IOT Project 4: Clustering Techniques

Ojas Barve - ovbarve@ncsu.edu

Task: SVM classification

This task involved the classification of data using the support vector classifier.

Figure 1. Below shows visualisation of the dataset provided containing two classes labelled in different colors.

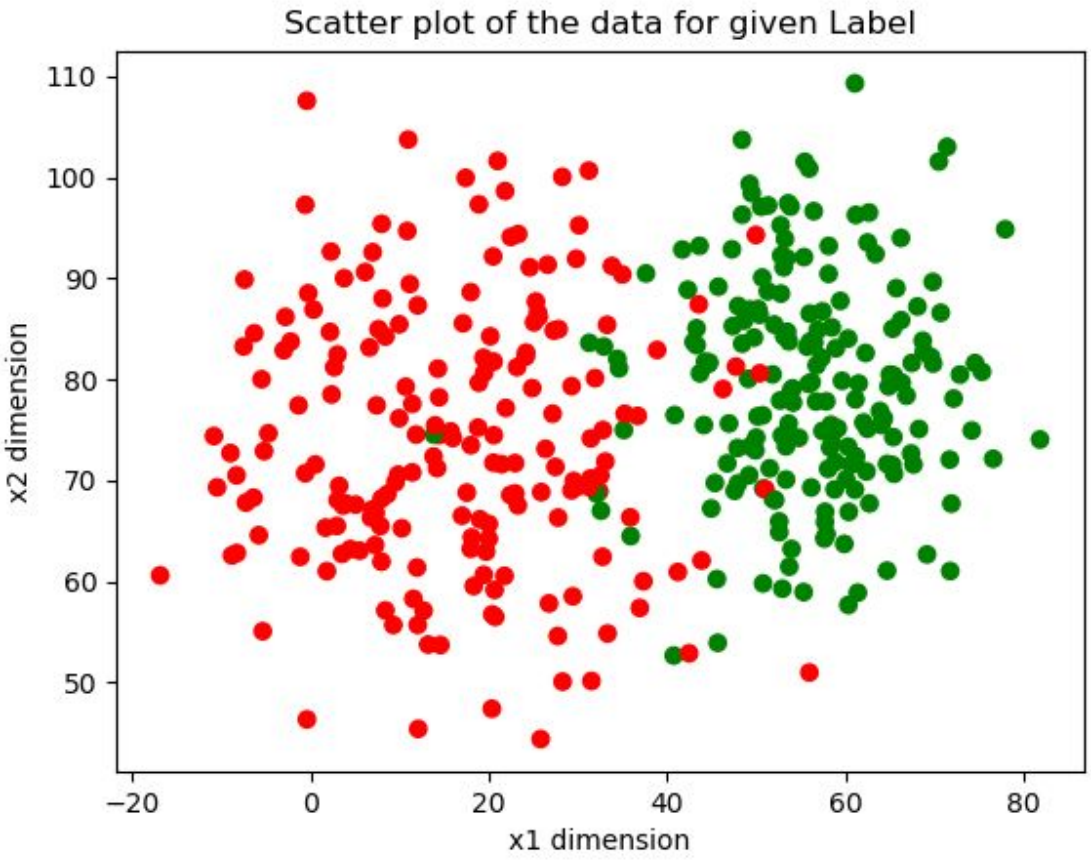

Figure 1. Scatter plot of given data.

Based on the dataset scatter diagram plotted we can conclude that the dataset in the given dimension is not linearly separable. Hence we need to use a kernel function to

change the dimensionality and take the data to a high dimensional space where the two clusters will be linearly separable.

For this purpose we use the Radial Basis Function as a kernel.

We apply SVC technique using this classification. We divide the dataset into 5 folds to allow cross validation.

Below is the cost function for soft margin svm:

$$f(\mathbf{w}) = \| \mathbf{w} \| / 2 + C * (\sum^{N}_{i=1} \xi)^{k}$$

And the RBF kernel is

$$K_{RBF}(x, x') = exp[- \gamma * | x - x' |^{2}]$$

In the above 2 functions we optimise the value of the C and γ parameters.

Task: Initial Search for C and γ :

In this case we are interested in for a large range of value of C and γ so that we can pick a combination where we achieve a relatively better accuracy to run a finer search latter on.

Our initial 3 dimensional plots for scatter plots are given as follows. We can provide heat charts to be be able to better understand the results of the grid search over the two hyperparameters.

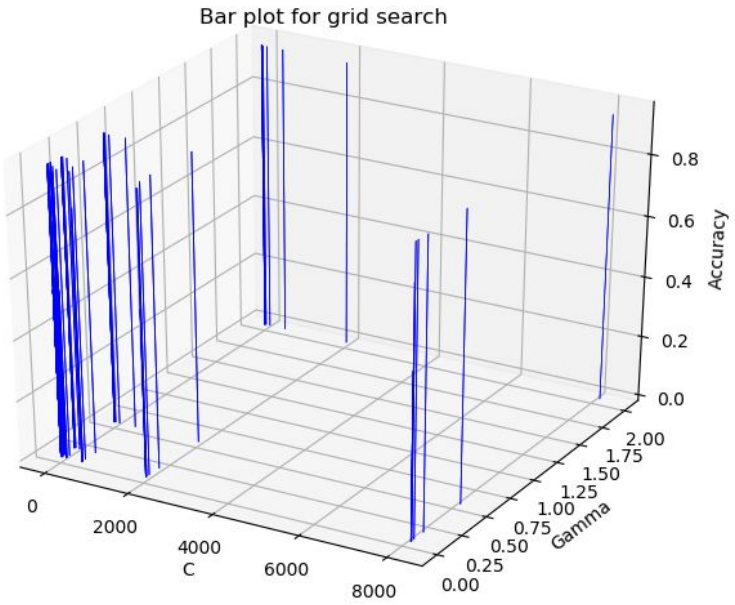

Figure 2. 3d bar chart

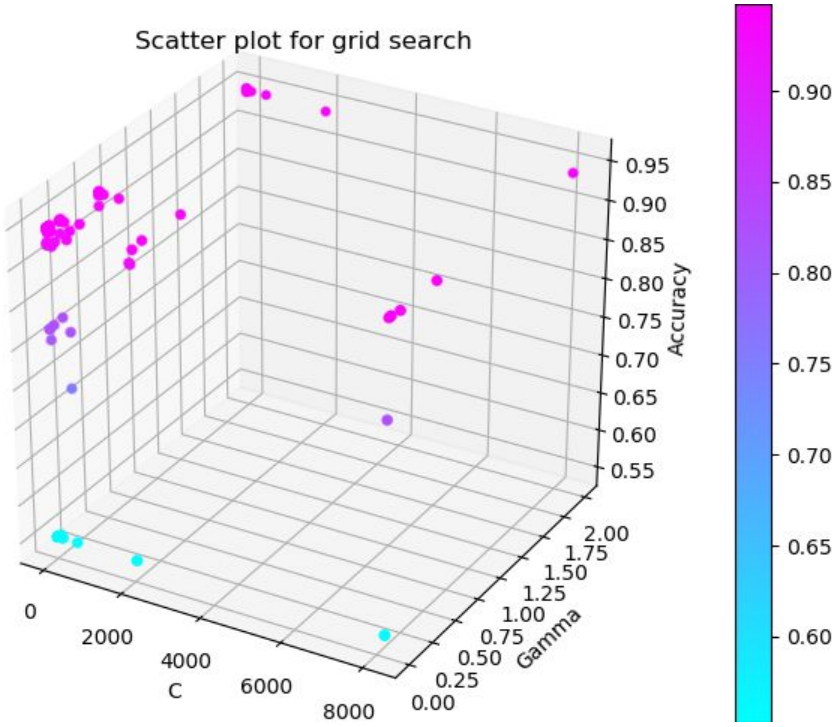

Figure 3 . 3d scatter plot

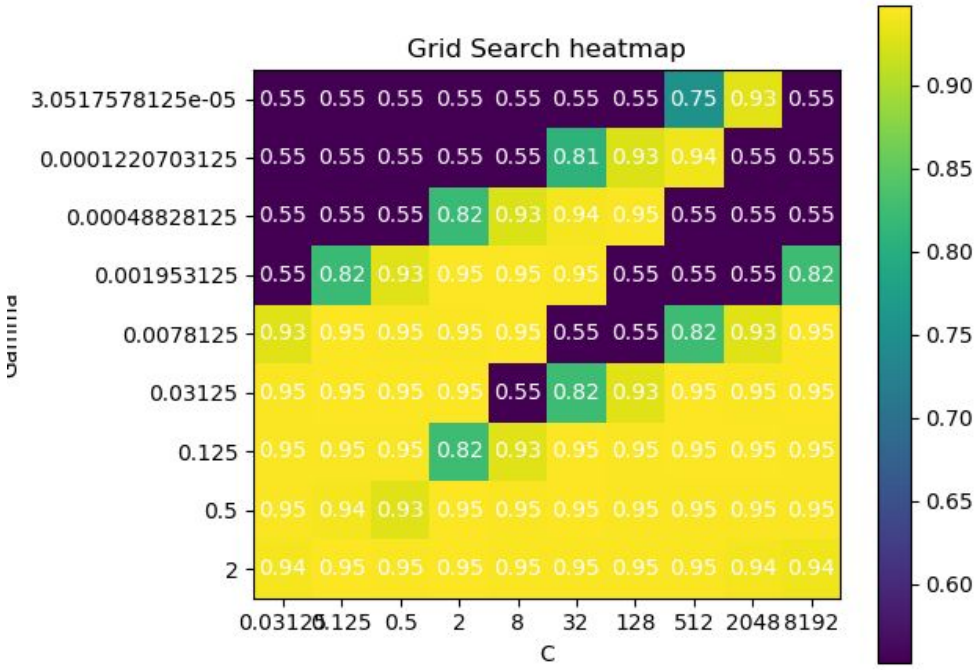

Figure 4. Heatmap of grid search

BEST values of C and Gamma over initial search over wider range was found for a value of

C = 2

Gamma = 0.125

Accuracy = 0.9477785372522215

Task : Finer search over the values obtained in the above iteration:

In this case we are interested in a finer range of value of C and γ from the values that we picked up from the first iteration of grid search we applied in the previous task. Our aim is to improve on the accuracy achieved over the first grid search iteration.

Below is shown results obtained.

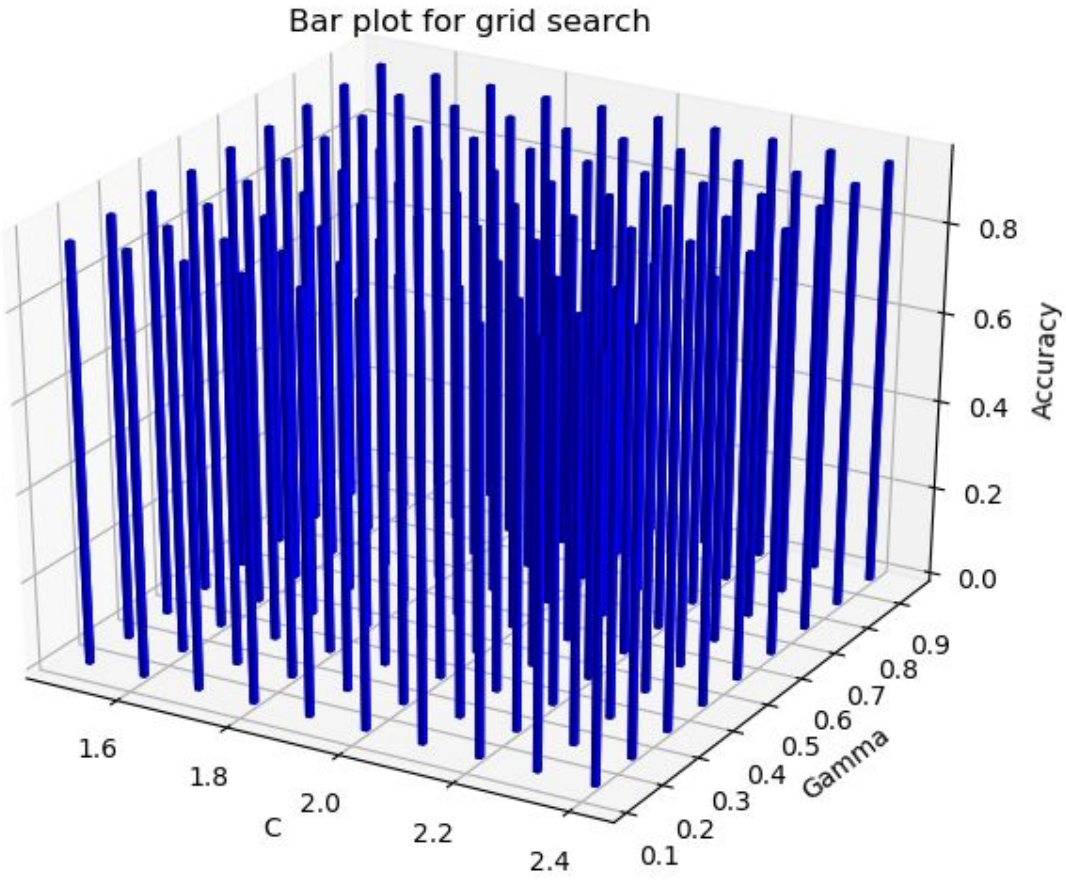

Figure 5. 3d bar chart

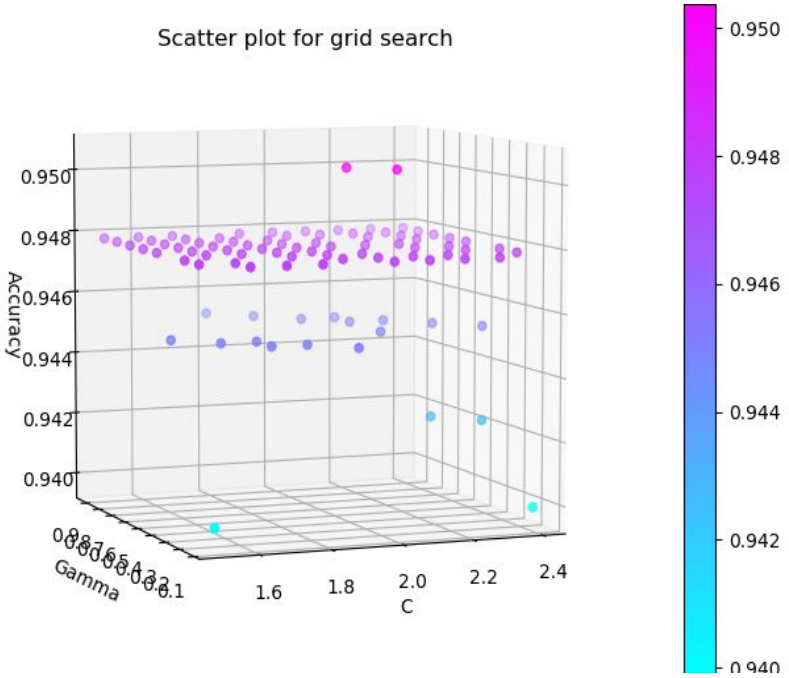

Figure 6. 3d scatter plot

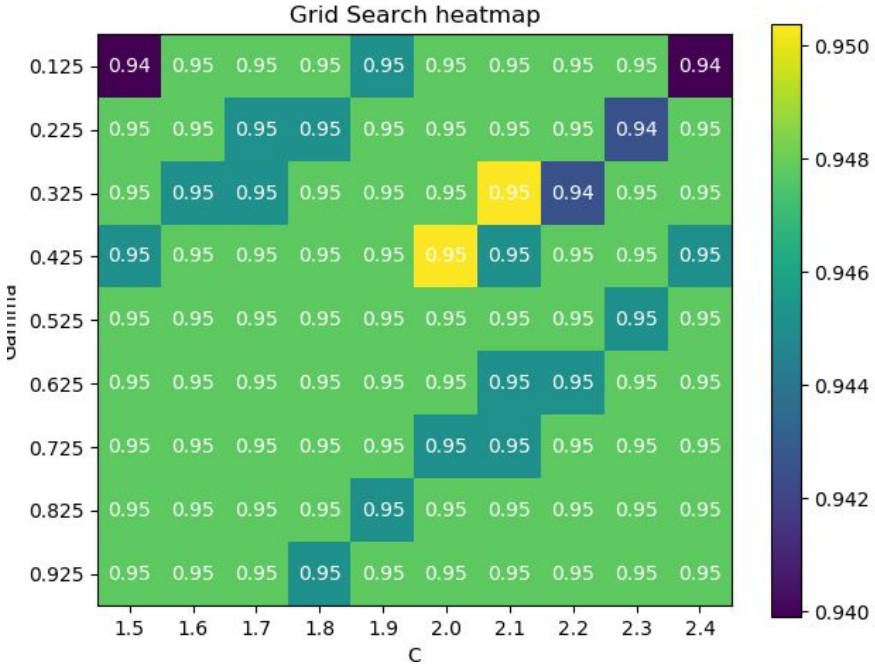

Figure 7. Heat map of grid search on finer range

BEST values of C and Gamma after fine tuning the range are as follows;

C = 1.7

Gamma = 0.925

Accuracy = 0.9503759398496241

Task : Final Comment

The grid search was performed for an initial values of the range being C = 2 ^–5 , 2^ –3 , …, 2 ^15 , and γ = 2^-15, 2^-13, … 2^3.

After fine tuning we change the range to

C: [1.5, 1.6, 1.7, 1.8, 1.9, 2.0, 2.1, 2.2, 2.3, 2.4]

Gamma: [0.125, 0.225, 0.325, 0.425, 0.525, 0.625, 0.725, 0.825, 0.925]

The accuracy improves from 94.77% to 95.03%. The algorithm was applied successfully and the hyperparameters were optimized.