CS 5565
Intro to Statistical Learning: Final Project

Objectives and **Deliverables**:

1) You will **reasonably attempt** to **optimize** the model for each of the methods requested.
2) You will provide **an explanation (min 2-3 sentences)** of the output **for each model**.
3) All methods should implement a **train** / **test** split. (70/30 is a good **starting** point*)
4) All results should **evaluate** your model using the testing data.
5) Provide at least one paragraph of **analysis** detailing the **accuracy** and **reliability** of your models for each task.
6) Attach the **output** of the summary for each of the models.
7) You should generate and include a **relevant image** or **screen shot** for each of your models. (This should be a **linear plot, ROC, Confusion Matrix, or diagram** of your results for each of the models.)
8) Set all generative seeds as the last 4 digits of your student ID [this will replace the rand value used for obtaining lambda and other tuning parameters.

** split ratios can be adjusted to evaluate the model when trained using a different amount of the total set. Increasing testing data will give a better evaluation of future performance, while at the same time limiting the model's access to representation data (it's always a trade-off).*

Instructions:

Section A:

Part 1: Regression

Complete the following objectives utilizing the data set you selected as ideal for**: Linear Regression**

Note: you may use the exercise from Chapter 3 as a reference **link here** --> Chapter 3 (Linear Regression).

1) Pick a set of feature value(s) you prefer. Be consistent such that you **keep** the *previous* features when adding more for the higher dimensional models. Each model should be fit to the same **response/target**.

a) Linear Regression

b) Polynomial Regression

c) Multi-Linear Regression

Part 2: Feature Selection / Model Optimization Methods

Complete the following objectives utilizing the data set you selected as ideal for**: Feature Selection / Model Optimization Methods**

1. Perform a Forward Stepwise Selection and a Backward Stepwise Selection, for reference Chapter 6 lab, change the direction for backward selection. You may utilize K-fold methods to compare ideal selections for each case. If you are using categorical data, make sure you convert the source data using the appropriate function, and that you are using the correct evaluation metrics.

2. Using the models generated for the feature selection, generate the plots of RSS and Adjusted R^2 for forward and backward features (as given in the exercise for chapter 6).

3. PCR
   Perform a PCR on your selected data as detailed in Chapter 6 lab. Most of the datasets should have sufficient features to make this task interesting. Generate a plot of the components relative to their fit target.

Part 3: Classification

Complete the following objectives utilizing the data set you selected as ideal for**: Classification**

Note: you should use the exercise from Chapter 4 and 5 as a reference
link here --> Chapter 4 (Classification)Links to an external site. & Chapter 5 (Sampling)Links to an external site.
Generate **two** Classification Models for based on your given data.
These models will include Logistic Regression, Linear Discriminant Analysis, as given in the chapter 4 lab exercise.
Note: You'll need to make sure you **properly encode any categorical data** for your classes or predictions.

Continued [Section B] on next page:

Section B:

Part 1: Splines

Complete the following objectives utilizing the data set you selected as ideal for**: Natural and Cubic Splines**

Note: you may use the exercise from Chapter 7 as a reference **link here** --> Chapter 7 (Splines)

- **Polynomial Regression and Step Functions:**
  - Try out different degrees and deliver the changes observed through discussion/comments.
- **Splines**
  - of Basis-Spline and Natural-Spline (as given in the lab) and **two new fits** of **df 9** and **22**. (substitute the values as given in the assignment) What would you conclude from the results generated from the fits given by df 16,22 compared to the default in the lab for both the B-spline and N-spline?

Part 2: Tress and SVM

Complete the following objectives utilizing the data set you selected as ideal for**: Trees and SVM**

Note: you may use the exercise from Chapter 8 and 9 as a reference **link here** -->

Chapter 8 Lab , Chapter 8_R_Lab / Chapter_9_Lab , Chapter_9 R Lab

- **Trees**
  - Generate a Tree Classifier or Regressor for the classes and predictors of the chosen dataset.
  - Perform Regression/Classification Tree, Random Forest, and Regression/Classification Boosting based on the dataset chosen.
- **Support Vector Machines**
  - Perform SVC and SVM for the multiclass classification with outputs resulting in decision boundary plots and confusion matrices.

Part 3: Neural Nets

Perform Neural Nets on image classification task on **MNIST** dataset, loaded as the part of the provided script. Use the script provided to perform the required tasks below and the instructions to achieve the task would be available as comments:

Set the seed with the last 4-digits of your Student ID.

- Try out the topology [neural net architecture] provided in the script [256,96,32,10] and implement the following **topologies** using the '**sigmoid'** activation function:
    1. [96, 32, 10]
    2. [128,64,32,10]

  ○ Remember to **add** / **remove** the layers as per the requested topologies.
- Repeat, the three topologies above with a different **activation function** – '**relu'**.
- Now, out of the [2x3 = 6] total experiments select the **best** performed topology with activation function for the further process (Look out for overfitting!).
- Hereon, with the chosen model with best performing topology and activation function – perform the same task with change in **learning rate** different from the provided script originally. The learning rates to try are:
    1. Learning rate to try with [0.01, 0.001, 0.0001]
    2. Out of the 4 learning rates tried [default+3 new] select the learning rate with which the best-chosen performed better.
- Now, we have the best performing topology, activation function, and learning rate so far, now as the final trail try with different batch size different from the provided script originally. The batch sizes to try with are:
    1. [32, 64]

Provide the configurations tried using the task sheet provided as part of the instructions and paste the table before answering the below questions into the document you are supposed to submit.

Provide the final best topology, learning rate, activation function, and batch size again the MNIST dataset and provide your views on the questions below:

- Compare the topologies and provide the reason the chosen topology worked out to be the best along with the activation function?
- Does topology and activation function depend on each other? If yes, then how and why?
- What learning rate felt like the best choice and why?
- What impacted the batch size on the performance of the model?
- Are the learning rate and batch size correlated? What notable changes were observed during the entire process?

Script Link: MNIST_FC_ANN