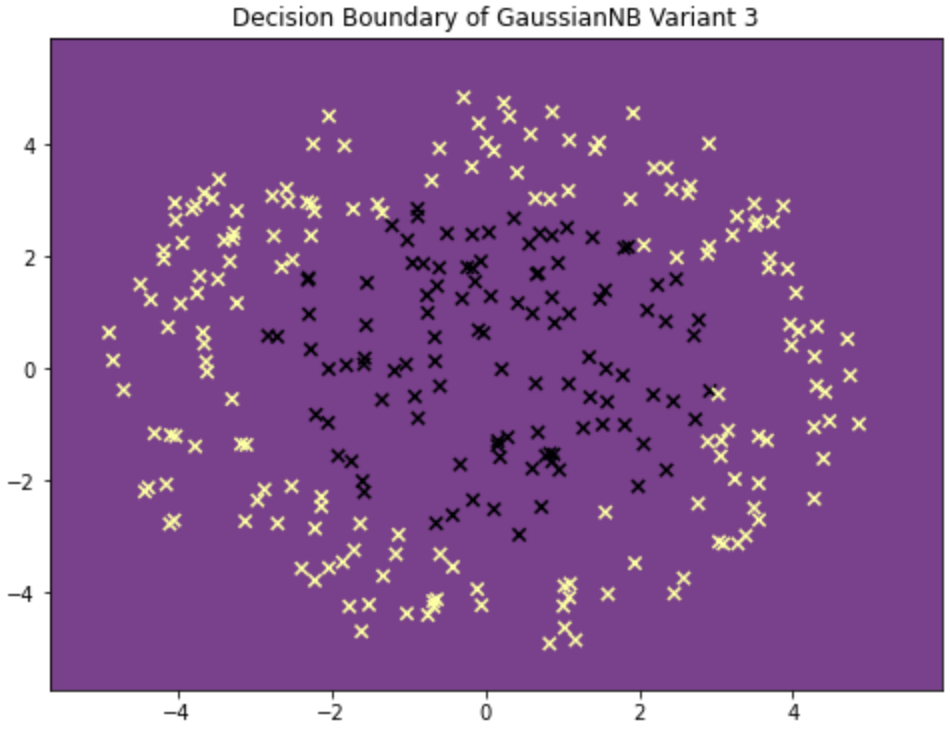


The Decision boundary in **variant 1** and **variant 2** are **linear** and that for **variant 3** is **non-linear**. We can infer from this that if the covariance matrix is same for all class then the decision boundary is linear. If not then it's non-linear.
The Accuracies for each of the three Variants is as follows:
Variant 1: 80%
Variant 2: 80%
Variant 3: 84.44%
We can infer from this that the Variant 3 has performed best among the three variants. All the three variants performed decent on this dataset.

4) The **average accuracy** and **variation in accuracy score** for **5-fold cross validation** of each variant is as follows:
   a) Variant 1: Avg.accuracy = 78%; Variation in accuracy = 0.00471
   b) Variant 2: Avg.accuracy = 80%; Variation in accuracy = 0.00355
   c) Variant 3: Avg.accuracy = 76.67%; Variation in accuracy = 0.00444

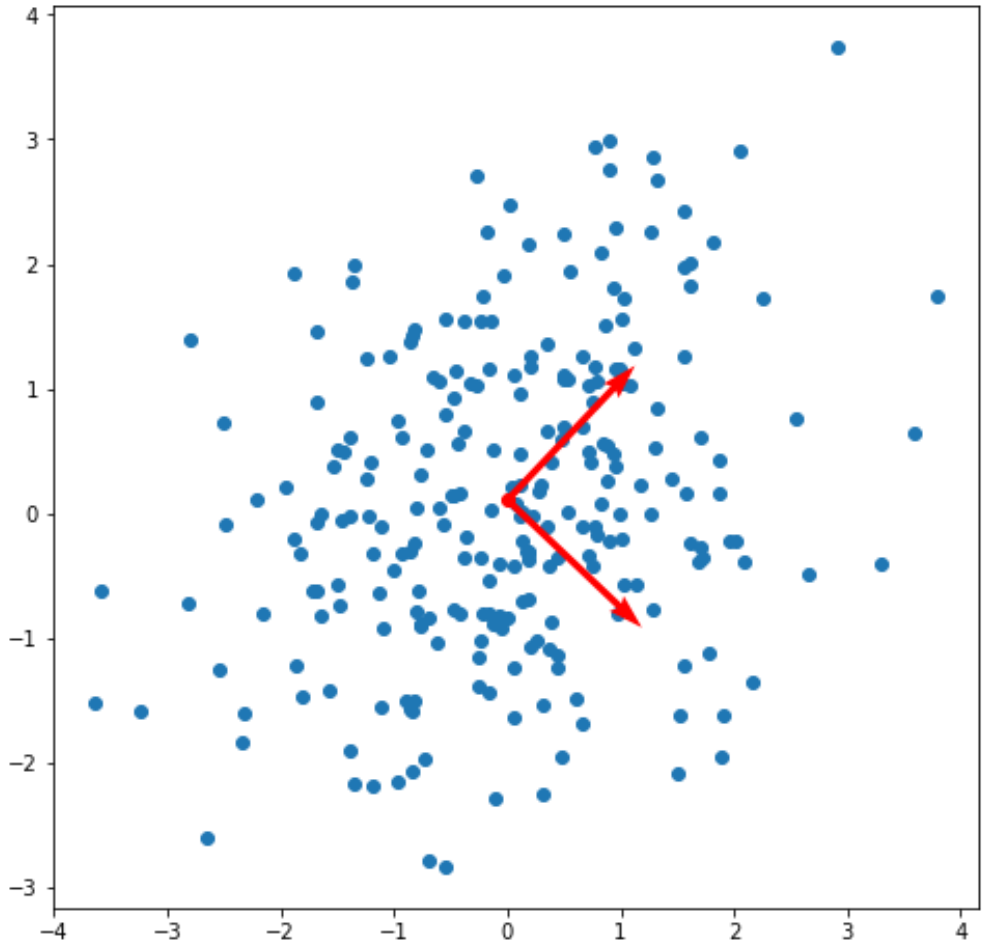
   So from the above scores we can see that variant 2 has least variability so its most generalisable and variant 1 is least generalisable with highest variability.

5) Decision Boundary for Synthetic Dataset created (variant 3):



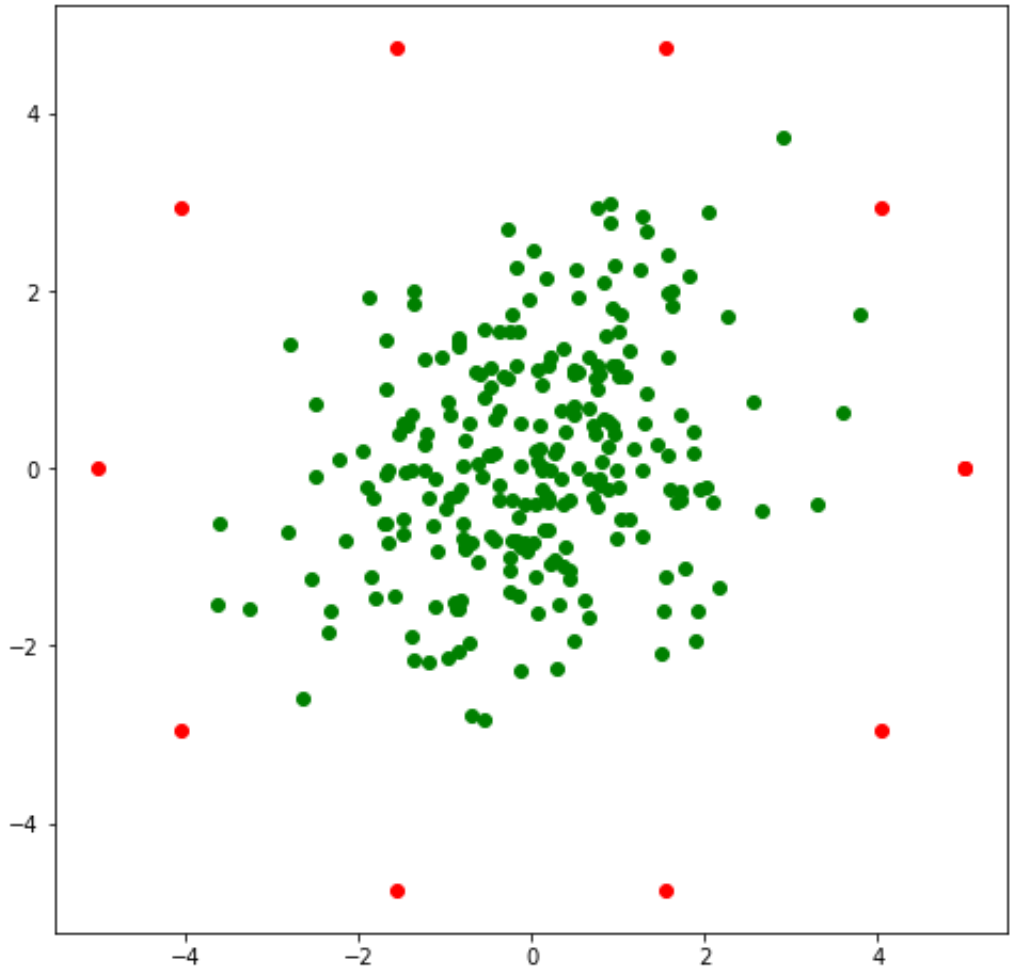
2)

1) The plotted Eigen vector on the scatter plot of dataset points:

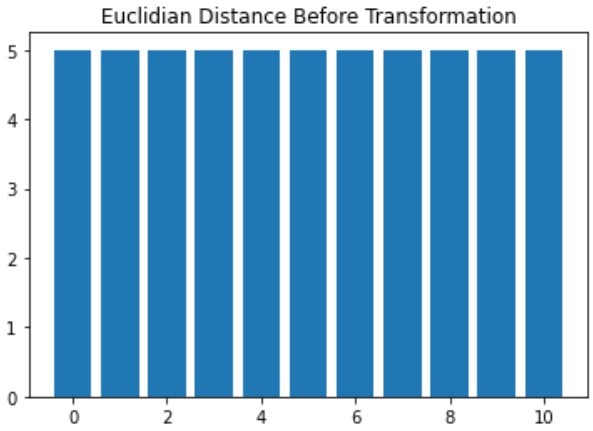


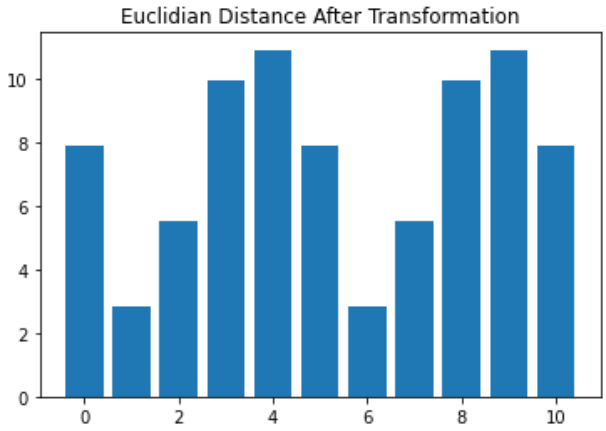
2) The Covariance matrix formed after Transformation is almost an Identity matrix with diagonal element to be exact 1 and other non-diagonal elements to be very small close to 0. This also shows that the transformation has transformed the original dataset X into a dataset with uncorrelated variables with unit variances.
The purpose of transforming the datapoints X to Y is to standardize the variables so that they all have the same variance and are uncorrelated. This is useful in data analysis because it simplifies the interpretation of the data and allows for easier comparison between variables. Additionally, the transformation can be used as a pre-processing step.

3) Plotted the points on circle and coloured it differently:



4) The bar plot of Euclidian distance before and after the prediction:





From the above two graphs we can infer that before the transformation the Euclidian distance was equal for all the points and after transformation Euclidian distance varies widely over a range. But it still has a pattern. Point 4 and 9 are farthest from the mean and point 1 and 6 are both closest. The graph seems periodic with time-period as 5 units.