CARNEGIE MELLON UNIVERSITY
DATA, INFERENCE & APPLIED MACHINE LEARNING (COURSE 18-875)
ASSIGNMENT 6

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 student completed the assignment 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 still needs to be repaired.

The submission deadline is **Monday 18th, November, 2024 16:59 Eastern Time (ET) / Monday 18th, November, 2024 23:59 Rwandan Time (CAT) .**

1. Nonlinearity (25 points)

1.1 Can you explain what is nonlinearity and why might it be necessary to consider nonlinear relationships between variables? (5)

1.2 Write down the mathematical equation for a nonlinear model and provide an example of an application where it might be appropriate. (5)

1.3 Can a nonlinear model be more parsimonious than a linear model? Write down mathematical formulae for both the linear and nonlinear models to support your answer. (5)

1.4 Surrogate data are used for testing for nonlinearity. What characteristics are typically preserved when generating surrogates? Give the names of two surrogate techniques and describe the approaches for implementing them.(5)

1.5 Define information, entropy and mutual information and also provide mathematical formulas. Describe how entropy can be used for constructing a feature for measuring regularity and give an example of an application. Explain how mutual information can be used for feature selection and why it might be better than correlation. (5)

2. Classification using trees (25 points)

2.1. Decision trees are often used to transform a set of observations into a specific recommended action. Describe the components (nodes, branches) of a decision tree. What is pruning and why might it be necessary to prune the tree? Why are decision trees an attractive method for classification in practical applications?(5)

2.2. Suppose an organization has built a rule-based classifier using domain knowledge. After collecting a large amount of data, outline the steps required to improve upon the existing approach by constructing a data-driven classifier. How would you advise to test the validity of the new model?(5)

2.3. Consider the challenge of classifying the likelihood of survival using the Titanic dataset. Construct a decision tree and display the structure of this tree using a graphic. (5)
2.4. Evaluate the performance of the tree (before and after pruning) and provide results using cross-validation. (5)

2.5. Compare the performance of the final tree with logistic regression and comment on the advantages and disadvantages of both. Which model is best for competing in the Kaggle competition? (5)

3. Classification using KNN (25 points)

3.1 By focusing on small neighborhoods of state space it is possible to construct parsimonious models. Describe the concept behind this general approach and a step by step procedure for implementing such a model. (5)

3.2 Consider the challenge of classifying the likelihood of survival using the Titanic dataset. In order to construct a KNN classifier, how will you transform the available variables? (5)

3.3 Calculate the performance of the classifier versus the number of neighbors used and provide a graphic to display the result. What is the optimal number of neighbors using cross-validation? (5)

3.4 Explain why some distance metrics are sensitive to the kind of features used.
Evaluate the performance using 3 different distance metrics. (5)

3.5 Compare the best KNN classifier with logistic regression and comment on the advantages and disadvantages of both. Which model is best for competing in the Kaggle competition? (5)

4. Regression – wine quality (25 points)

The wine quality database provides information about the quality of wine. There are two datasets, one for red wine and one for white wine, which contain quality ratings, from one to ten, along with their physical and chemical properties. The challenge is to use these features to predict the rating for a wine and to assess performance. It is advisable to study white and red wine separately:
You can get the datasets using this link

4.1 Calculate the average of each feature for the red and white wines separately and make a comparison using a bar graph showing the two wines together. How do the results relate to common sense (or the intuition of a wine expert) based on the features that are available?(5)

4.2 What is the correlation between each feature and the dependent variable using a separate analysis for white and red wine? Which variable is most relevant for each wine?(5)

4.3 **Use Lasso and cross-validation to provide a plot of MSE against lambda and the parameter estimates versus lambda for white and red wine.** How do the features selected by LASSO compare with an approach of setting a threshold on the absolute correlation coefficient? (5)

4.4 Use the features identified by LASSO to construct a **KNN regression** model for red wine. (5)

4.5 What is the performance of a linear regression model and the KNN model, measured by MSE and R^2? Describe the advantages and disadvantages of both models.(5)

Extra credit: You are encouraged to enter the Kaggle challenge referencing this dataset. 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. After this assignment, you should have two new approaches for classifying survival on the Titanic: a tree and a KNN model.