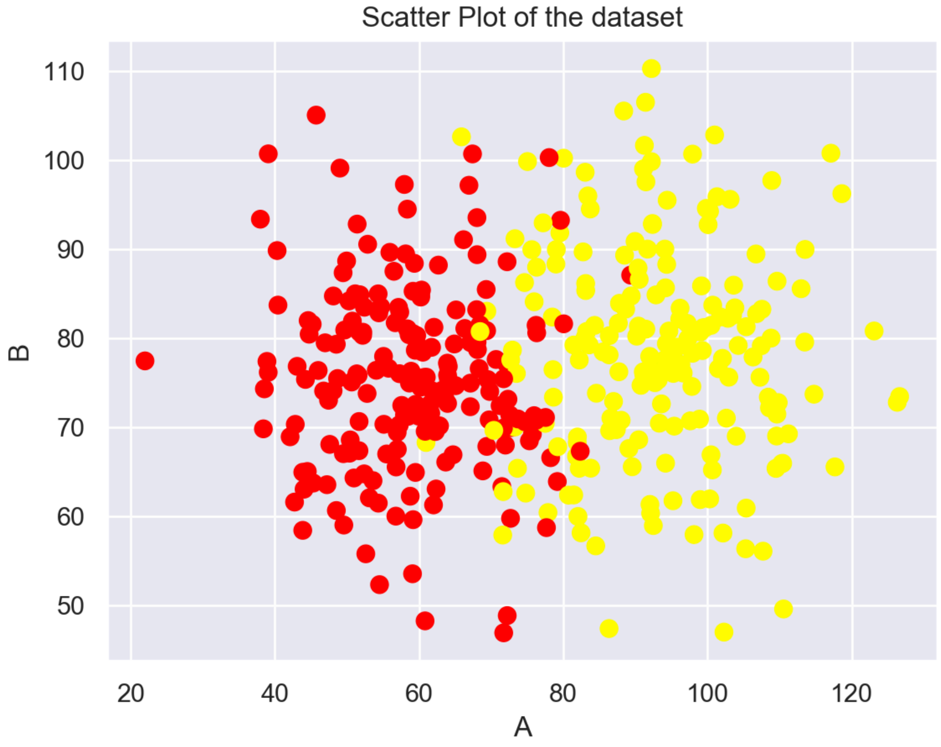
# ECE 592: IoT Analytics

## PROJECT 5: Support Vector Machines

**STUDENT ID: 200203773**
**STUDENT NAME: ESHAAN VENKAT KIRPAL**

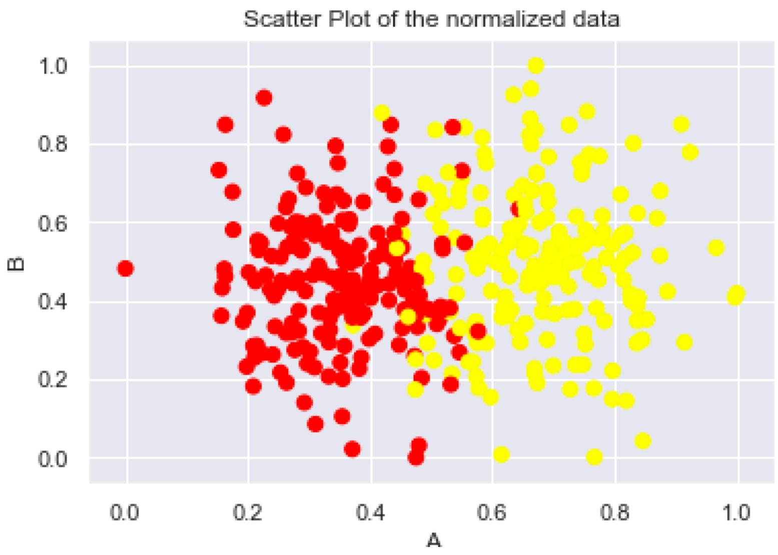
**SUMMARY OF THE INPUT DATA:**

The input data has 384 observations and two features, say A and B. The observations are classified into two classes, y=1 (in red) or y=2 (in blue). As can be seen from the scatter plot below, the data is not linearly separable.



Thus, to find a classifier using the Support Vector Machine model, we will project the data into a higher dimension such that a linear separator would be sufficient. For this we will use the RBF kernel.

Though first, we will normalize the data. Below is the scatter plot of the normalized data.



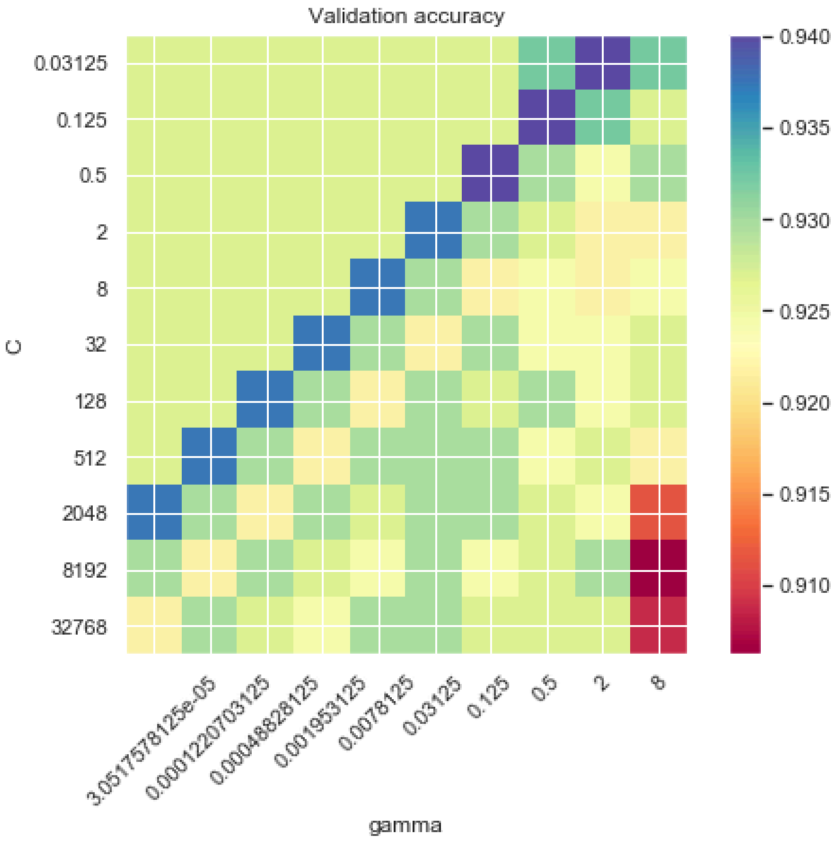
### COMPUTING BEST VALUE OF C AND GAMMA

We use grid search cross-validation to explore combinations of parameters. Here we will adjust C (which controls the margin hardness) and gamma (which controls the size of the radial basis function kernel), and determine the best model.

The values of C and gamma were varied in the range shown below:
C = $2^{-5}$, $2^{-3}$, ..., $2^{15}$
γ = $2^{-15}$, $2^{-13}$, ..., $2^{3}$

**The best parameters are {'C': 0.03125, 'gamma': 2} with an accuracy of 94.01%.**

The best parameter can be observed from the heatmap of the classifier’s cross-validation accuracy as a function of C and gamma.

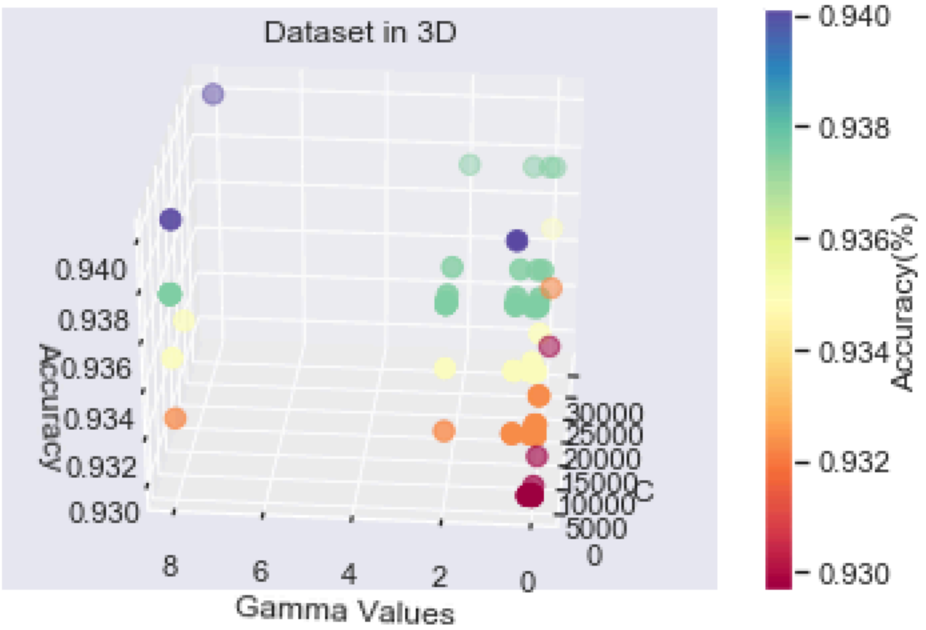


From the heatmap we observe that the accuracy is not near optimal when the gamma values are less than 0.5 and at the same time C is also small. Also, accuracy is bad when C is too high.

For intermediate values, we can see in the heatmap that good models can be found on a diagonal of C and gamma. Smooth models (lower gamma values) can be made more complex by increasing the importance of classifying each point correctly (larger C values) hence the diagonal of good performing models.

We get good accuracy values along the diagonal where the parameter values with high accuracy are the ones in the top right corner in dark blue color. Since the best parameters lies on the boundaries of the grid, we will extend in that direction in a subsequent finer search.

Next, we visualize the accuracy against the hyperparameter grid using a 3D plot. On the x and y axis are the different values of C and γ while the accuracy is plotted on the z-axis.



The regions of varying accuracy have been differentiated using colors. It is clear from the above plot that the maximum accuracy is obtained with gamma=2 and C=0.03125, i.e. 94.01%.

There are two other points in the scatter plot that are in dark blue and have comparable accuracies to our best value.

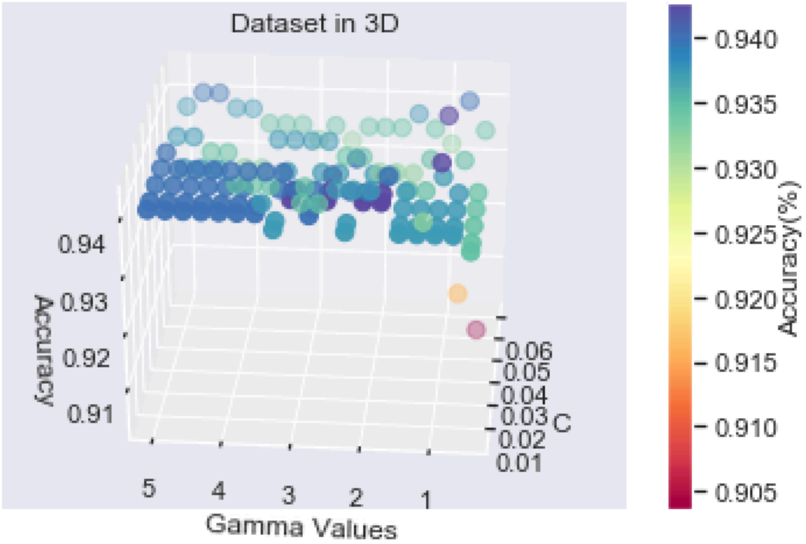
We next perform fine grid search for the two parameters C and gamma to get the best model.
A finer grid search was performed around all the three points and the best accuracy result was reported. The results for the fine search for the other two points has not been reported here though its results can be obtained from the code provided with this report.

Here we only show results from the fine search performed around the region where the grid search returned the maximum accuracy. The following range of values were used for the fine search:

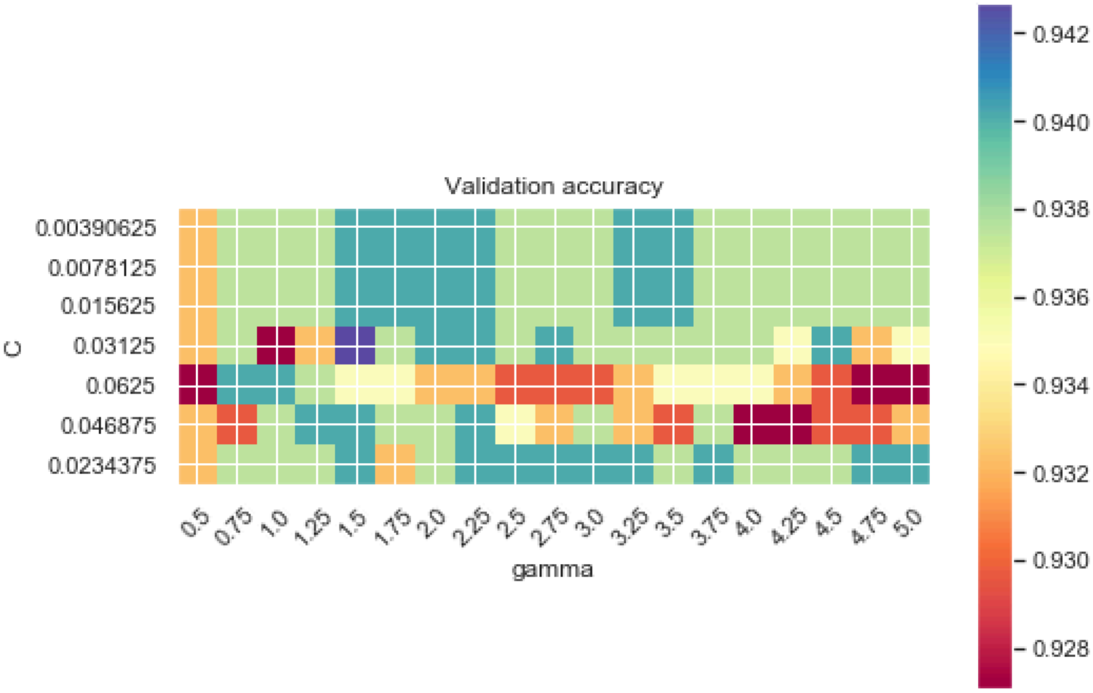
C = [0.00390625, 0.0078125, 0.015625, 0.03125, 0.0625, 0.046875, 0.0234375]
γ = [0.5 , 0.75, 1. , 1.25, 1.5 , 1.75, 2. , 2.25, 2.5 , 2.75, 3. ,
3.25, 3.5 , 3.75, 4. , 4.25, 4.5 , 4.75, 5. ]

**The best parameters are {'C': 0.03125, ‘γ’: 1.5} with an accuracy of 94.27%.**

Regions with same accuracy are plotted with similar color in the 3D scatter plot above. We have far more points in this plot since we used a bigger range of values for the hyperparameters C and gamma during the fine grid search.



To better identify the point of maximum accuracy, we can look at the heatmap below. The dark blue point in the left middle of this map is where we have the maximum accuracy, i.e. when C=0.03125 and gamma=1.5. Note the best model obtained using grid search was with gamma=2.



One can observe that for values of gamma between 1.5 to 2.25 we get comparably equally performing models as we vary C. For low and high values of gamma, we get models with lower accuracies, though these are not very low accuracy values.

**CONCLUSION:**

This report analyses the effect of the parameters gamma and C of the Radial Basis Function (RBF) kernel SVM.
Intuitively, the gamma parameter defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'. The gamma parameters can be seen as the inverse of the radius of influence of samples selected by the model as support vectors.

On the other hand, the C parameter trades off correct classification of training examples against maximization of the decision function's margin. For larger values of C, a smaller margin will be accepted if the decision function is better at classifying all training points correctly. A lower C will encourage a larger margin, therefore a simpler decision function, at the cost of training accuracy. In other words``C`` behaves as a regularization parameter in the SVM.

From the results of our grid and fine search with cross validation for RBF kernel SVM, we conclude that the best accuracy of **94.27%** is obtained for the values of **C = 0.03125 and γ = 1.5**. **Also, we learn that using finer search helped us in finding a model with higher accuracy.**

The value of C and γ for which the highest accuracy was obtained was selected as our final model and it serves as the best fit for the classification of the given dataset.