

SYDE675: Introduction to Pattern Recognition**Assignment 3**

Due: 11:59 PM (EST), March 22, 2024, submit on LEARN.

Include your name and student number!

Submit your write-up in pdf and all source code in a zip file (with proper documentation). Write a script for each programming exercise so that the TAs can easily run and verify your results. Make sure your code runs!

[Text in square brackets are hints that can be ignored.]

NOTE For all the exercises, you are only allowed to use the basic Python, Numpy, and matplotlib (for plotting) libraries, unless specified otherwise. For any of the exercises, if timing is an issue, you can reduce the number of training images or the number of epochs.

Exercise 1: Logistic Regression (30 pts)

In this exercise you will be using the MNIST dataset. Use **PCA** in Scikit learn to convert the 784×1 vectors to 2×1 vectors (2-d data). Use the training set for implementing the classifiers in the exercise. Also, use only the data for digits/classes 3 and 4 in the dataset.

1. (10 pts) Implement the logistic regression algorithm to classify the MNIST dataset. Let $\Pr(C_1|x) = \sigma(\mathbf{w}^T \mathbf{x} + w_0)$ and $\Pr(C_2|x) = 1 - \sigma(\mathbf{w}^T \mathbf{x} + w_0)$. Learn the parameters \mathbf{w} and w_0 using the gradient descent algorithm. Use the maximum number of epochs to be 100 for GD. You can also use any appropriate convergence criteria to stop the GD loop before 100 epochs. Use a step size of 0.1 (or 0.01 if it is converging poorly) for GD.
2. (3 pts) After the training process, create the following plots:
 - (a) test error vs the number of epochs
 - (b) training error vs the number of epochs
 - (c) test loss vs the number of epochs
 - (d) training loss vs the number of epochs
3. (6 pts) Retrain your classifier but without any activation function i.e. no sigmoid at the end. How does the classifier perform compared to when there is a sigmoid activation? Explain.
4. (6 pts) Now let's fix your bias to any non-zero real value (you can choose 1 for convenience). Keep the sigmoid activation in the classifier. Repeat the training and testing as in steps 1 and 2. What happens to the performance of the classifier? Explain.
5. (5 pts) Compare the final test error with your classifiers in previous exercises (MED, MMD, kNN). Which classifier performs the best? Why?

Exercise 2: CNN Implementation (50 pts)

Note: Please mention your Python version (and maybe the version of all other packages).

In this exercise, you are going to run some experiments involving CNNs. You need to know **Python** and install the following libraries: **Pytorch**, **matplotlib** and all their dependencies. You can find detailed instructions and tutorials for each of these libraries on the respective websites.

For all experiments, running on the CPU is sufficient. You do not need to run the code on GPUs. Before starting, we suggest you review what we learned about each layer in CNN, and read at least this **tutorial**.

1. (17 pts) Implement and train a VGG11 net on the complete **MNIST** dataset. VGG11 was an earlier version of VGG16 and can be found as model A in Table 1 of this **paper**, whose Section 2.1 also gives you all the details about each layer. The goal is to get the loss as close to 0 as possible. Note that our input dimension is different from the VGG paper. You need to resize each image in MNIST from its original size 28×28 to 32×32 [why?].

For your convenience, we list the details of the VGG11 architecture here. The convolutional layers are denoted as Conv(number of input channels, number of output channels, kernel size,

`stride, padding`); the batch normalization layers are denoted as `BatchNorm(number of channels)`; the max-pooling layers are denoted as `MaxPool(kernel size, stride)`; the fully-connected layers are denoted as `Linear(number of input features, number of output features)`; the drop out layers are denoted as `Dropout(dropout ratio)`:

```
- Conv(001, 064, 3, 1, 1) - BatchNorm(064) - ReLU - MaxPool(2, 2)
- Conv(064, 128, 3, 1, 1) - BatchNorm(128) - ReLU - MaxPool(2, 2)
- Conv(128, 256, 3, 1, 1) - BatchNorm(256) - ReLU
- Conv(256, 256, 3, 1, 1) - BatchNorm(256) - ReLU - MaxPool(2, 2)
- Conv(256, 512, 3, 1, 1) - BatchNorm(512) - ReLU
- Conv(512, 512, 3, 1, 1) - BatchNorm(512) - ReLU - MaxPool(2, 2)
- Conv(512, 512, 3, 1, 1) - BatchNorm(512) - ReLU
- Conv(512, 512, 3, 1, 1) - BatchNorm(512) - ReLU - MaxPool(2, 2)
- Linear(0512, 4096) - ReLU - Dropout(0.5)
- Linear(4096, 4096) - ReLU - Dropout(0.5)
- Linear(4096, 10)
```

You should use the following in your training process unless specified otherwise: cross-entropy loss `torch.nn.CrossEntropyLoss`, and optimize using SGD optimizer with momentum=0.9.

[This experiment will take up to 1 hour on a CPU, so please be cautious of your time. If this running time is not bearable, you may cut the training set to 1/10, so only have ~600 images per class instead of the regular ~6000.]

2. (3 pts) Once you've done the above, the next goal is to inspect the training process. Create the following plots:
 - (a) test accuracy vs the number of epochs
 - (b) training accuracy vs the number of epochs
 - (c) test loss vs the number of epochs
 - (d) training loss vs the number of epochs

[If timing is an issue, train the neural networks for 20 epochs. Although 20 epochs are not sufficient to reach convergence, it is sufficient to see the trend. You can reduce the number of epochs further if it is infeasible to train for 20 epochs.]

3. (10 pts) Train your CNN with Adam. Plot the graphs for training and test accuracies vs the number of epochs for Adam. Compare the accuracies of Adam with the SGD accuracies in Ex 1.2. Which optimizer performed better and why?
4. (10 pts) Replace ReLU activation with Sigmoid in your CNN and train it again. Plot the graphs for training and test accuracies vs the number of epochs. Compare it with your results in Ex 2.2. Do you see any change because of Sigmoid? If so, explain why.
5. (10 pts) Remove Dropout from your CNN and train it again. Plot the graphs for training and test accuracies vs the number of epochs. Compare it with your results in Ex 2.2. Did Dropout have any effect on your network performance? Explain.

Exercise 3: MLP Implementation (20 pts)

In this exercise, you are going to run some experiments involving multi-layer perceptrons (MLPs). You can implement them in Pytorch as well (see the instructions at the beginning of Exercise 2).

1. (12 pts) Implement a three-layer MLP (details below) and train it on the MNIST dataset. The details of the MLP architecture are:
 - `Linear(784, 512)` - ReLU
 - `Linear(512, 512)` - BatchNorm(512) - ReLU

- `Linear(512, 10)`

You should use the cross-entropy loss `torch.nn.CrossEntropyLoss` at the end. Also, use the SGD optimizer with `momentum=0.9`.

[If you used 600 images per class in Exercise 2, then use the same images here.]

Note: You will need to flatten your images from 28×28 to 784×1 vectors, because the MLP expects a vector as an input.

2. (3 pts) After the training process, create the following plots:

- (a) test accuracy vs the number of epochs
- (b) training accuracy vs the number of epochs
- (c) test loss vs the number of epochs
- (d) training loss vs the number of epochs

[If timing is a concern, train the neural networks for 20 epochs. Although 20 epochs is not sufficient to reach convergence, it is sufficient to see the trend. You can reduce the number of epochs further if it is infeasible to train for 20 epochs.]

3. (5 pts) Compare the accuracies of the MLP and VGG11. Which network is better for MNIST classification and why?