ASSIGNMENT 2 BACKPROPAGATION

CISC/CMPE 452/COGS400

20 marks

General Instructions for Code and Submission (for all assignments)

1. You can use any programming language (preferred Python, java, C, C++, Java)
2. Make one zip file named as Asgn2_studentID which should include
   a. *A PDF file (template given on the next page) containing the ANN design, accuracy in precision, recall, and confusion matrix, comparison of accuracies of Model 1 and 2. Include a critical analysis about possible causes for the difference in accuracies.*
   b. *Program code with comments in the code to explain what each of your program files and functions are for*
3. Upload zip file to the OnQ site
4. Mark will be deducted based on late policy (-2 per day after due date until the end date after which no assignments will be accepted)

Assignment 2: The Original Dataset

- You will use the handwritten character database for this assignment published by Lecun et al available at the following URL yann.lecun.com/exdb/mnist/ (accessible on OnQ as MNIST handwritten digit database, Yann LeCun, Corinna Cortes an.pdf). You will know about all the different models that were implemented to reduce the error of the initial model.

Assignment 2 – Your Task

- You will implement the Backpropagation learning algorithm for a multilayer (input, one hidden layer and one output layer) in the following two different ways to learn to classify handwritten characters using the MNIST training dataset and the corresponding label dataset.
  - *Model 1: Using raw programming without using any predefined libraries to create, train or test the models. You can use a library to create a visualization or a chart from the predicted class values.*
  - *Model 2: Using predefined libraries in data analytic tools like Keras or Scikit-learn. A) Implement one model with the same parameters as you use in Model 1. B) Try to improve Model 2A by trying other parameter values.*
- The input should be the training data and the output should be the corresponding digit class. Use sigmoidal output function at every layer and node. In other words, you must decide the following design criteria:
  - *Initial weights and learning rate*
  - *Training iterations and termination criteria*
  - *Number of nodes in the layer*
  - *Momentum*
- Use the predefined libraries as shown on the next slide to load the dataset in your program.
- Compare the accuracies of models 1 and 2.

Assignment 2: Loading the Data in Python

- Using Keras package:

```
from tensorflow.keras.datasets import mnist
# Load data
(X_train, y_train), (X_test, y_test) = mnist.load_data()
```

- Using scikit-learn package:

```
from sklearn.datasets import load_digits, fetch_openml
from sklearn.model_selection import train_test_split

# Load data from https://www.openml.org/d/554
X, y = fetch_openml('mnist_784', version=1, return_X_y=True)

# Split data into train and test splits
X_train, X_test, y_train, y_test = train_test_split(
    X, y, train_size=60000, test_size=10000, shuffle=False)
```

- Both will load same data but input features (i.e. pixel values for an image) will have different dimensions
  - *i.e. Multi-dimensional array of shape 28x28 Vs. One-dimensional array of shape 784*
- For more details, please check:
  - https://www.tensorflow.org/api_docs/python/tf/keras/datasets/mnist/load_data
  - https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_openml.html

Mark Distribution and Rubrics –
Total 20 Marks

- 8 marks – For properly documented (with comments) fully functional executable code of Model 1 achieving high accuracy. (5 marks if model functions but less optimally with some missing documentation, 0 for non-functional models)
- 6 marks – For properly documented (with comments) fully functional executable code of Model 2A (3 marks) and 2B (3 marks). (3 marks for non optimal coding and low accuracy, 0 for non-functional models)
- 6 marks – Complete report (pdf file) containing confusion matrix, overall precision, recall, and a critical discussion about lessons learned (choice of design factors and parameter values) from design choices, models' performances and classification accuracies. You must present the results for both the training and testing phases. (3 marks if results and parts of above reporting are missing, 0 if report is missing or none of the required information is presented in the report).