Homework Assignment 1

Teo Zeng

April 03, 2022

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

   supervised learning: For each observation of the predictor measurement(s) $x_i, i = 1, ..., n$ there is an associated response measurement y_i. We wish to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference). (p26, ISLR)

   unsupervised learning: For every observation $i = 1, ..., n$, we observe a vector of measurements x_i but no associated response y_i. We wish to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference). (p26, ISLR)

   In **supervised learning**, there is the presence of outcome variable to guide the learning process. On the other hand, in **unsupervised learning** we observe only the features and have no measurements of the outcome. (p2, ESL)

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

   Since **Quantitative variables** take on numerical values and **qualitative variables** take on values in one of *K* different classes. We tend to refer to problems with a **quantitative** response as *regression problems*, which uses *regression models*. Those involving a **qualitative** response are often referred to as *classification problems*, which uses *classification models*. (p28, ISLR)

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

   (Skipped)

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:

     Descriptive models chooses models to best visually emphasis a trend in data (Class Powerpoint)

   + Predictive model:

     Predictive model aims to predict *Y* with minimum reducible error. (Class Powerpoint)

   + Inferential models:

     Inferential explores the significant feature: that is, it aims to test theories or casual claims or state the relationship between outcome and predictors.(Class Powerpoint)

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?
   + In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice.
   + Describe how the bias-variance tradeoff is related to the use of mechanistic or empirically-driven models.

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.

     In the first scenario, the question is *predictive* since each person's likelihood to vote can be predicted, which is a number between 0 and 1. On the other hand, in the second scenario, the question is *inferential* since it asks for a relationship between outcome and predictors.

Exercises

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

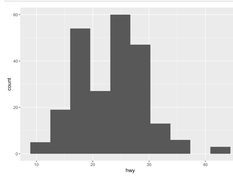
```
## ✔ ggplot2 3.3.5     ✔ purrr   0.3.4
## ✔ tibble  3.1.6     ✔ dplyr   1.0.8
## ✔ tidyr   1.2.0     ✔ stringr 1.4.0
## ✔ readr   2.1.2     ✔ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
```

1.

```
hist <- ggplot(mpg, aes(x=hwy)) + geom_histogram(bins = 10)
print(hist)
```

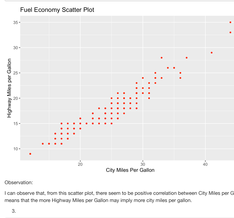


Observation:

In this histogram where "Highway Miles per Gallon" is plotted on the x-axis and its frequency is plotted on the y-axis. I see that it has highest frequency in 25 ~ 30 Gallon, and most of the highway miles per gallon is in 15 ~ 30 Gallon

2.

```
scatterplot <- ggplot(mpg, aes(hwy, cty)) +
  geom_point(color='red', size= 1) +
  labs(x='City Miles Per Gallon', y='Highway Miles per Gallon', title='Fuel Economy Scatter Plot')
print(scatterplot)
```

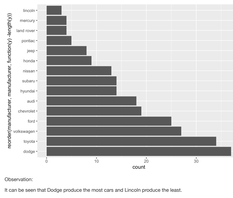


Observation:

I can observe that, from this scatter plot, there seem to be positive correlation between City Miles per Gallon and Highway Miles per Gallon. This means that the more Highway Miles per Gallon may imply more city miles per gallon.

3.

```
plot3<-ggplot(mpg, aes(y=reorder(manufacturer, manufacturer,function(y)-length(y)))) + geom_bar()
plot3
```

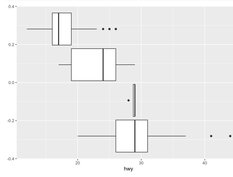


Observation:

It can be seen that Dodge produce the most cars and Lincoln produce the least.

4.

```
barc<- ggplot(mpg,aes(group = cyl, x=hwy))+geom_boxplot()
barc
```



Observation:

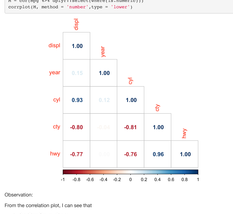
One possible pattern may be that the more cylinders a car have, the less highway miles per gallon this car may have.

5.

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
M = cor(mpg %>% dplyr::select(where(is.numeric)))
corrplot(M, method = 'number',type = 'lower')
```



Observation:

From the correlation plot, I can see that

- **Positive Correlations**:
  - displ-cyl
  - cty-hwy
- **Negative Correlations**
  - displ-cty
  - displ-hwy
  - cyl-cty
  - cyl-hwy

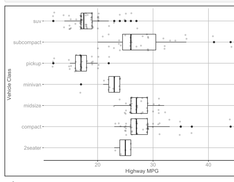
And these correlations make sense because more engine displacement should mean more cylinderder the car have. And cars that have more city miles should have more highway mileage also.

However, I am surprised to see that engine displacement is negatively correlated to city and highway mileage, so does cylinder.

PSTAT 231 Exercises

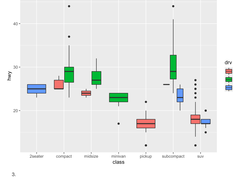
1.

```
library(ggplot2)
library(ggthemes)
p <- ggplot(mpg, aes(x=hwy, y=class)) +
  geom_boxplot() +
  geom_jitter(shape=16, position=position_jitter(0.2),alpha = 0.25) +
      labs(x = "Highway MPG",
           y = "Vehicle Class") +
      theme(panel.grid.major = element_line(colour = "grey"),
      panel.background = element_rect(fill = 'NA'),
      plot.background = element_rect(fill = 'NA',colour = 'black',size =1),
      axis.line.x = element_line(),
      axis.ticks = element_line(colour = "grey50",size = rel(1)),
      axis.title = element_text(colour = "grey50",size = rel(1)))
```



2.

```
p <- ggplot(mpg, aes(x=class,  y=hwy, fill=drv))+
  geom_boxplot()
p
```



3.

```
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point(aes(colour = drv)) +
  geom_smooth(aes(linetype = drv), se = FALSE)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```