

# Homework 3

Due on 04/04/2024

In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the dataset “auto.csv”. The dataset contains 392 observations. The response variable is “mpg\_cat”, which indicates whether the miles per gallon of a car is high or low. The predictors are:

- cylinders: Number of cylinders between 4 and 8
- displacement: Engine displacement (cu. inches)
- horsepower: Engine horsepower
- weight: Vehicle weight (lbs.)
- acceleration: Time to accelerate from 0 to 60 mph (sec.)
- year: Model year (modulo 100)
- origin: Origin of car (1. American, 2. European, 3. Japanese)

Split the dataset into two parts: training data (70%) and test data (30%).

- (a) Perform a logistic regression analysis using the training data. Are there redundant predictors in your model? If so, identify them. If none is present, please provide an explanation.

- (b) Based on the model in (a), set a probability threshold to determine the class labels and compute the confusion matrix using the test data. Briefly interpret what the confusion matrix reveals about your model's performance.
- (c) Train a multivariate adaptive regression spline (MARS) model. Does the MARS model improve the prediction performance compared to logistic regression?
- (d) Perform linear discriminant analysis using the training data. Plot the linear discriminant variable(s).
- (e) Which model will you use to predict the response variable? Plot its ROC curve using the test data. Report the AUC and the misclassification error rate.