Unit: Advanced Mathematics and Statistics for Data Science and AI
Assessment Element1: In-Class Assignment

Note: This is part 2 of Element1 aimed to check the knowledge of later part of the lecture after Week 5

Housekeeping Information

Before you start, I would recommend you please read the whole document once including the caution note, before you dive in straight.

Assessments: There are two Elements where Element 1 is In Class Assignments and Element 2 is Practical Exam

Element 1: In Class Assignments

A collection of assignments applying concepts and skills from the unit, as specified in the Assessment Brief (50%).

Element 2:

Practical Exam You will be individually asked to generate a report integrating text, code and visualisations (e.g., in the form of a Jupyter notebook) detailing statistical insights given a previously unseen dataset (50%).

Submission Date:

Element 1: In Class Assignments:

By 6pm (1800) GMT on Monday 4th December 2023

Element 2: Practical Exam:

By 6pm (1800) GMT on Friday 8th December 2023. Open 7 days prior to deadline; open from Friday 1st December 2023.

Submission Method:

Element 1: In Class Assignments:

A Zip folder (100MB max) containing:

1. a *link to your GIT code repository,*
2. a link to a video (with voice over/detailed captions) demonstration of your work and
3. *Readme* file to contextualise the overall design and development that includes a 300-word mini project description written in your style highlighting on
    a. What is the task about?
        i. Understanding of the bigger picture and each sub task,
    b. What Maths and Statistics involved?
        i. Identifying and applying the appropriate concepts learnt
    c. How did you implement?
        i. Possible libraries, packages involved.
        ii. Logic used and why?
        iii. Cite the references used
    d. Outcomes of each sub task
        i. Interpret and provide detailed explanation about the arrived results at each step
    e. Challenges and How you resolved?
        i. Challenges if any and how it was overcome

f. References
    i. Acknowledge by listing all of the resources that helped to complete the task

4. The GIT repository must include
    a. the programming tasks,
    b. dataset used,
    c. a link to a video demonstration of your work, and
    d. Readme file to contextualise the design and development.

Element 2: Practical Exam: Zip folder via Moodle:

A Zip folder (100MB max) containing your portfolio of works, along with a link to your implementation notebook with necessary comments, exported as PDF and submitted in the Zip.

Element 1: In Class Assignments

Please find below the task for second part of the Element 1 In-Class Assignment
1. This assessment will have a series of smaller sub-tasks to be completed.

Dataset: Should use(choose) In Class Assignment Element1 datasets
Plagiarism: Avoid Plagiarism Essentials
Preferred Language: Python
Platform/IDE: Any

Overall Task brief:

For this In-Class Assignment, you may have to complete the below tasks on all the possible and relevant datasets.

1. For each dataset, Identify the a. type (eg: Linear/NonLinear; Single/Multilabel) and b. tasks possible on the dataset (Classification/Regression) and justify your answer with the following evidences
    a. Inspect and report based on the type of variables based on your basic domain knowledge/context
    b. Any one exploratory analysis technique
    c. Any one Inferential analysis technique
    d. Any one predicative analysis technique
2. Once you are confident of the above step apply the following loss functions appropriately only on the (all possible) appropriate datasets
    a. L1 loss
    b. L2 loss
    c. Log loss
    d. Categorical cross entropy loss
    e. Hinge loss
3. For each dataset provide appropriate visual plots to show a comparison of better loss function in case there was possibility to apply more than one loss function.
4. Do assess with appropriate metrics based on Classification/Regression applied on the datasets. (Hint: Not limited to accuracy, R^2, Precision, Confusion matrix etc….)
5. Also choose any one non-linear dataset from the datasets provided and try any kernel transformation to linear space and then fit model and assess accuracy.
6. Choose any one suitable dataset and perform the following:
    a. Create a scenario for Overfitting in the context of regression (Hint: You can emulate this either by adjusting the subset of features or the size of the training dataset to create scenarios where overfitting is more likely to occur)

b. Prove the overfitting with evidence (Hint: Metrics and Plots)
c. Now apply any two regularization methods and evaluate performance before and after Regularization (Hint: Metrics and Plots)

7. Choose any one suitable dataset and perform the following:
    a. Create a scenario for Overfitting in the context of Classification (Hint: You can emulate this either by adjusting the subset of features or the size of the training dataset to create scenarios where overfitting is more likely to occur)
    b. Prove the overfitting with evidence (Hint: Metrics and Plots)
    c. Now apply any two regularization methods and evaluate performance before and after Regularization (Hint: Metrics and Plots)
8. Perform the below task on MASTER_PhonesmartdataAll_CCI_AdvStats.csv and wine dataset
    a. Apply Decision Tree on both of it without and with pruning and record your observations.

Caution Note: *Should provide sufficient comments throughout the Assignment. Should credit every reference used. Having said that too much of external code use may lead to loss of grade, care to be taken to show case originality of work. You are allowed to develop the necessary logics to restructure the dataset to complete task without distorting the meaning of the dataset (eg: Adding necessary columns, Transposing, Grouping etc.,). If this step is carried out upload this dataset to the git else the original would do. But should be supported with appropriate comments and justification, else it will be considered deviation from the expected task leading to loss of grade.*