ADTA 5550.400: Deep Learning with Big Data

Midterm Assessment

1. Overview

The midterm covers all the topics that have been discussed in the first half of the course. The materials in any format, should be considered and used for the midterm. Additionally, the student can use any other source of information that he/she can gather, providing it is relevant and supporting the student's answers.

The student is required to create an MS Word document named "**ADTA5550_midterm.docx**" that will contain all his/her midterm work, except for the Python coding.

IMPORTANT NOTES:
--) *--) If an MS Word document is specified as the required format of the submitted document, the student should **back up** the MS Word document by saving it as a PDF file before submitting it.*

IMPORTANT NOTES:
*--) If an MS Word document is specified as the required format of the submitted document, the student should submit it, **not** submit a PDF.*

--) All the submission requirements are expected to be submitted in an MS Word document, except for Python code or being specified otherwise.

--) When discussing a topic or answering a question, it is expected that the student has to provide adequate explanations and supporting details.

IMPORTANT NOTES:
*--) For the Python code, the student is required to write the **code of each step** in **one cell** of the Jupyter Notebook document, as shown in the lectures. Then the student is required to **run the code of each cell** to **show the results of each step** in each submitted Jupyter Notebook document.*

*--) For Python code in Jupyter Notebooks, the student is required to run the code and submit the Jupyter Notebook document that contains the results. The student should **not** copy the results of Python code into the MS Word document.*

2. Datasets

IMPORTANT NOTES:
--) All data sets can be found in the Canvas module**:…/DATA_SETS**

2.1 Dataset: pima_diabetes.csv

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old.

The dataset consists of 8 medical predictor variables and one target/outcome variable (class). Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

The variables of this data sets:

1. preg: Number of times pregnant (preg)
2. plas: Plasma glucose concentration a 2 hours in an oral glucose tolerance test (plas)
3. pres: Diastolic blood pressure in mm Hg (pres)
4. skin: Triceps skin fold thickness in mm (skin)
5. test: 2-Hour serum insulin in mu U/ml (insu)
6. mass: Body mass index measured as weight in kg/(height in m)^2 (mass)
7. pedi: Diabetes pedigree function (pedi)
8. age: Age in years (age)
9. class: 0 or 1 (no or yes)(negative or positive)

IMPORTANT NOTES:
--) The student needs to download the data set pima_diabetes.csv from the Canvas module: .../DATA_SETS, and then upload it to the remote deep learning server

--) The steps of how to upload the data file to the remote server are discussed in the document:
***HOWTO_upload_files_to_remote_server_using_GCP_SSH**.pdf (Canvas module: .../SW_DOCS)*

--) In one of the first steps of coding, the student needs to load the data set into a Pandas data frame. The details of how to do this are discussed in the document:
***HOWTO_load_dataset_into_dataframe_in_remote_server.pdf** (Canvas module: .../SW_DOCS)*

IMPORTANT NOTES:

Python code to load the dataset into a Pandas data frame:

```
# Specify what and where is the data file
# MUST SPECIFY THE LOCATION OF THE FILE – USING THE CORRECT PATH
filename = '…/pima_diabetes.csv'

# Specify the fields with their names
col_names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']

# Load the data into a Pandas DataFrame
df = read_csv(filename, names=col_names)
```

3. PART I: AI Deep Learning (20 Points)

SUBMISSION REQUIREMENT #1:

Question 1.1:
--) Provide an overview (at a **minimum** of **2 pages, including images**) of the history of artificial intelligence, including its sub-fields, machine learning and deep learning.

Question 1.2:
--) Provide an overview (at a **minimum** of **1.5 pages, including images**) of deep learning, including (but **not** limited to) the relationship between deep learning and machine learning, artificial intelligence.

Question 1.3:
--) Explain (at a **minimum** of **1.5 pages, including images**) why Deep Learning is very popular in recent years.

4. PART II: MLPs (Fully Connected Neural Networks) with Keras (50 Points)

TO-DO

Build, train, and **evaluate** a deep neural network MLP that has **two layers**. The training is done on the dataset **pima_diabetes.csv** using the Keras sequential model in a Jupyter Notebook document.

IMPORTANT NOTES:
--) It is expected that the student knows how to count the layers of a deep neural network.

--) First, **design** the MLP neural network, i.e., it has two layers, to train a deep learning model on the dataset. Then, using MS PowerPoint or Draw Tool in MS Word to **draw a diagram** of the neural network with all the layers, the neurons, and the feed-forwarding connections.

--) **Build** and **train** the model using the Keras sequential model in a Jupyter Notebook document.

IMPORTANT NOTES:
*--) It is expected that the student completes all the steps, including the **dataset introduction**, the data preprocessing, EDA of the dataset, etc.*

--) **Evaluate** the model using the KerasClassifier and the **10-fold** cross-validation technique.

--) Obtain the accuracy level from the training process, say **accuracy_training**.
--) Obtain the accuracy level from the evaluation process, say **accuracy_evaluation**.

--) **Write a report** on these results:
- Present these results
- Compare these results: accuracy_training versus accuracy_evaluation.
- Using critical thinking to provide a **reasonable** explanation of the gap between them.

--) **Write another report** to **compare** the accuracy level from the evaluation process, i.e., accuracy_evaluation, obtained in this project (MLP on **pima_diabetes.csv**) with the accuracy level from the evaluation process of the project discussed in the lecture (MLP on **Iris.csv**)

- Present these results.
- Compare these results.
- Using critical thinking to provide a **reasonable** explanation of the gap between them.

SUBMISSION REQUIREMENT #2

--) Add **one section** to discuss the design of the MLP into the MS Word document above, "**ADTA5550_midterm.docx,"** including the **diagram** of the neural network.

--) **Run all the steps** of the project in the Jupyter Notebook document to get the results of each step.

--) Add **another** section to the MS DOCS document "**ADTA5550_midterm.docx**" for the above report on the accuracy levels of the model obtained from both the training and the evaluation process.

--) Add **one more** section to the MS DOCS document "**ADTA5550_midterm.docx**" for the above report to **compare** the accuracy level from the evaluation process in this project (MLP on **pima_diabetes.csv**) with the accuracy level from the evaluation process in the lecture (MLP on **Iris.csv**).

5. PART III: Redesign the MLP (30 Points)

TO-DO
To improve the performance of the MLP on the dataset pima_diabetes.csv, it is assumed that the student plans to use the trial and error approach experimenting with a new design of the MLP. There are many ways to redesign a neural network.

--) First, based on the knowledge of the deep neural network MLP and using critical thinking, the student **redesigns** the MLP neural network, then **build, train,** and **evaluate** the redesigned MLP to find out if it produces a higher accuracy level.

--) Using MS PowerPoint or Draw Tool in MS Word to **draw the diagram** of the redesigned neural networks with all the layers, the neurons, and the feed-forwarding connections.

--) **Redo all the steps** of the project "**MLP on pima_diabetes.csv**" in **another** Jupyter Notebook document.

SUBMISSION REQUIREMENT #3

--) To discuss the new design of the MLP, add **one new** section into the MS DOCS document above: "**ADTA5550_midterm.docx"**. The discussion should include:

- The **diagram** of the redesigned neural network

- Discussion in detail of **how** the MLP is re-designed
- Discussion in detail of **why** such a redesigned network can potentially produce improved performance, i.e., higher accuracy level.

--) **Run all the steps** of the project in the Jupyter Notebook document to get the results of each step.

--) Add **another** section to the MS DOCS document: "**ADTA5550_midterm.docx**" to discuss the results obtained from the redesigned MLP, especially comparing them with those from PART II.

IMPORTANT NOTES:
*--) With the assumption that the student uses the trial and error approach, it is **OK** if the results of training and evaluating the redesigned neural network do not show any significant improvement in the network performance.*

6. HOWTO Submit

Due date & time: 11:59 PM – Thursday 07/04/2024

The student is required to submit the midterm – all the documents: Microsoft Word and Jupyter Notebooks – on Canvas in the appropriate section (Week4 : Midterm).