Wright State University
Computer Science and Engineering

CS4850-6850– Assignment3
Foundation of AI

Instructor: Dr. Tomojit Ghosh

Due by 2024/10/24

This assignment is intended to apply the knowledge of Data Driven Learning on some practical data. Both the section CS4850 and 6850 are required to complete it. Good luck and happy coding work!

Distribution of Marks

Question	Points	Score
1	25	
2	40	
3	35	
Total:	100	

1. (25 points) In this assignment you will work on a small MNIST data sets with 1000 training samples (100 from each digit classes) and 200 test samples. These are handwritten digits of $0, 1 \ldots 9$. Run Autoencoder on the data. In this experiment use one hidden layer with two/three linear nodes. (You can use the code provided, but you need to change it). Project the data in two-dimensional and three-dimensional bottleneck space. Show the projected data using a scatter plot. Your figure should have proper legend for each digit class. Compare the projection with 2D and 3D PCA plots and comment on these plots.

2. (40 points) In this question you will explore the effect of depth, i.e., no. of hidden layers, towards the embedding of Autoencoder. You will use the following network architecture (d: input data dimension):

   - 1. $d \to 50 \to 2 \to 50 \to d$
   - 2. $d \to 100 \to 50 \to 2 \to 50 \to 100 \to d$
   - 3. $d \to 200 \to 100 \to 50 \to 2 \to 50 \to 100 \to 200 \to d$

   Set the activation function in the bottleneck layer as 'linear' and in other hidden layers as 'tanh'. Train each Autoencoders on the same training set. After that pass the training and test data through the encoder to get hidden codes and calculate 5NN (five nearest neighbor) accuracy for each netwrork and report the result in a table with explanation. This time you will run the experiment ten times and calculate the mean accuracy with standard deviation. Commnent on the result.

   Note: Nearest Neighbor is a simple classifier. To predict the class label of a test sample, you need to find out its k-nearest neighbors from the training set, where k is user select (could be 1,3,5,7,...etc). Given the label information of the training set, assign the label of the test samples by majority voting. Example: Assume $k = 5$ and you want to assign the class label of a test sample x^i_{test}. Find out the five nearest neighbors of x^i_{test} from the training set using Euclidean distance. If the majority of the neighbors of x^i_{test} belongs to class C_j, then assign the sample x^i_{test} to the j^{th}-class. For this assignment, use scikit-learn package: `https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html`

3. (35 points) In this question, you will work with an ANN classifier (the code is provided, you can change the activation function if your wish) with one hidden layer to observe the effect of no. of hidden units on the test data's misclassification rate (error rate). Vary the number of nodes in a logarithmic scale, i.e., $10, 20, 40, 80, \ldots$ up to 1280. For each network configuration, train the model ten times and compute the average error rate. After that, plot the error rate for each network configuration and comment on the plot.

   Note: You should submit a pdf file with you answers. Make sure to include the code snippet, written in Python/PyTorch for each questions. The figures should have proper legend, labels to X and Y axixes.