ADTA 5340: Discovery and Learning with Big Data

Thuan L Nguyen, PhD

Midterm Assessment

1. Overview

The midterm covers all the topics discussed in the first half of the course. Materials in any format, including in-class discussion, should be considered and used for the midterm. Additionally, the student can use any other source of information that he/she can gather, provided it is relevant and supports the student's answers.

The student is required to create an MS Word document named "**ADTA5340_midterm.docx**" that contains 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 otherwise specified.

IMPORTANT NOTES:
*--) For the Python code, the student must 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 must run the code and submit the Jupyter Notebook document containing the results. The student should refrain from copying the results of Python code into the MS Word document.

2. Data Sets

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

IMPORTANT NOTES:

--) *Without further instructions, the student must **use all the attributes/variables** of a data set to train the model, i.e., **no feature selection** should be made.*

3. PART I: AI and Machine Learning (15 Points)

SUBMISSION REQUIREMENT PART I:

Researchers need coding to build, train, and test AI/machine learning models.

Discuss **in-depth** (including images, diagrams, etc.) the similarities and differences between general programming and AI/machine learning, focusing on the differences.

4. PART II: AI Machine Learning: Learning Styles and Process (15 Points)

SUBMISSION REQUIREMENT PART II:

Discuss **in-depth** (including images, diagrams, etc.) **three primary learning styles** of AI machine learning.

5. PART III: Preprocessing Data (10 Points)

TO-DO
--) **Preprocess data**: Missing values

- Clean the dataset **loan_approval.csv** by removing records that contain missing values.

--) **Preprocess data**: Mixed up data types of the values of the same attribute (variable) (See the document ADTA5340_MIDTERM_important_notes.pdf)

- Clean the dataset to address the issue

IMPORTANT NOTES:
--) To detect and handle the missing values in the dataset, the student can apply any approach that he/she finds appropriate and comfortable, including writing Python code (if he/she knows how to do it), using tools (if he/she knows any tool), or doing it manually, e.g., open the CSV file with Excel, looking for the missing value and remove the record from the file.

SUBMISSION REQUIREMENT PART III:

--) **Add a section** to the above MS Word document ("**ADTA5340_midterm.docx**") to discuss **in detail** how the student detects and handles the missing values in each record, the number of records in the original dataset, and the number of records in the clean one.

--) The **clean** dataset **loan_approval.csv** after preprocessing the data

IMPORTANT NOTES for the next sections **(PART IV** and **V):**
--) For Exploratory Data Analysis (EDA), **each step** of this analysis must be coded in **one cell**.
--) For Exploratory Data Analysis (EDA), **univariate data visualization**, **each chart** of **each applicable variable** must be displayed in **its own plot**.

6. PART IV: AI Machine Learning: Supervised: Linear Regression (30 Points)

TO-DO

--) Build, train, and test a supervised machine learning model on the **full** dataset **housing_boston.csv,** i.e., **all** the **14 attributes,** using the **linear regression** algorithm with Python library Scikit-Learn in a Jupyter Notebook document.

IMPORTANT NOTES:
*--) The student is expected to complete all the steps, including the **dataset introduction**, the data preprocessing, EDA of the dataset, etc.*

--) Make two new records with **reasonable** values of all the predictors of the full dataset. Each record represents the house-pricing data of a new community in the Boston area.

--) Use the trained machine learning model to predict the "Median value of owner-occupied homes in 1000 dollars" for these two new records.

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

--) **Add a section** to the above MS Word document ("**ADTA5340_midterm.docx**") to report and discuss the results of each significant step:

- Building the model
- Train the model
- Making up two new records (presenting the value of each attribute)
- Predicting the median housing prices of the new record and interpreting the results.
- Evaluating the model
- Interpret the prediction results again based on the results of evaluating the model

SUBMISSION REQUIREMENT #4:

IMPORTANT NOTES:
*--) **Write a report** (in the cell immediately following the cell in which the data set is visualized) to discuss what insight into the dataset can be obtained from each of the visualizations of the data set.*

--) The student is expected to provide all necessary comments for the code in each cell.

--) Show the work in a native Jupyter Notebook document.
--) Each step is coded in one cell.
--) Run the code of each step to show the results.
--) Report and discuss the results of each major step (in the **MS Word** document)

7. PART V: AI Machine Learning: Supervised Logistic Regression (30 Points)

TO-DO

--) **Build** and **train** a supervised machine learning model on the **clean** dataset **loan_approval.csv** using the **logistic regression** algorithm with Python library Scikit-Learn in **another** Jupyter Notebook.

IMPORTANT NOTES:
*--) The student is expected to complete all the steps, including the **dataset introduction** (as seen in the lectures on the dataset housing_boston.csv), the data preprocessing, EDA of the dataset, etc.*

--) **Make two new records** with **reasonable** values of all the predictors of the dataset. Each record represents the predictor data of a new loan application.

--) Use the trained machine learning model to **predict** the outcome of these two loan applications, i.e., whether they are approved or not.

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

--) **Add a section** to the above MS Word document ("**ADTA5340_midterm.docx**") to report and discuss the results of each significant step:

- Building the model
- Testing the model
- Making data for two new loan applications (presenting the value of each attribute)
- Predicting the outcome of the new applications and interpreting the results.
- Evaluating the model
- Interpret the prediction results again based on the results of evaluating the model

SUBMISSION REQUIREMENT #5:

IMPORTANT NOTES:
*--) **Write a report** (in the cell that immediately following the cell in which the data set is visualized) to discuss what insight into the dataset can be obtained from each of the visualizations of the data set.*

--) It is expected that the student provides all necessary comments for the code in each cell.

--) The **clean** dataset **loan_approval.csv** after preprocessing the data
--) Show the work in a native Jupyter Notebook document.
--) Each step is coded in one cell.
--) Run the code of each step to show the results.
--) Report and discuss the results of each major step (in the **MS Word** document)

8. HOWTO Submit

Due date & time: 11:00 PM Saturday 04/13/2024 – Total: 100 points

The student is required to submit the midterm – the **MS Word** documents, **Jupyter Notebook** documents, and the **cleaned dataset**, i.e., the *.csv file – as attachments to a UNT email sent to the instructor (Thuan.Nguyen@unt.edu).

The subject of the email must be: "ADTA 5340: Midterm Assessment – Submission."