PSTAT 131 Homework 1

Caleb Mazariegos

2022-03-31

Machine Learning Main Ideas

Question 1:

Define supervised and unsupervised learning. What are the difference(s) between them?

Supervised learning is when input data is given, we can see an answer key to help train our model. Unsupervised learning is when the input data is not given, therefore the model needs to learn and figure out patterns on it's own.

Question 2:

Explain the difference between a regression model and a classification model, specifically in the context of machine learning.

In a regression model, the response variable Y is quantitative, which means that the values will be numerical values such as price. In a classification model, the response variable is qualitative which means that the values will be split into categories such as survived/died.

Question 3:

Name two commonly used metrics for regression ML problems. Name two commonly used metrics for classification ML problems.

Two commonly used metrics for regression Machine Learning problems are Mean Squared Error and Mean Absolute Error. Two commonly used metrics for classification machine learning problems are accuracy and precision.

Question 4:

As discussed, statistical models can be used for different purposes. These purposes can generally be classified into the following three categories. Provide a brief description of each.

Descriptive models: These models visually emphasize a trend in data

Inferential models: States relationship between outcome and predictors. The aim of this model is to test theories.

Predictive models: The aim of this model is to predict the response variable Y with minimum reducible error. Not focused on hypothesis tests.

Question 5:

Predictive models are frequently used in machine learning, and they can usually be described as either mechanistic or empirically-driven. Answer the following questions.

Define mechanistic. Define empirically-driven. How do these model types differ? How are they similar?

A mechanistic model uses a theory to predict what will happen in the real world. An empirically-driven model does the opposite, it studies real world events to develop a theory. A mechanistic model is more flexible because you can add parameters, an empirically-driven model requires a large number of observations. They are similar because they can be over fitted, which means that the statistical model exactly fits against the training data.

In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice.

In my opinion, it is easier to understand an empirically-driven model. My reasoning is that we can use what has happened in previous experiments to have an accurate model.

Describe how the bias-variance trade-off is related to the use of mechanistic or empirically-driven models.

The bias-variance trade-off is related to the use of mechanistic or empirically-driven models because if there is low variance and high bias, the models will be over fitted.

Question 6:

A political candidate's campaign has collected some detailed voter history data from their constituents. The campaign is interested in two questions:

Given a voter's profile/data, how likely is it that they will vote in favor of the candidate?

How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate?

Classify each question as either predictive or inferential. Explain your reasoning for each.

The first question is predictive because it is predicting whether someone is likely to vote based on their data.

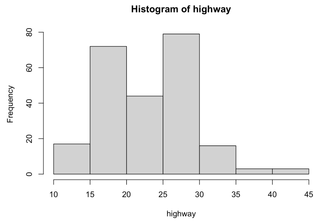
The second question is inferential because it is testing the theory on whether a voter's contact with a candidate will affect their support.

Exploratory Data Analysis

Exercise 1

We are interested in highway miles per gallon, or the hwy variable. Create a histogram of this variable. Describe what you see/learn.

```
highway <- mpg$hwy
hist(highway)
```



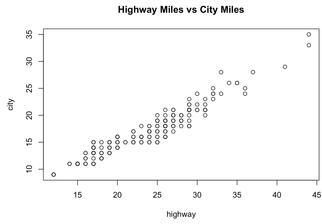
From what I can see in this histogram, it is skewed to the left. The most frequent value is between 25 and 30 highway miles.

Exercise 2:

Create a scatterplot. Put hwy on the x-axis and cty on the y-axis. Describe what you notice. Is there a relationship between hwy and cty? What does this mean?

```
city <- mpg$cty

plot(highway, city, main = "Highway Miles vs City Miles")
```



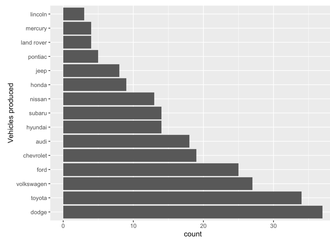
The scatterplot looks very neat and many of the points are arranged in what appears to be vertical lines. There is a positive relationship between highway and city miles which means that as highway miles increase city miles also increase.

Exercise 3:

Make a bar plot of manufacturer. Flip it so that the manufacturers are on the y-axis. Order the bars by height. Which manufacturer produced the most cars? Which produced the least?

```
data <- mpg
data$manufacturer <- as.factor(data$manufacturer)

ggplot(mpg,aes(x=reorder(manufacturer,manufacturer,function(x)-length(x)),), horizontal = TRUE,) + geom_bar() + c
oord_flip() + xlab("Vehicles produced")
```



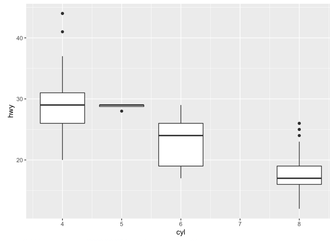
ggplot source: https://murraylax.org/rtutorials/barplots.html

Dodge produced the most vehicles and Lincoln produced the least

Exercise 4:

Make a box plot of hwy, grouped by cyl. Do you see a pattern? If so, what?

```
ggplot(mpg, aes(group = cyl, x = cyl, y = hwy)) + geom_boxplot()
```



source:

https://abhishekstats.com/2020/08/05/box-plot-with-ggplot2/

There is a negative relationship between highway miles per gallon and the number of cylinders. In other words, the more cylinders in a vehicle, the less highway miles per gallon.

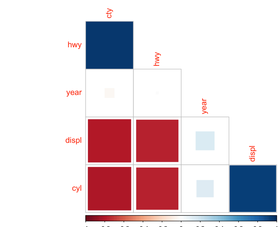
Exercise 5:

Use the corrplot package to make a lower triangle correlation matrix of the mpg dataset.

```
mpg_num <- mpg %>%
  select(displ, year, cyl, cty, hwy)

M = cor(mpg_num)

corrplot(M, method = 'square', order = 'FPC', type = 'lower', diag = FALSE)
```



Which variables are positively or negatively correlated with which others? Do these relationships make sense to you? Are there any that surprise you?

City miles per gallon is positively correlated with highway miles per gallon and negatively correlated with year, engine displacement, and cylinders. Highway miles per gallon is negatively correlated with engine displacement and cylinders. Year has a slight positive correlation with engine displacement and cylinders. Engine displacement has a positive correlation with cylinders.

I have limited knowledge of cars, but most of these make sense to me. The only relationship that surprised me was that the variable "Year" does not have a strong correlation with any variable.