Machine Learning Laboratory – Assignment V

Using built in functions to implement Support Vector Machines (SVMs) with different kernel functions – Linear and non-linear (Gaussian (RBF), and polynomial) – and analyse the impact of these kernels on classification performance and decision boundaries when applied on Spambase dataset to classify emails as either spam or non-spam based on various features provided in the dataset.

The Spambase dataset contains 4601 email messages, out of which 1813 are labeled as spam. The dataset consists of 57 numeric features including the frequency of various words, characters, and other attributes in the emails and the last column denotes a binary label, 1 for spam, 0 for not spam. The run-length attributes (55-57) measure the length of sequences of consecutive capital letters.

1) Load the dataset and do relevant pre-processing.

2) Train a Linear SVM (using Sci-kit libraries)

3) Train Non-Linear SVMs with Different Kernels (using Sci-kit libraries)

a) Gaussian/RBF(Radial Basis Function)Kernel:

b) Polynomial Kernel: Start with degree 3 and then experiment with different degrees(e.g., 2,4)

4) Evaluate the Models - Calculate the accuracy and plot confusion matrices for each model.

5) Plot the Decision Boundaries and visualize how each SVM separates the data. (Use a 2D representation of the dataset (reduce dimensionality if necessary using PCA)

6) Compare the performance of SVMs based on all the kernels and find out which kernel performs better on the given dataset.