CSCE 4930 - Introduction to Machine Learning
Assignment 3

This is an individual assignment. Rules governing academic integrity are strongly enforced without exception.

1. In this problem, you will be using a dataset representing digital images of handwritten digits of different customers of a large retailer. Each digit is represented in the library using 16x16 pixel black/white images, where the value of every pixel is 0 if it is black, or 1 if it is white. Hence, we can consider every image in our dataset as a sample with 16x16 = 256 binary features, while the labels represent the actual number that the image represents. For each digit, we have 600 training samples and 500 testing samples that you can find on Blackboard in the assignments folder. The training samples are stored in the files called "training_features.txt" and "training_labels.txt", where each line in the labels file corresponds to the label (actual digit) for the features represented in the same line in the features file. Same format is used to store the test data.

    a. Propose a neural network structure to predict the digit given its hand-written image. Make sure to justify your choices for the structure (specifically the input and output layers) as well as your activation function. **(5 points)**

    b. Implement a neural network model with back-propagation using the algorithms we discussed in class and the appropriate data structures. Make sure your implementation allows for different network structures that the user can provide (either interactively or through a config file). Your code has to be implemented from scratch, not using any of the existing libraries **(25 points)**

    c. Use your implementation to train a model using the training data from the provided dataset and report the average precision, recall, and accuracy of your model on the test data **(10 points)**

    d. Implement a distortion model for the data and re-run your model training and testing on the distorted dataset at different distortion levels. Plot the average precision and recall of the your test runs at different distortion levels (At a distortion level of X%, your distortion model should basically produce a distorted dataset by randomly picking X% of the pixels in each image and assigning random black/white (0/1) values to them). **(10 points)**

2. Using the following text embedding table for different schools and cities , answer the questions below, illustrating all your steps and computations. Note that MIT is located in Boston, Stanford and Berkeley are both located in San Francisco

    a. Compute the embedding vector for a school that is similar to Stanford but located in San Diego. **(5 points)**

    b. Which school is the most similar to MIT among the ones in San Francisco? **(5 points)**

MIT	[0.3, 0.5, -0.8, 0.4]
San Francisco	[0.1, 0.2, 0.9, 0.6]
San Diego	[0.1, 0.3, 0.8, 0.7]
Stanford	[0.4, 0.7, -0.6, 0.4]
Berkeley	[0.4, 0.6, -0.7, 0.3]
Boston	[0.1, 0.1, -0.5, -0.1]

3. For the following 2-D dataset, use Euclidean distance to answer the following questions

point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

   a. Construct a complete-link hierarchical clustering for the points in the dataset. Show your full dendrogram, and illustrate how the points in the dataset will be grouped into two clusters **(10 points)**

   b. Assuming we start from cluster centroids p1 & p2, trace k-means clustering with k=2 for 5 iterations and show how the points in the dataset are assigned to the two clusters and how the centroids change **(10 points)**

   c. Assuming that the results from hierarchical clustering are the true assignments of the points (their true clusters). Compute the precision and recall for the k-means assignment with respect to the assumed true classes of the points. **(10 points)**

   d. Using affinity propagation, go through 5 iterations of computing the availability and responsibility matrices for this data set. Discuss the impact of the different choices of the self-similarity values on the projected clustering output (hint: you can write a small script to compute the matrices and try different self similarity values for the discussion - Make sure to illustrate at least one iteration by hand) **(10 points)**