CSCE 4930 - Introduction to Machine Learning
Assignment 4

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

1. In this problem, you will be using the MNIST dataset (https://www.kaggle.com/oddrationale/mnist-in-csv) representing images of handwritten digits, where each digit is represented using 28x28 pixel grey scale images (a number between 0 and 255). Each line contains 785 values, the first value is the label / digit (0-9) and the remaining 784 values are the values of the pixels in row major. The file to be used is mnist_train.csv containing 60,000 images, available at the link above.

    a. Implement an Expectation-Maximization model using the algorithm we discussed in class. Make sure your model can accept an input in a standard format (according to your specification) and output the cluster assignments of the input data (with probabilities for each cluster). Your code has to be implemented from scratch, not using any of the existing libraries. **(20 points)**

    b. Use your implementation to cluster the above dataset into 10 clusters corresponding to the digits from 0-9. Make sure to remove the label / digit from the dataset before running your E-M algorithm. Report the cluster assignments. **(5 points)**

    c. Using the clustering you obtained in part b), assign each point to the cluster with the highest probability, and report the precision, recall, and accuracy of your cluster assignments as compared to the true digit representation. **(5 points)**

    d. Use the mean vector for each cluster to draw an image representing the cluster representative value, with the pixel values of the mean vector for each of the clusters. **(5 points)**

    e. Implement an affinity propagation model using the algorithm we discussed in class. Your code has to be implemented from scratch, not using any of the existing libraries. **(15 points)**

    f. Vary the self-similarity parameter in your affinity propagation model, and show how that affects the resulting clustering of the above dataset in terms of precision, recall, and accuracy with respect to the true labels of each digit. **(10 points)**

    g. Plot the images of the cluster representatives of affinity propagation (for a value of self similarity that result in the closest number of clusters to 10) **(5 points)**

2. The following CPTs represent the conditional probabilities measured for a dataset of students behavior during the school year. The variables in the data are all binary and measure the corresponding feature of the students.

$$\begin{aligned}\mathbf{P}(\text{creative} = T) &= 0.69932\\ \mathbf{P}(\text{smart} = T) &= 0.70472\\ \mathbf{P}(\text{party} = T) &= 0.60216\end{aligned}$$

creative	smart	$\mathbf{P}(\text{project} = T \mid \text{creative}, \text{smart})$
T	T	0.90484
T	F	0.40307
F	T	0.79326
F	F	0.10731

project	hw	$\mathbf{P}(\text{success} = T \mid \text{project}, \text{hw})$
T	T	0.89633
T	F	0.20737
F	T	0.30714
F	F	0.05066

smart	party	$\mathbf{P}(\text{hw} = T \mid \text{smart}, \text{party})$
T	T	0.80252
T	F	0.89790
F	T	0.09447
F	F	0.30556

a. Draw a Bayesian Network that can be represented by the above CPTs **(5 points)**

b. Answer the following questions based on the provided network (Make sure to show your steps):
   i. What is the probability of a random student succeeding in the course? **(5 points)**
   ii. What is the probability of success for a student who is smart, parties often but not very creative? **(5 points)**
   iii. What is the probability of a smart student to finish his/her project? **(5 points)**

3. For the following dataset, show how an AdaBoost classifier works with 4 classifiers. For the individual classifiers, use a single variable decision tree that splits the data on a single variable. For each loop iteration, report the following variables ϵ_t, α_t, $\forall i\ D_t(i)$.
Calculate the precision, recall, and accuracy of the final voting classifier on the training data. Compare these metrics with the single variable decision trees classifiers. **(15 points)**