Math 5700 Project: Data Creation and Model Evaluation

1. Create Some Data

For this project, you will generate your own synthetic dataset. Follow the guidelines below:

Data Requirements

- Create at least 300 observations.
- Include one response variable, denoted by y.
- Include at least five input variables/features, labeled x_1, x_2, ..., x_5.

Feature Construction

Your input variables should include variety, such as:

- Generating features over different intervals
- Drawing features from different probability distributions

Response Construction

The response variable y should have some connection to the input features, but not a perfect one. This means you should introduce noise. At least one of the variables should be left out of the model.
Example model:
$y = 7x_1 + 3(x_2)^2 - 5x_3 + \varepsilon$,
where ε is random Gaussian noise.

Train/Test Split

- Use 80% of your data as the training set.
- Use the remaining 20% as the test set.

2. Fit a Model Using Least Squares

(a) Fit the Model

Use least squares to fit a linear model:
$\hat{y} = w_0 + w_1x_1 + w_2x_2 + \ldots + w_5x_5$
Fit the model using only the training data.
Recall this uses the formula $(A^T)Ax = (A^T)b$,
which you can solve directly using MATLAB or
python.

(b) Compute Training Error

Compute the training mean squared error (MSE):
$MSE_{train} = (1/n_{train}) \Sigma (y_{train} - \hat{y}_{train})^2$

(c) Compute Test Error

Compute the test MSE using the held-out data:
$MSE_{test} = (1/n_{test}) \Sigma (y_{test} - \hat{y}_{test})^2$

3. Add Flexibility and Repeat

Fit a more flexible model by adding higher polynomial terms or other feature transformations. For example:

$\hat{y} = w_0 + w_1x_1 + w_2x_2 + w_3(x_2)^2 + \ldots$

Repeat steps (b) and (c) for this more flexible model. Compare training and test MSE.

Do this for a few different, but increasingly more flexible, models.

At the end, you can show a graph of the error versus the degree of flexibility, and discuss the impact of model flexibility and overfitting.