CS 5834: Urban Computing
Fall 2025
Homework 3
Date Assigned: Sep 30, 2025
Date Due: Oct 9, 2025

1. (20 points) You are training a logistic regression model and you notice that it does not perform well on test data. Could the poor performance be due to underfitting? Explain. Could the poor performance be due to overfitting? Explain.
2. (15 points) Which of the following types of curves can be estimated accurately from a given dataset (assuming it satisfies the functional form) using linear regression methods? Here y is the dependent variable, and x is the independent variable. “Estimate accurately” means if we substitute values for the coefficients in the curves below, generate data, and learn a regression model from that data, the model will recover the values we used to generate the data in the first place.
   a. $y = \sum_{i=0}^{n} a_i x^i$
   b. $y = ax + b \cdot sin(x) + c \cdot log(x) + d$
   c. $y = ae^{(bx)}$
   d. $y = ax + bx + c$
   e. $y = (ax + b)/(cx + d)$

3. (20 points) Consider the following data and assume you are trying to fit a linear model $y = ax_1 + bx_2 + c$. Derive least squares estimates for a, b, and c.

x_1	x_2	y
0	0	0
0	1	1.5
1	0	2
1	1	2.5

4. (45 points) Consider the Boston Housing dataset available at: https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html. The goal is to predict MEDV using the other variables. Take some time to explore the data and understand its distributions. See which variables have predictive utility.

a. (15 points) Implement linear regression using a subset of features that you have identified.
b. (15 points) Implement ridge regression.
c. (15 points) Implement the LASSO algorithm.

For full credit, draw plots, give statistics, explain in detail your conclusions, and compare the above algorithms.

What to turn in:

- A PDF containing answers to all the above questions. The PDF should contain a hyperlink to any code you have written or generated for possible perusal and evaluation.