PSTAT 131 Homework 1

Matthew Zhang

Spring 2022

Machine Learning Main Ideas

Question 1: Supervised learning can be defined as a subset of machine learning that uses labeled data to map inputs to outcomes by training an algorithm. It is often inclusive of prediction, estimation, classification, etc. and adjusts the model until it is appropriately fit.

While unsupervised learning has unlabeled data which can be used to discover patterns utilized in clustering and association problems; particularly when there exists uncertainty of properties within the data set. Moreover, there are no "answers" and response. The differences can be seen within the labeled (functional training data) versus unlabeled data (raw) for supervised and unsupervised learning, respectively.

Question 2: In the context of machine learning, regression is associated with a continuous and real-valued number output, while classification is for categorical and class outputs. In other words, in regression, Y variable is quantitative such as temperature and for classification, Y is qualitative such as binary classification (spam/not spam) or multi-class. Regression gives viable solutions to questions such as "How much?" and "How many?" and classification gives viable solutions to questions regarding True/False and "Survived/Not Survived". To expand from the previous question, the field of supervised learning can be broken down into regression and classification.

Question 3: For regression ML problems, two commonly used metrics are the Mean Squared Error (MSE) which is representative of the residual error and R-Square Value which can be interpreted as the model explaining a certain percent of variation.
For classification ML problems, two commonly used metrics are the accuracy of the classification, which measures the total number of predictions a model gets right, including both True Positives and True Negatives. Second, F1 score which is a harmonic mean of both the precision and recall score and informs that for F1 to be high, precision and recall must also be high.

Question 4: -Descriptive models: Descriptive models demonstrate unsupervised approaches in summarizing and classifying previous events.
-Inferential models: Inferential models often assess quality of outputs in prediction and estimation rather than being able to predict future events.
-Predictive models: Predictive models utilizes machine learning approaches for forecasting future data based on previous data and predicts future responses given the features.

Question 5: -Mechanistic predictive models tend to use theory and concepts in order to express a real-world problem or process. Some properties include the assumption of a parametric form, adding parameters for flexibility in which too many could lead to overfitting the model. In comparison, empirically-driven models emphasize studying real-world events in the past to develop a theory of principles with characteristics like no assumptions of function f, requiring a large number of observations, are innately more flexibile, and also the possibility of overfitting. In other words the difference between the two are that, in an empirical model, observation is the focus as opposed to theory where if you observe an outcome based on a condition, then you can predict that outcome in the future; mechanistic models need only a few input data points for a prediction.

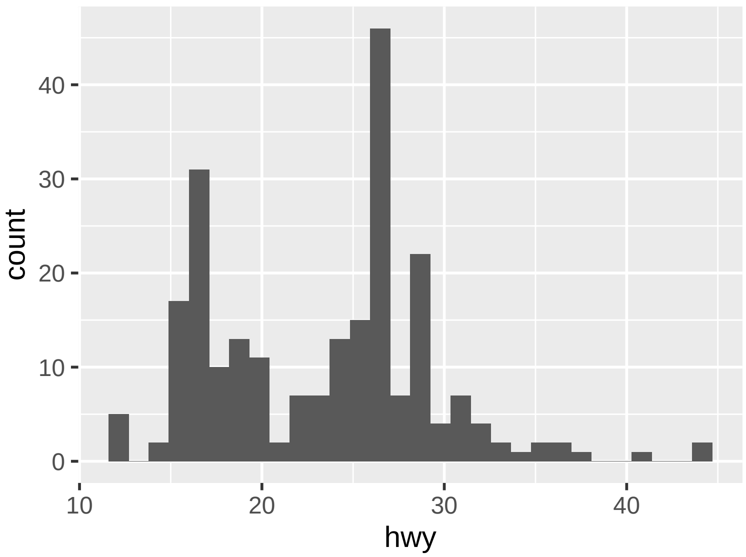
-In general, empirical models are easier to understand because they rely on simple observation and not mathematical/statistical theory formulas. For example, observing that when the sunsets, the light outside becomes dark. Yet, it is important to note that both empirical models and mechanistic models include elements of each other and that neither are the "right answer," but the utility of each model can lead to a better solution.

-The bias-variance trade-off is related to the use of empirical models and mechanistic models because it is in part, model evaluation. We think of bias as the error between average model prediction and the ground truth where we determine the ability of the model to predict values; variance as the average variability in the model prediction for the given data that tells how much the function can adjust to change in data. High bias associates with an overly-simplified model and under-fitting with high error in the training and testing sets. Like such, high variance associates with an overly-complex model and over-fitting with low error on training data and high error on test. Because we want to find the best model where the error is reduced, we look for the middle ground of these two. This trade-off in bias and variance helps us determine the empirical models and mechanistic models that have properties of overfitting and number of predictors.

Question 6 -Predictive, because based on past history, we want an algorithm to learn the patterns in the voting history to able to predict and determine who they are more likely to vote for in the candidacy.
-Inferential, because if a voter has personal contact with a candidate, we want to observe the influence or change in voter's perspective and how this may impact their decision. The "likelihood" of support may or may not change and we want to infer how it may change.

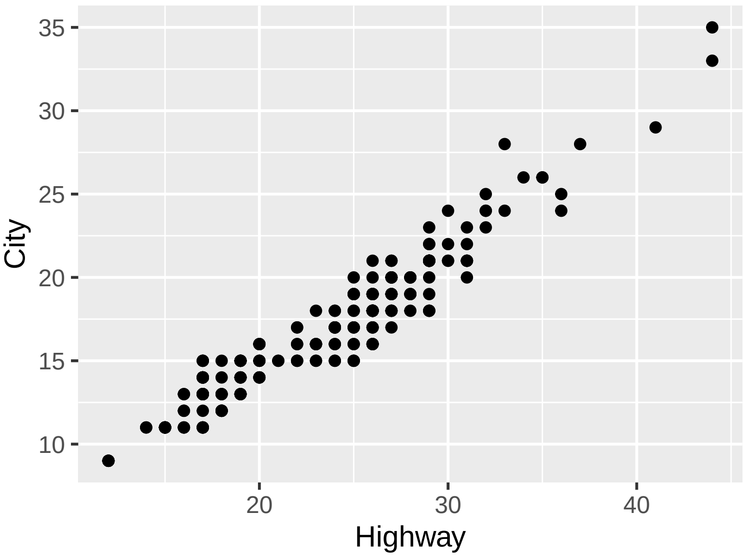
Exploratory Data Analysis

Exercise 1



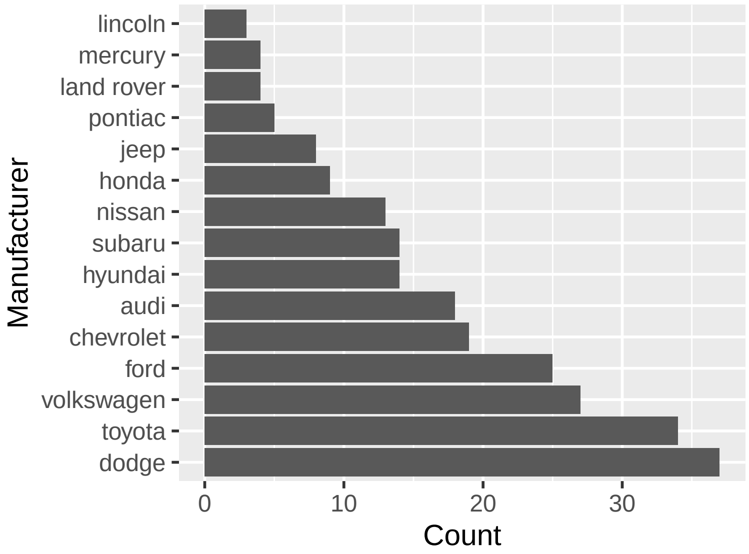
The histogram for highway miles per gallon does not seem normal, however the majority of points lie around 27. Additionally, there seems to be a few outliers past 40. These observations will all have an impact on our models because they may cause inaccuracy within the model as a result of a skewed variance.

Exercise 2



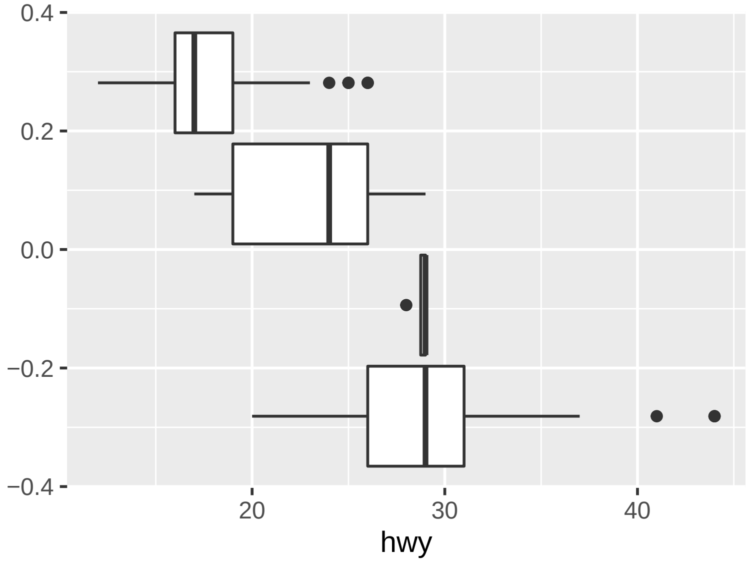
Yes, looking at the scatterplot, there is a positive linear relationship between hwy and cty. We can see that as hwy increases, cty increases as well. Furthermore, we can say that there exists a strong positive correlation between hwy and cty which will eventually help us in performing feature selection depending on our model.

Exercise 3



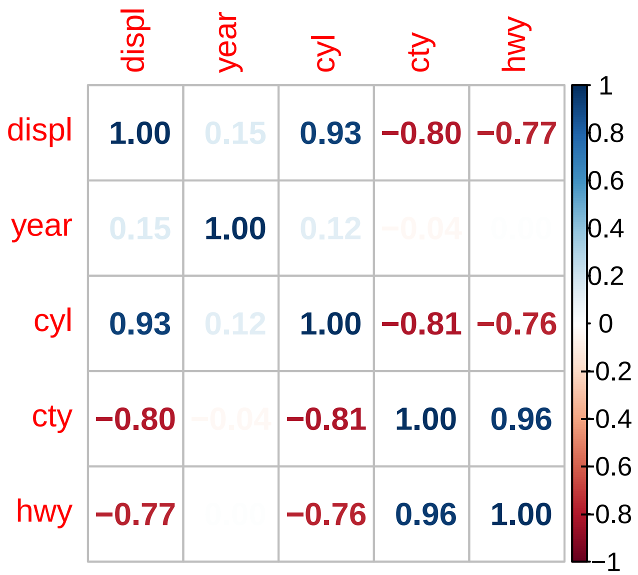
Looking at the bar plot, we can see that Dodge produced the most amount of cars and Lincoln produced the least amount of cars.

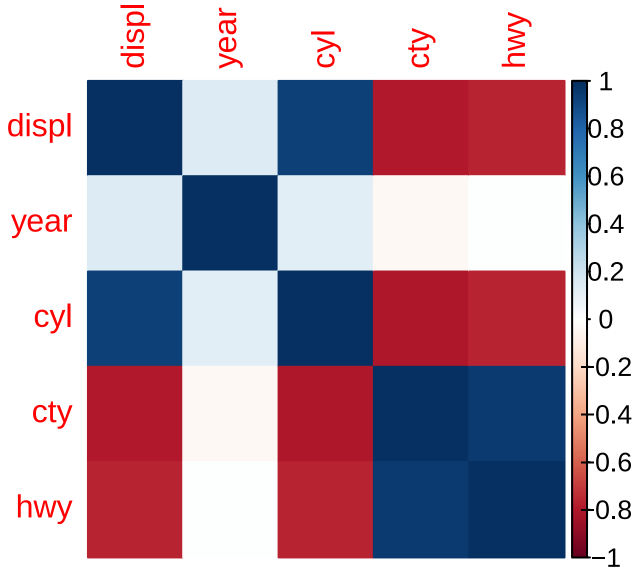
Exercise 4



Based on the box plot of hwy, grouped by cyl, there seems to be a common trend in outliers as there are points that deviate from the boxes.

Exercise 5





We can see that according to the correlation matrix, (year and displ), (cyl and displ), (cyl and year) and (hwy and city) are positively correlated with one another. Further, (cty and displ), (hwy and displ), (cty and year), (cty and cyl) and (hwy and cyl) are negatively correlated with one another. These relationships intuitively make sense such as the number of cylinders being highly correlated with engine displacement in vehicle engineering. I did not expect city and highway miles per gallon to be so correlated so I am interested in the exact underlying reason for such a high relationship.