Assignment 2 (CA2: 40%)

The objective of the assignment is to learn unsupervised learning and deep learning.

Guidelines

1. Submit your code and in a compressed package (zip file).

2. Students are required to submit their assignment using the assignment link under the Assignment folder.

3. The normal SP's academic policies on Copyright and Plagiarism applies. Please note that you are to cite all sources. You may refer to the citation guide available at: http://eliser.lib.sp.edu.sg/elsr_website/Html/citation.pdf

Submission Details

Deadline: August 6, 2021, 23:59H
Submit through: Polymall

Late Submission

50% of the marks will be deducted for assignments that are received within ONE (1) calendar day after the submission deadline. No marks will be given thereafter.
Exceptions to this policy will be given to students with valid LOA on medical or compassionate grounds. Students in such cases will need to inform the lecturer as soon as reasonably possible. Students are not to assume on their own that their deadline has been extended.

PART A: UNSUPERVISED LEARNING (40 marks)

Background

a) Given the iris dataset, if we knew that there were k types of iris, but did not have access to a taxonomist to label them: we could try a clustering task: split the observations into well-separated group called clusters.

Dataset

Use the iris dataset from scikit-learn

Tasks

1. Write the code to solve the clustering task. Normally you would be using scikit-learn, but if you'd prefer to work with your own implementation of learning algorithms, or some other toolkit, that is fine.
2. Write a short report (e.g. in Jupyter Notebook) detailing your implementation, your experiments and analysis.
3. Test your clustering with different possible values of k
4. Determine the best possible value of k. And show how you are able to determine that this is the best value for k.
5. Use more than just one clustering (k-means) algorithm.
6. Create a set slides with the highlights of your Jupyter notebook report. Explain the unsupervised machine learning process, model building and evaluation. Write your conclusions.

PART B: DEEP LEARNING (50 marks)

Background

Implement an image classifier using a deep learning network. [Hint: You may wish to refer to papers on successful DL architectures such as AlexNet]

Dataset

You are to use the MNIST dataset.

Tasks

1. Write the code to solve the prediction task. Normally you would be using TensorFlow/Keras, but if you'd prefer to work with your own implementation of learning algorithms, or some other toolkit, that is fine.
2. Write a short report (e.g. in Notebook) detailing your implementation, your experiments and analysis. In particular, we'd like to know:

   - How is your prediction task defined? And what is the meaning of the output variable?
   - How do you represent your data as features?
   - Did you process the features in any way?
   - Did you bring in any additional sources of data?
   - How did you select which DL model to use?
   - Did you try to tune the hyperparameters of the learning algorithm, and in that case how?
   - How do you evaluate the quality of your system?
   - Can you say anything about the errors that the system makes? For a classification task, you may consider a confusion matrix.
   - Is it possible to say something about which features the model considers important? (Whether this is possible depends on the type of classifier you are using)
   - Provide a reference section for any papers, online articles, books, publications that you have referenced.

3. Create a set of slides with the highlights of your Jupyter notebook report. Explain the entire machine learning process you went through, data exploration, data cleaning, feature engineering, and model building and evaluation. Write your conclusions.

Submission requirements

1. Submit a zip file containing all the project files (Jupyter notebook), all data sets used, and the slides (PPTX or pdf).
2. Submit online via the Assignment link.

Evaluation criteria:

Application of suitable algorithms	20%
Suitable evaluation of algorithms	20%
Background research	20%
Presentation/Demo	20%
Quality of report (Jupyter)	20%

PART C: DEEP LEARNING CHALLENGE (10 marks)

Background

Implement an image classifier using a deep learning network. Image classification is very useful for E-commerce and healthcare providers because it allows images to be automatically placed in the correct shopping aisle or automatically tag unlabelled images such as possible detection of X-rays of COVID-19 patients. [Hint: You may wish to refer to papers on successful DL architectures such as AlexNet]

Dataset

Find **ONE** suitable industrial data set from Kaggle (e.g. the **Fashion MNIST** dataset) or other public repositories. Another example would be the COVID-19 lung X-ray images.

Tasks

- Similar to Part B
- No need for presentation
- Submit slides/ppt and Jupyter i.pynb files

— End of Assignment —