Homework 5

Calista Harris

Task 1: Conceptual Questions

Question 1

Cross-validation helps us estimate how well a random forest model is likely to perform on unseen data by repeatedly splitting the data into training and validation sets. This process reduces variability in the performance estimate that could arise from a single train/test split and allows for more reliable model assessment. Although random forests have built-in error estimation through out-of-bag observations, cross-validation can provide a complementary or alternative method, especially when comparing multiple models or tuning hyperparameters.

Question 2

The bagged tree algorithm, short for bootstrap aggregation, involves generating multiple bootstrap samples from the original dataset, fitting a decision tree to each sample, and aggregating their predictions. For regression tasks, the final prediction is the average of individual tree predictions, while for classification, it is typically determined by majority vote. This ensemble method helps reduce variance and improves predictive performance compared to a single tree model.

Question 3

A general linear model (GLM) is a statistical model where the expected value of the response variable is modeled as a linear combination of the explanatory variables. It has the form $E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$, and is used for tasks such as multiple linear regression and ANOVA. The model assumes the errors are normally distributed and have constant variance, making it a foundational approach for modeling relationships between variables.

Question 4

Adding an interaction term to a multiple linear regression model allows the effect of one explanatory variable on the response variable to depend on the level of another explanatory variable. This enables the model to capture relationships where the combined effect of two variables is not merely additive. In mathematical terms, it adds a term like $\beta_3 X_1 X_2$ to the model, allowing the slope of one variable to change based on the value of the other.

Question 5

Splitting the data into a training and test set allows us to evaluate how well a model generalizes to new, unseen data. The training set is used to fit the model, while the test set provides an unbiased estimate of the model's predictive performance. This helps guard against overfitting and ensures that the model is not simply memorizing the training data, but instead learning patterns that can apply more broadly.