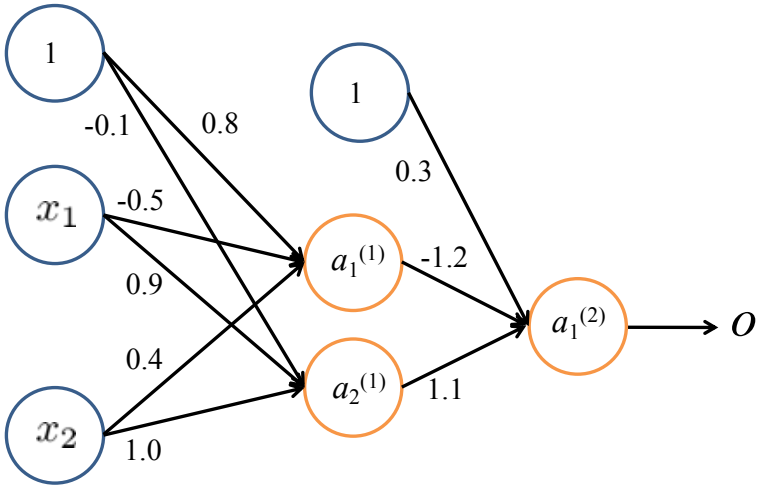
CS 5661: Topics in Data Science
Homework2, Due Date: Mon, Mar 19
Instructor: Dr. Mohammad Pourhomayoun

Question1: Backpropagation

In the following Neural Network, we have initialized the weights randomly. Use a training sample (X,y) = ((1,1), 0) to update the weights (perform one round of backpropagation using one training sample). Use learning rate parameter α = 0.1. Note that we have bias terms with value of 1 in this network (no need for coding for this question).



Question2: Handwriting Recognition Using ANN:

Write and submit your python codes in "Jupyter Notebook" to perform the following tasks. Make sure to provide proper descriptions as MarkDown for each section of your code.

a- Download the dataset "Digit" from CSNS. Check out the dataset. It includes 1797 small images (8x8 pixels), each one includes a hand-written digit (0-9). You have to download the corresponding csv file that includes the labels of the images. The goal is to build an Artificial Neural Network that can recognize the hand-written digits.

Import the following two libraries to work with images:
import matplotlib.image as mpimg
import matplotlib.pyplot as plt

you can use:
mpimg.imread(file_name) to load an image, and
plt.imshow(image_name, cmap=plt.cm.gray_r, interpolation='nearest') to show an image. Add **%matplotlib inline** at top of your code to make sure that the images will be shown inside the Jupyter explorer page.

b- Build the feature matrix and label vector: Each image is considered as a data sample with pixels as features. Thus, to build the feature table you have to convert each 8x8 image into a row of the feature matrix with 64 feature columns for 64 pixels.

c- Use sklearn functions to split the dataset into testing and training sets with the following parameters: **test_size=0.1, random_state=2**.

d- Design and Train an ANN with one hidden layer with **80 neurons** to recognize the digits based on the training dataset that you built in part (c). Use **random_state=1, learning_rate_init = 0.002**. Then, Test your ANN on testing set (from part(c)), and calculate and report the accuracy. Also, calculate and report the Confusion Matrix.

e- Now, use **GridSearchCV** to find the best number of neurons for your 1-hidden layer network. Search in the range of 50-200 neurons, and use the following parameters for your NN: activation='logistic', solver='adam', alpha=1e-5, random_state=1, learning_rate_init = 0.002. What is the best accuracy, and best number of neurons?

Note: Since the size of the dataset is not small, it may take a long time to finish the process (depending on your system it can be up to hours). So, don't leave it for the last minutes (late submissions will not be accepted under any circumstances!)