CS312:Artificial Intelligence Laboratory

Lab 5 Report Group 24

Utkarsh Prakash - 180030042 Manjeet Kapil - 180010021

1. Introduction

In this assignment, we were expected to use a SVM to classify emails into spam and non-spam categories and report the classification accuracy for various SVM parameters and kernel functions.

2. Library Used

Scikit-Learn -:Scikit-learn is largely written in Python, and uses numpy extensively for high-performance linear algebra and array operations. Furthermore, some core algorithms are written in Cython to improve performance. Support vector machines are implemented by a Cython wrapper around LIBSVM; logistic regression and linear support vector machines by a similar wrapper around LIBLINEAR.

Pandas -: Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. Pandas is mainly used for data analysis. Pandas allows importing data from various file formats such as comma-separated values, JSON, SQL, Microsoft Excel.

3. Details of SVM implementation in Scikit Learn

The Linear Kernel for SVM is implemented in svm.LinearSVC in Scikit-Learn. The model by default applies L2 regularization, and the strength of regularization is controlled by the parameter C. The higher values of C correspond to less regularization. In other words, when we use a high value for the parameter C, LinearSVC try to fit the training set as best as possible, while with lower values of the parameter C, the model puts more emphasis on finding a coefficient vector (w) that is close to zero. The default value of C is 1.

The polynomial Kernel for SVM is implemented in svm.SVC in Scikit-Learn. The ‘kernel’ argument specifies the kernel type to be used in the algorithm. If the kernel is defined to be

'poly' then the algorithm uses a polynomial kernel. The 'degree' argument defines the degree of polynomial to be used in the algorithm if the kernel specified is polynomial.

The Gaussian Kernel for SVM is implemented in svm.SVC in Scikit-Learn. The 'kernel' argument specifies the kernel type to be used in the algorithm. If the kernel is defined to be 'rbf' then the algorithm uses a Gaussian kernel or Radial Basis Function (RBF). The RBF as implemented in Scikit Learn:

$$k_{rbf}(x_1, x_2) = \exp(\gamma||x_1 - x_2||^2)$$

The 'gamma' argument to SVC provides the gamma parameter to the RBF.

4. Experimental Results:

- **Linear Kernel**

C-Value	Training Accuracy	Testing Accuracy
0.01	92.42%	91.81%
0.03	93.26%	91.74%
0.1	93.66%	92.32%
0.3	93.78%	92.54%
1.0	93.91%	92.54%

- **Quadratic Kernel**

C-Value	Training Accuracy	Testing Accuracy
0.01	62.91%	62.99%
0.03	67.14%	66.90%
0.1	72.82%	72.12%
0.3	79.53%	78.49%
1.0	86.05%	83.41%

- **Gaussian Kernel**

C-Value	Training Accuracy	Testing Accuracy
0.01	70.15%	71.03%
0.03	88.47%	87.25%
0.1	91.39%	90.44%
0.3	93.47%	91.67%
1.0	94.81%	92.90%

- For a linear kernel the best value of C is 1 with training accuracy 93.91% and testing accuracy 92.54%.
- For quadratic kernels the best value of C is 1 with training accuracy 86.05% and testing accuracy 83.41%.
- For a RBF kernel the best value of C is 1 with training accuracy 94.81% and testing accuracy 92.90%.