CS 5834: Urban Computing
Fall 2025
Homework 4
Date Assigned: Oct 14, 2025
Date Due: Oct 23, 2025

1. (10 points) Between ridge regression and the LASSO, which regularization method is more likely to help interpretability, and why?

2. A city's police department wants to classify each street segment as "high risk" or "low risk" for property crime using features such as: land-use mix, population density, lighting index, nearby bar density, prior-week incident counts, and socioeconomic variables.

   Both a decision tree and a logistic regression model are trained. The decision tree highlights "presence of bars" and "lighting index" as dominant splits. The logistic regression assigns moderate weights to several correlated social variables (income, unemployment, population density).

   (10 points) Explain conceptually why the decision tree may appear to overemphasize a few discrete features compared to logistic regression.

   (10 points) How can logistic regression better represent the combined influence of correlated or interacting social factors, and what are its limitations in doing so?

   (10 points) If you wanted a model that captures nonlinear interactions *and* remains interpretable for urban policy decisions, what modeling strategy (i.e., decision trees or logistic regression?) would you use?

3. Sensors in office buildings record temperature, CO_2, sound levels, and Wi-Fi access counts every 5 minutes. You need to predict whether a given room is occupied or vacant to optimize HVAC scheduling. Let us suppose you train a regression model using one week of sensor data and evaluate it on the following week.

   (5 points) Explain why this prediction task may be difficult even though the input variables are continuous and seemingly informative. What sources of noise or variability might reduce accuracy?

   (10 points) Suppose you simplify the model (e.g., by pruning, regularizing, or limiting depth/number of coefficients) and observe slightly lower training accuracy but higher test accuracy. Explain why this might be.

4. Consider the Metro Interstate Traffic Volume Dataset from: https://archive.ics.uci.edu/dataset/492/metro+interstate+traffic+volume. Download this data and take some time to understand its format and contents.

Our goal is to setup an ML problem where we predict traffic volume based on weather conditions, holiday flag, hour of day, temperature, visibility, wind speed, precipitation, etc.

(15 points) Conduct an exploratory analysis of the data. Plot different attributes, their distributions and study how the target variable varies w.r.t. these attributes. Make qualitative conclusions about relationships you observe. Form your own understanding of which attributes are likely going to be predictive.

(15 points) Develop 3 ML regressors: Linear Regression, Random Forests, XGBoost to predict the traffic volume. Engineer time features sensibly (e.g., encode hour as cyclical; create weekend/holiday flags). Do a chronological train/test split (train on earlier dates, test on later) to avoid temporal leakage.

(15 points) Develop 3 ML classifiers by first creating a categorical target by binning the volume (e.g., into “low”, “medium”, and “high” levels). Clearly state your binning rule. Again use Logistic Regression, Random Forests, and XGBoost to predict the traffic volume class. Follow the same sensible directions as above.

Use a suitable k-fold cross-validation (decide k yourself) and evaluate your regressors and classifiers using standard measures.

Make qualitative conclusions/interpretations about which methods work well (and why), and which framing of regression or classification produced more actionable/robust results?

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.