Support Vector Machine Classification Project

Dataset Description:

• ***mnist data.mat*** contains data from the MNIST dataset. There are 60,000 labeled digit images for training, and 10,000 digit images for testing. The images are flattened grayscale, 28 × 28 pixels. There are 10 possible labels for each image, namely, the digits 0–9.

Figure 1: Examples from the MNIST dataset.

• ***spam data.mat*** contains featurized spam data. The labels are 1 for spam and 0 for ham. The data folder includes the script featurize.py and the folders spam, ham (not spam), and test (unlabeled test data); you may modify featurize.py to generate new features for the spam data.

• ***cifar10 data.mat*** contains data from the CIFAR10 dataset. There are 50,000 labeled object images for training, and 10,000 object images for testing. The images are flattened to 3×32×32 (3 color channels). The labels 0–9 correspond alphabetically to the categories. For example, 0 means airplane, 1 means automobile, 2 means bird, and so on.

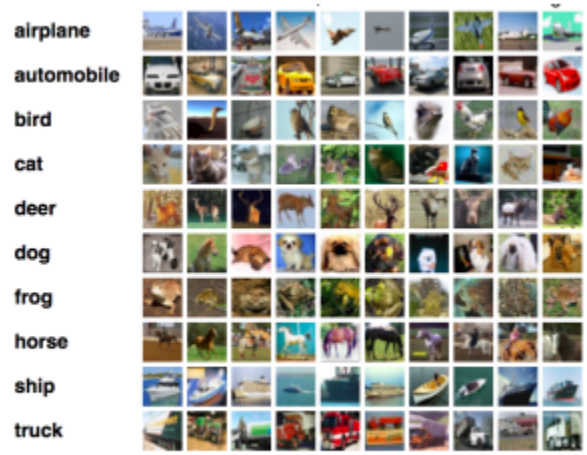

Figure 2: Examples from the CIFAR-10 dataset.

Methods for Optimization:

Using the grid search function from SKLearn to fine tune hyperparameters of my model significantly improved my accuracy score on all three datasets. Employing cross validation limited overfitting, gave me better results on the test sets, and helped in choosing the optimal hyperparameters. Another important factor was the size of the training sets I used in my grid search. For instance, running a high number of iterations with a smaller training set to pick the best parameters yielded far worse results than running fewer iterations with a larger training set in my grid search. Additionally, I found that the way I chose to normalize/standardize (ie preprocess) my data significantly impacted accuracy results on my test set. Also, for the spam dataset I was able to boost my score substantially by adding new features to the input vector such as length of the text, frequency counts of certain spam/ham keywords, and punctuation counts. One idea that did not work well was increasing my C value very high while using a very low gamma parameter. Although my grid search suggested these to be the best parameters in some cases, I found that the results were better when keeping C in the range of 1-5 and gamma in the range of .01 - .0001. Lastly, for all three datasets the “rbf” kernel yielded the best results for my SVC model.

Kaggle Competition and Results:

MNIST Dataset accuracy: 0.98580 = 98.58%
Placed in the top 3% of the class (over 700 students in CompSci 189)
https://www.kaggle.com/c/hw1-mnist-competition-cs189sp22/leaderboard

SPAM Dataset score: 0.86921 = 86.92%
Placed in the top 10% of the class (over 700 students in CompSci 189)
https://www.kaggle.com/c/hw1-spam-competition-cs189sp22/leaderboard

CIFAR10 Dataset score: 0.56470 = 56.47%
Placed in the top 9% of the class (over 700 students in CompSci 189)
https://www.kaggle.com/c/hw1-cifar-10-competitioncs189sp22/leaderboard