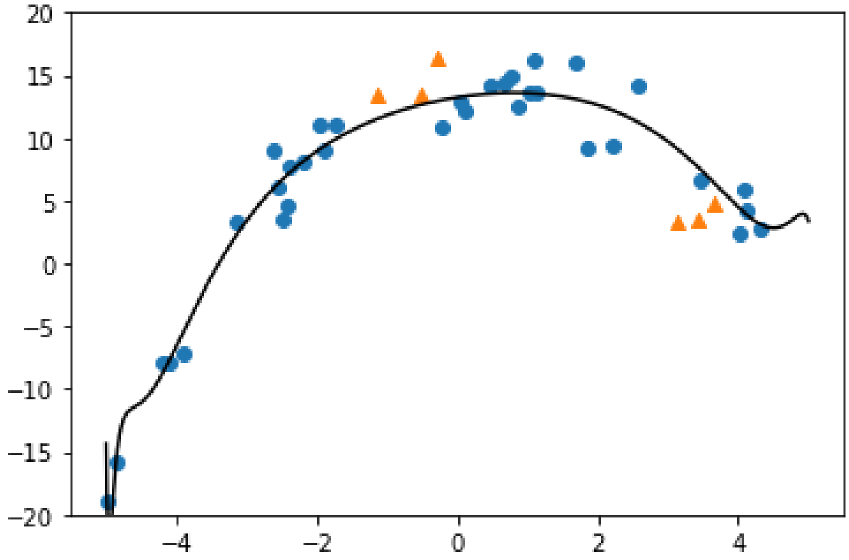
# COMP-2704: Supervised Machine Learning



Assignment 2: Polynomial Regression

## Setup

Complete the following steps to setup this assignment.

i. Open Jupyter Notebook in a web browser.
ii. Within the "SupervisedML" folder, create a folder named "Assignment2" and navigate into this folder.
iii. Open the link: https://github.com/luisguiserrano/manning/tree/master/Chapter_4_Testing_Overfitting_Underfitting
iv. Download the file:
    a. *Polynomial_regression_regularization.ipynb*
v. Upload this file to the folder you created "~/SupervisedML/Assignment2".
vi. Open the *Polynomial_regression_regularization.ipynb* notebook and run all code cells. Fix any errors that occur.

## Problem

Create a notebook with filename *SML_a2.ipynb* within the folder "~/SupervisedML/Assignment2". Add cells with markdown text to the notebook to complete the following steps. Add text cells to answer the questions. You may copy relevant lines of code from *Polynomial_regression_regularization.ipynb*.

First, import the necessary libraries and modules. Then:

1) *[2 marks]* Create an array of coefficients named *coefs* to define a new polynomial

$$f(x) = (a + 1)x^3 + (b + 1)x^2 + (c + 1)x + d$$

where $a$ is the average last digit of your group's student numbers, $b$ is the average second last digit, $c$ is the average third last digit and $d$ is the average fourth last digit. Use the polynomial to generate 500 rows of x-y data (with a random component added to the y-values) and provide a scatter plot.

2) *[2 mark]* Add columns to the dataset to hold powers of $x$ up to $x^n$ and set $n=10$. Split the data so that 70% is used for training, 20% for validation, and 10% for testing.

3) *[3 marks]* Create a new model called *model_elastic_reg* that uses an L1 penalty of 0.1 and an L2 penalty of 0.1. (Using both L1 and L2 penalties is known as elastic regularization.) Write a modified *display_results()* function that uses the training and validation data to calculate errors and use it to display the results.

4) *[2 marks]* Compare the training and validation error of *model_elastic_reg*. Comment on what this implies in terms of overfitting and/or underfitting.

5) *[5 marks]* Create a new model named *optimal_model*. Using nested for loops, write code to try a range of values for $n$, L1, and L2, using three different values for each

hyperparameter. (This is called a *grid search*.) For each combination of hyperparameters, calculate and display the training and validation RMSE and provide a labelled scatterplot. (Number each model with an index and use this as the *x*-coordinate for your scatter plot. For each index, plot both the training and validation RMSE.)

6) *[2 marks]* By referring to training error, validation error, complexity, and scatter plots, select the best set of parameters for *optimal_model* and justify your choice.

7) *[2 marks]* Use the modified *display_results()* function to display the results of *optimal_model*. Compare this with the results of *model_elastic_reg*, and explain which model is better. After selecting the best model, use that to calculate the error on the testing data.

8) *[2 marks]* With the exception of import and print statements, add a comment before each line of code in *SML_a2.ipynb* to explain what it does.

## Contributions

Using markdown at the end of the notebook, list the contribution of each student to the assignment, referencing specific question numbers and other tasks such as formatting and submitting. Ideally, both students will contribute to all questions.

## Submission

Upload your notebook to the Assignment 2 dropbox on the course website. Late submissions will lose 10%.

**Total marks = 20**