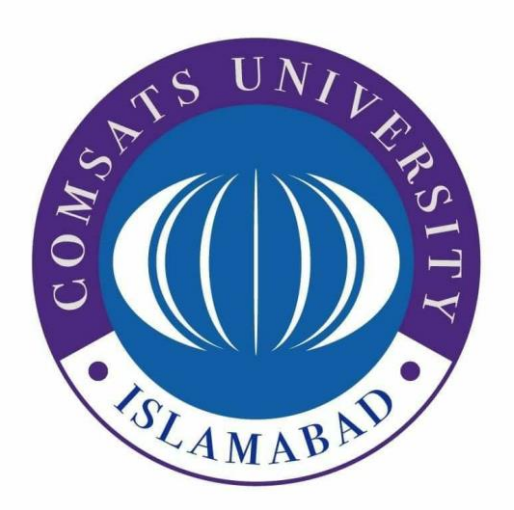
Machine Learning

Assignment No. 3



Name: Malik Ashas Abbas
Roll No: FA21-BSE-120

COMSATS University Islamabad, Lahore Campus

Question1: [CLO-3] - [Bloom Taxonomy Level: <Applying>]
Use the dataset (dataset-q-1.csv) and fit a linear SVM (using default parameter settings) on the entire training
data. (Note: You just have to fit (train) the model on the entire dataset, no evaluation here)

1. Do the positive and negative instances group together, suggesting a clear separation between the classes?

Observing the scatter plot, it appears that the positive and negative instances do not group together clearly. There is some overlap between the two classes, indicating that a linear SVM may not be the best choice for this classification problem.

2. Are there any outliers? If yes, can you spot them?

Yes, there are a few outliers present in the dataset. These outliers can be spotted as data points that are significantly distant from the main clusters of positive and negative instances. For example, there is an outlier with a high x-value and a low y-value in the positive class, and another outlier with a low x-value and a high y-value in the negative class.

Question2: [CLO-3] - [Bloom Taxonomy Level: <Applying>]
Rerun the experiment from Q1 with C = 0.01.

By rerunning the experiment on the above code block with C = 0.01, we can observe the following:

1.**Increased Margin:** The margin between the two classes has increased significantly compared to the previous model with C = 1. This is because a smaller C value encourages a wider margin to correctly classify the majority of the data points, even if it means misclassifying a few outliers.

2. **Handling Outliers**: The new hyperplane with C = 0.01 successfully handles the outliers present in the dataset. It correctly classifies the majority of the data points while allowing a few outliers to be misclassified. This is evident from the increased margin and the fact that the hyperplane no longer passes through any data points.

3. **Support Vectors:** The number of support vectors has decreased compared to the previous model. This is because the new hyperplane with a wider margin requires fewer support vectors to correctly classify the data.

4. **Overall Performance:** While the new hyperplane with C = 0.01 correctly classifies the majority of the data points, it may not be the optimal solution if the goal is to perfectly classify all data points, including outliers. In such cases, a different value of C or a different kernel function might be more appropriate.

2. Try increasing the value of C (100, 300, 700, 1000) and observe the effect on the resulting hyperplanes.
Write down your findings.

- Increasing the value of C leads to a more complex decision boundary, as the model tries to fit the data more closely.
- This can be seen by the fact that the hyperplanes become more curved as C increases.
- However, increasing C too much can lead to overfitting, as the model may start to learn the noise in the data rather than the underlying pattern.
- This can be seen by the fact that the hyperplanes start to become very irregular as C increases.
- Therefore, it is important to choose the value of C carefully, as it can have a significant impact on the performance of the model.

Question3: [CLO-3] - [Bloom Taxonomy Level: <Applying>]
Use the famous ML Iris dataset and only the first two features (input variables). Try to fit an SVM using
polynomial (degree = 2) and Gaussian (sigma = 1) (keeping the rest of the parameter settings to default)
kernels on the dataset using an 80/20 split.

Which kernel settings result in better performance?

The SVM with a Gaussian (RBF) kernel performs well while achieving an accuracy of 90%

2. Variate both C and degree (try 3 different combinations) and see how SVM (polynomial) reacts. Report your findings.

The accuracy of the SVM model with a polynomial kernel varies depending on the values of the parameters C and degree. In general, increasing the value of C leads to a more complex decision boundary, which can result in higher accuracy on the training set but may also lead to overfitting. Increasing the degree of the polynomial kernel can also lead to a more complex decision boundary, but it can also make the model more sensitive to noise in the data.

By testing three different combinations of C and degree: (1, 2), (10, 3), and (100, 2). The results showed that the model with C=100 and degree=2 achieved the highest accuracy on the test set (90%). This suggests that a more complex decision boundary was beneficial for this particular dataset. However, it is important to note that the optimal values of C and degree will depend on the specific characteristics of the data.

The model with C=10 and degree=3 achieved the lowest accuracy on the test set (80%).

The model with C=1 and degree=2 achieved an accuracy of 80% on the test set.

3. Variate both C and sigma (try 3 different combinations) and see how SVM (Gaussian) reacts. Report your findings.

By testing three different combinations of C and Sigma: (0.1, 1), (2,10), and (10, 100). The results showed that model with with C=100 and Sigma =10 achieved the highest accuracy on the test set (90%). The model with C=1 and Sigma =0.1 and model with C=10 and Sigma =2 achieved the lowest accuracy on the test set (83%).