Artificial Intelligence Lab Assignment 6

Group Number: 27

Pavan Kumar V Patil 200030041

Karthik J Ponarkar 200010022

March 16, 2022

Contents

1 Problem Description

- According to the problem statement, we have to write a program which has to classify a random email into either spam or not spam.
- Basically, we have to build a spam email classifier using support vector machines.
- After using **SVM** to classify the data set of emails into spam or not spam categories, we have to observe and report the accuracy of our classification in terms of various **SVM** parameters and kernel functions.
- Since, this strategy comes under Machine Learning, the machine itself optimises the value of SVM parameters and kernel functions from the experience E gained during the training phase from the training set data such that the accuracy is the maximum.

2 Requisite inputs for email classifier program

- A **csv** file should be given as the input to this program which contains necessary characteristics of the emails such as number of capital letter alphabets, presence of certain key words etc. It also contains the information about the true classification of the mail.
- We randomly use 70% of the data as the training data to train our model and remaining 30% as the test data to test the accuracy of our classification.
- The input file contains 4601 lines (rows) which is the number of emails given as the input.

3 Libraries and packages used in the program

1. import pandas
2. from sklearn import svm
3. from sklearn.model_selection import train_test_split

3.1 Pandas

- Pandas is an open source Python package that is most widely used for data science and data analysis and machine learning tasks. It is built on top of another base package named **numpy**, which provides support for multi-dimensional arrays.
- Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis/manipulation tool available in any language. It is already well on its way toward this goal.

3.2 Scikit-Learn Package

- Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python.
- It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

- It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

4 Brief description of Kernels

- A kernel is a function used in SVM for helping to solve problems. They provide shortcuts to avoid complex calculations. The amazing thing about kernel is that we can go to higher dimensions and perform smooth calculations with the help of it. We can go up to an infinite number of dimensions using kernels.
- In machine learning, a kernel refers to a method that allows us to apply linear classifiers to nonlinear problems by mapping non-linear data into a higher-dimensional space without the need to visit or understand that higher-dimensional space.

Note : The underlying sections briefly explain about the different kernels used in the code for implementation of SVM for the given data set.

4.1 Linear Kernel

- Linear Kernel is used when the data is linearly separable, that is, it can be separated using a single line. It is one of the most common kernels to be used.
- It is mostly used when there are a large number of Features in a particular data set.
- The linear kernel is the simplest kernel function. It is given by the inner product $< x, y >$ plus an optional constant C.

4.2 Polynomial Kernel

- In machine learning, the polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.
- In this problem, we use a special case of polynomial kernel which is quadratic kernel which is of degree 2.
- Intuitively, the polynomial kernel looks not only at the given features of input samples to determine their similarity, but also combinations of these. In the context of regression analysis, such combinations are known as interaction features.

4.3 RBF (Gaussian Radial Basis Function) Kernel

- In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification.
- RBF Kernel is popular because of its similarity to K-Nearest Neighborhood Algorithm.
- It has the advantages of K-NN and overcomes the space complexity problem as RBF Kernel Support Vector Machines just needs to store the support vectors during training and not the entire dataset.

5 Observation and Analysis

C	Kernel					
	Linear		Quadratic		RBF	
	Training Set Accuracy	Test Set Accuracy	Training Set Accuracy	Test Set Accuracy	Training Set Accuracy	Test Set Accuracy
0.0001	0.731987578	0.712527154	0.609627329	0.627805938	0.60310559	0.612599566
0.001	0.867391304	0.863866763	0.629813665	0.641564084	0.60310559	0.612599566
0.01	0.913043478	0.910934106	0.637888199	0.648081101	0.667701863	0.679217958
0.1	0.931677019	0.922519913	0.657453416	0.661839247	0.69068323	0.703837799
1	0.923291925	0.919623461	0.67826087	0.684286749	0.706521739	0.706521739
10	0.889440994	0.896451846	0.703416149	0.709630702	0.734782609	0.746560463
100	0.75931677	0.753801593	0.727018634	0.730629978	0.839440994	0.82404055
1000	0.72623547	0.715623235	0.779813665	0.774076756	0.911801242	0.908761767
10000	0.697515528	0.680666184	0.854658385	0.856625634	0.935714286	0.927588704

- From the observation table for the three different models with varied values of C, we can observe the following points.
- The accuracy rate of the linear model is maximum for the middlemost values of c i.e. 0.1. When we plot accuracy with respect to C, we get a bell shaped curve with maximum accuracy at $C = 0.1$.
- The accuracy rates for both quadratic (polynomial) and RBF increases uniformly with increase in the value of C. The point to note here is that as C increases to a very high value like $C = 10000$, RBF model tends to show more accurate classification than the quadratic model.
- Linear model is least accurate at very high values of C.

6 Conclusion

- The training data can be classified into linear separable form.
- The reason for this classification is because for kernel other than linear model, they perform poorly with lower accuracy rate for low value of C. For very minute (small) values of C, we should get misclassified examples, even if our training data is linearly separable.
- We observed best test and training data accuracy for median values of C in case of linear model as compared to the other two models.
- So, this is the final observation obtained across different C parameter and various kernels for the given dataset.