DS 310 Machine Learning
Fall 2020 / Amulya Yadav
HW #5: 50 points

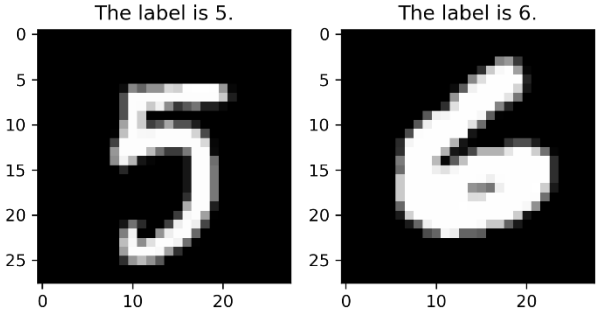
Please submit two files on Canvas:
- *your answer to this homework;*
- *the code for problem 1 (named "{your_name}_HW5.ipynb").*

Remember to submit the two files by clicking on "add another file" in Canvas, instead of submitting one zipped file of the aforementioned two files.

Please reach out to TA Hangzhi Guo (email: hangz@psu.edu) if there is any question about the assignment.

1. (40 points) In this problem, we will apply machine learning models to classify hand-written digits. The `example_notebook.ipynb` is a good starting point.

Download the digit data set from the course website. The zip archive contains two files: Both files are text files. Each file contains a matrix with one data point (= vector of length 784) per row. The 784-vector in each row represents a 28 × 28 image of a handwritten number. The data contains two classes—the digits 5 and 6—so they can be labeled as 0 and 1, respectively. The image on the right shows the first row, re-arranged as a 28 × 28 matrix and plotted as a gray scale image.



(A) First, you should try to classify digits using SVM model. In specific, you will

- Randomly select about 25% of the data and set it aside as a test set.
- Train a linear SVM with soft margin. Cross-validate the margin parameter.
- Train an SVM with soft margin and RBF kernel. You will have to cross-validate both the soft-margin parameter and the kernel bandwidth.
- After you have selected parameter values for both algorithms, train each one with the parameter value you have chosen. Then compute the misclassification rate (the proportion of misclassified data points) on the test set.

(B) Plot the cross-validation estimates of the misclassification rate. Please plot the rate as

(a) a function of the margin parameter in the linear case.

(b) a function of the margin parameter and the kernel bandwidth in the non-linear case (you are encouraged to use heat map here).

(v) Report the test set estimates of the misclassification rates for both cases, with the parameter values you have selected, and compare the two results. Is a linear SVM a good choice for this data, or should we use a non-linear one?

(vi) You will then implement a neural network for classifying digits. If you do not have experiences in implementing the neural network, we recommend using Keras.

(i) Implement a neural network with convolutional neural network layers. You should set batch size to be 128, learning rate to be 0.1, and the number of epochs to be 10.
   a) Plot the learning curve with respect to the training loss.
   b) Report the test set estimates of the misclassification rates.

(ii) Next, you should tune the learning rate. You should try *five other learning rates* with batch size to be 128 and the number of epochs to be 10. Plot the learning curve of each learning rates with respect to the training loss (there should be five plots in total).

2. (10 points) Consider the dataset: Stock market data, which include the prices and volumes of various stocks on different trading days. Then, for this dataset, give specific examples of both classification and clustering tasks that can be performed. For each task, state how the data matrix should be constructed (i.e., specify the rows and columns of the matrix).
• Task:
• Row:
• Column:

1. **Classification**
   A. **Task**: The specific classification task that we can do with this particular stock market data is classify if a stock will be past/within certain price thresholds based on the volume and day. Price ranges will need to be encoded to single value numeric (e.g., $5.00 - $10.00 = 1).
   B. **Row**: The rows will represent each stock (i.e., each row will be a stock ticker).
   C. **Columns**: Each row will have four columns which include the ticker, volume, day, and the output column, price.
2. **Clustering**
   A. **Task**: A clustering task could be finding patterns in the day of the week based on the volume and price of stocks. We can try to find five clusters representing a day of the week. This task may reveal various patterns such as that more people purchase/sell stocks at the end/beginning of the week compared to the middle of the week.
   B. **Row**: The rows will be the volume.
   C. **Column**: The columns will be price which contain the values of the days of the week.