6th Sep 2022

Attendance: 10%, Continuous evaluation: 70%, Viva-20%

Assignment No. 4

i. Download Titanic Dataset (https://www.kaggle.com/heptapod/titanic/version/1#) and do initial pre-processing including normalization, na or zero column handling, train test split, and others (Write an explanation of each in the report).

ii. Train the SVM using the below kernels with parameters, present the support vectors in the table of the comparison of the model along with accuracy.
   a. Linear
   b. Polynomial: where degree d is set to 2, 3 and 5
   c. RBF
   d. Sigmoid

iii. Take only two features from the dataset and train the models with the same parameters and plot the graphs to show the boundaries. Also, create a custom kernel function of your own using a mathematical function for suggestion Lograthmic or Tangent function.

iv. For RBF kernel vary the control parameter C with a binary search technique to reach an optimal C value. Plot the graph for validation accuracy. Using this, mention the situation of overfitting and underfitting. Set Gamma to 0.5. Create a function for the whole process. [Maximum 20 runs]

v. Using the above-created function now varies the Gamma parameter with the same binary search techniques as above for the C value which has maximum validation accuracy. Explain, whether the above calculated maximum test accuracy is the optimal test accuracy or there can be a better value of C and Gamma.

vi. Download the Forest Cover Type dataset (https://www.kaggle.com/uciml/forest-cover-type-dataset) and preprocess the dummy variables to create training, test, and development set. Reduce the train data size if the system is unable to process the whole dataset.

vii. Train the one vs rest and one-vs-one SVM model on the above dataset for multiclass classification. Plot and Analyze the Confusion matrix for the above models. Show the accuracy in the graph. State the difference of the two approaches using the model parameters.

Submit a report with results.