**Introduction**

For this homework assignment we were tasked with using the LibSVM for MATLAB library to recognize the digits in the optdigits dataset. First, I formatted the data in a format that the SVM functions could understand, then I used cross-validation and performed a grid search to find the optimal SVM parameters (gamma and cost). Finally, I trained the model on the training data using these optimal parameters and ran the classification algorithm on testing data. I then used a confusion matrix to display the accuracy of the algorithm.

**Methods**

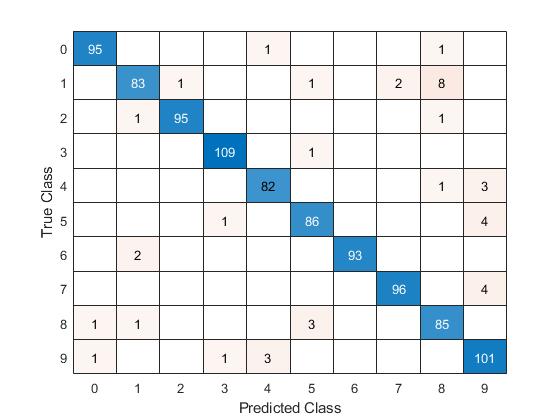
Below are algorithms that are critical to the implementation of the algorithm:

1. function [trainlabels,trainfeatures,testlabels,testfeatures] = splitData(class\_label,data)
   1. Separate the data into training and testing data. This is done to evaluate the model’s performance on unseen data.
2. function [c,g,accuracy,params] = gridSearch(trainlabels, trainfeatures,n
   1. Performs a cross-validation grid search on the data. Essentially, this function finds the optimal gamma and cost parameters. These parameters will be used in the classification process.
3. function model = train(trainlabels,trainfeatures,params)
   1. Uses the training labels, features, and parameters to create a model.
4. function [predicted\_label,cm] = predict(testlabels,testfeatures,model)
   1. Uses the model and svmpredict() to predict value of the unseen digits.
   2. Displays confusion matrix of the data.

**Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Datapoints** | **Cost** | **Gamma** | **Four-Fold Cross Validation Accuracy** | **Prediction Accuracy** |
| 20 | 2.548494e-03 | 2.000058e-15 | 9.813857e+01 | 95.6567% (925/967) |

**Figure 1**. Optimal Cost and Gamma Parameters (Grid Search Results)



**Figure 2.** Optdigit dataset Confusion Matrix

**Discussion**

Figure 1 shows the results from the grid search. The purpose of the grid search is to identify good Cost and Gamma values, so that the classifier can accurately predict unknown data. Four-fold cross-validation is used in the grid search. Cross-validation is used to prevent the model from overfitting the training set. To implement the grid-search, I varied the Cost values exponentially from 2.0e-15 to 2.0e+03 and the Gamma values from 2.0e-03 to 2.0e+15. The highest cross validation accuracy, 98.13857%, was achieved by setting cost to 2.548494e-03 and gamma to 2.000058e-15.

Figure 2 contains the confusion matrix for the Optdigit dataset. Confusion matrices are tables that are used to describe the performance of a classification model. In other words, for each digit, it shows the predicted value (predicted by using the model). Looking at the chart, numbers 1 and 9 were misclassified most often. Nine was often misclassified as a four. This is understandable since the two numbers share the same shape. Additionally, the number one was misclassified as an eight. It is not as clear why this took place; it could be due to a lack of those digits in the datasets.

**Software listing and executable software**

Start the program by pressing “Run”. The program will display the confusion matrix and the results from the grid search.