CARNEGIE MELLON UNIVERSITY
DATA INFERENCE AND APPLIED MACHINE LEARNING (COURSE 18-785)
ASSIGNMENT 5

INSTRUCTIONS

- Submissions should be made via canvas.
- **Single** Python/MATLAB code file(.ipynb or .m) **[Do not Submit checkpoints for .ipynb]**. In addition, each line of code should be documented by text. This demonstrates that the code is unique and owned by the student.
- Assignment report(.pdf) with full evidence that the assignment was completed by the student and demonstrate a full understanding of each step in the process including textual descriptions of each result (statistics, table, graph etc) represents and insights that can be gained.
- Indicate the libraries you have used in your code at the beginning of the report (After the title page).
- Using ChatGPT for any assignment is not allowed as it could lead to being flagged for plagiarism.
- Data files (as given).

Submission process:

1. Put source code **file and data files** in a single folder
2. Name of the folder should be the same as your andrew ID
3. **Zip this folder and attach the zipped file on assignment submission page (CANVAS)**
4. After attaching zipped file, click on "Add Another File" from assignment submission page and **attach your report**
5. Submit your assignment

N.B. This process will allow us to compile your reports in **Turnitin** to check for plagiarism.

Specific reasons for a submission being classified as incomplete include:

- Failure to correctly name your folder with your Andrew ID
- Failure to correctly name your report, and code file with andrewID_DIAML_AssignmentNo. For example, mcsharry_DIAML_Assignment1, mcsharry_DIAML_Assignment2 and mcsharry_DIAML_Assignment3.
- A missing report describing the steps, results, and insights
- A missing dataset required for running the code
- A missing code file such as .ipynb or .m file
- An error in the file path needed to run the code

The student is responsible for checking that their submission is complete. Students will lose 10% as for late submission even if the submission is repaired during the 24 hours after the deadline has passed, and receive 0 for the assignment if it is not repaired.

The submission deadline is **on Monday 06, November, 2023 16:59 Eastern Time (ET) /**

Monday 06, November, 2023 23:59 Rwandan Time (CAT).

1. Statistical learning (25 points)

1.1 Describe at least four steps to implementing a rule-based approach to decision-making and give an example. Is any domain knowledge required to establish a rule? Support your answer with an explanation.

1.2 Explain over-fitting and why it is a problem in statistical learning. If you have a small dataset containing ten data points, should you prefer a simple model with one parameter or a complex model with ten parameters? Support your answer with an explanation.

1.3 There are two commonly used approaches to avoid over-fitting; describe each one.

1.4 Provide two examples of metrics used to evaluate the performance of a model and give a formula for each one. Give two examples of applications and appropriate metrics for each case.

1.5 Why are benchmarks useful in machine learning and give two examples.

2. Machine Learning (25 points)

2.1 What is machine learning? Discuss its evolution over time and why is it popular?

2.2 Give three examples of machine learning techniques that can be viewed as either supervised or unsupervised approaches.

2.3 What is the difference between classification and regression?

2.4 What is the difference between supervised learning and unsupervised learning?

2.5 Give examples of successful applications of machine learning and explain what technique is appropriate and what type of learning is involved?

3. Diabetes data (25 points)

In this assignment we will be looking at data analyzing diabetes patients (N = 442). The data consists of 11 columns. These columns are AGE, SEX, BMI (body mass index), BP (average blood pressure), S1, S2, S3, S4, S5, S6 (the last six are blood serum measurements looking for example at the Glucose level in blood). The 11th column is the dependent variable y, a quantitative measure of disease progression 1 year after baseline (one year after incurring the disease).

3.1. Load the diabetes data into MATLAB or Python from here. Produce a correlation matrix of the explanatory variables. Make a heat-map of the matrix (using imagesc and colorbar) and describe the relationships between the variables.

3.2. What is collinearity? What effect does collinearity amongst predictor variables have on their estimated coefficient value?

3.3. Create a multivariate linear model using all ten variables and a constant. In the rest of this assignment this model will be referred to as model1. What are the Mean Squared Error and the adjusted R^2 for model1? Are all variables significant? Could this be a problem of collinearity?

3.4. What is the difference between forward selection and backward selection?

3.5. How does the approach stepwise work in the sense of selecting variables? Use the function stepwise to interactively compose a model using forward selection. Which variables are selected? How does this function work? What is the MSE and R^2 value for this new model?

4. Analyzing the Titanic data set (25 points)

The titanic3 data frame describes the survival status of individual passengers on the British passenger liner RMS Titanic when sunk. The data frame does not contain information for the crew, but it does contain actual and estimated ages for almost 80% of the passengers. More details about the data frame can be found on the file *titanic3info.txt.* For this assignment, use the titanic dataset stored in *titanic3.csv*. You can get the dataset from this link

4.1 What is the difference between logistic regression and linear regression?

4.2 Load in the titanic dataset and calculate the probability of survival for a passenger on the titanic.

4.3 Provide a table giving survival probabilities broken down by passenger class, gender and age.

4.4 Build a logistic regression model for survival rates based on passenger class, sex and age.

What are the parameter estimates and are these parameters statistically significant?
4.5 What is the performance of the model, measured by classification accuracy (number of correct classifications divided by total number of classifications) based on confusion matrix?

Extra credit: You are encouraged to enter the Kaggle challenge referencing this data set. At the end of this course, extra-credit will be given to students based on their final score on the challenge, coinciding with the deadline for the final assignment. Go to this link https://www.kaggle.com/c/titanic-gettingStarted and follow the instructions to register and enter the challenge.