CS1762: Experiment 2

Maximum Marks: 10

Submission deadline: 30 August, 2020

2. Write a program to experiment with the use of Support Vector Machines (SVMs) for binary and multiclass classification problem, and underst and the effects of varying various parameters therein.

To experiment with the use of SVMs for a multi-class classification problem, and understand the effects of varying various parameters therein. Your task is to try and learn an SVM classifier for the given data sets, using just the given features, and thereby also to assess the usefulness of the different features. Here is how you should proceed:

1. Data Visualization

Familiarize yourself with a standard SVM library using randomly generated data as given the attached code (SVM_Data_Visulization). You may wish to play with a simple Irish Data set (https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data) data set to get a feel for using the library, before you move on to the actual data for this assignment. The recommended one is LIBSVM (http:// www.csie.ntu.edu.tw/~cjlin/libsvm/), which is available for MATLAB, Python, and many other languages. Figure out how you can set various parameters, such as the value of C for the soft-margin SVM, the choice of kernel function, and the kernel parameters (if any).

2. **Binary classification:**

(i) Investigate linear SVM on the given "Bill" data sets using various hyper-parameters. Study the effects of changing the different parameter values, including the type of kernel function being used. How do they affect the accuracy?

(ii) Now, consider the data sets with multiple classes. Choose just 2 out of the 10 classes in your data, and train an SVM. Leave some data aside for validation, or ideally, use cross-validation. Study the effects of changing the different parameter values, including the type of kernel function being used. How do they affect the accuracy? Can you distinguish cases of overfitting, underfitting, and good fitting? Also try using only the first 10 features, instead of all 25, and compare the results in the two cases. Repeat this exercise for at least two more pairs of classes out of the 10 given to you. Do you consistently get the best results for the same parameter settings, or does it vary a lot depending on which pair of classes you're looking at?

3. **Multi-class classification** (Marks is not associated with this section and it is only for practice)
Now we would like to train a classifier for all 10 classes. Figure out what method(s) your chosen library uses for multi-class classification. For instance, LIBSVM uses one versus-one. Based on your choice, build a classifier for all the classes, and evaluate using validation or cross-validation. Again, study the effects of changing the various parameters and kernel function. Try and fine tune them to obtain the best possible performance. How do these tuned values compare to what you obtained in the binary classification setting? If there are major differences, what might be the reason? Also, try training the multi-class classifier using only the first 10 features, instead of all 25. How does this affect your results? What does this tell you about the usefulness of the different features?