131hw1

Camilla Zhai

2022-10-02

Machine Learning Main Ideas

Q1: Supervised learning is a machine learning method of training a computer algorithm on labeled input data for a specific output. The model is trained until it recognizes the underlying patterns and relationships between the input data and the output labels, allowing it to produce accurate labeling results when presented with previously unseen data.(response known) Unsupervised learning is a machine learning method in which models learn without being supervised by a training dataset. In this method, models use the given data to find hidden patterns and insights. In other words, the algorithm is given unlabeled data and is designed to find patterns or similarities on its own.(response not known) Differences: Unsupervised learning algorithms do not use labeled input and output data, whereas supervised learning algorithms do. The goal of supervised learning is to predict outcomes for new data by knowing what type of result to expect, whereas the goal of unsupervised learning is to gain insights from large amounts of new data by letting machine learning decide what is interesting from the dataset.

Q2: The output variable in a regression model can be either continuous or real, whereas the output variable in a classification model must be discrete. Furthermore, the task of the regression algorithm is to map the input value to the continuous output variable, whereas the task of the classification algorithm is to map the input value to the discrete output variable. Furthermore, the goal of a regression model is to find the best fit line that accurately predicts the output, whereas the goal of a classification model is to find the decision boundary that divides the dataset into different classes.

Q3: Regression ML problem: Mean Squared Error and Root Mean Squared Error Classification ML problem: F1 score and Confusion Matrix

Q4: From slide 39 of the lecture power point: Descriptive models seek to visually emphasize a trend in data by employing techniques such as plotting a line on a scatterplot. The goal of an inferential model is to test theories to see which features are significant; it describes the relationship between outcome and predictors. The goal of a predictive model is to predict Y with the least amount of reducible error in order to find the best combination of features. It does not focus on hypothesis tests.

Q5: Mechanistic model uses a theory to predict what will happen in the real world. Empirically-driven model studies real-world events to develop a theory. Mechanistic modeling describe the inner mechanism and phenomena of a system using first principles. Empirically-driven modeling is the regression of model parameters to fit the input-output behavior from measured data. Compare to mechanistic model, Empirically-driven model is easier to understand. This is because this modeling method is based on empirical observations that are straight forward to understand and use than the theory-based method.

Empirically-driven models have higher variance and lower bias; Mechanistic models have higher bias and lower variance.

Q6: The first question is predictive: it employs historical data (voter history data) to forecast and predict likely future outcomes (voters vote in favor of candidate); it does not rely on hypothesis tests. The second question is inferential: it infers properties from tests and estimates. Its goal is to test theories/hypotheses and determine the relationship between outcome and predictor. In this case, the hypothesis to be tested is the question itself. “Voters who had personal contact with the candidate” is the predictor, and “voters’ likelihood of support for the candidate” is the outcome/response.

Exploratory Data Analysis

```r
#install.packages("tidyverse")
#install.packages("ISLR")
#install.packages("tidymodels")
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

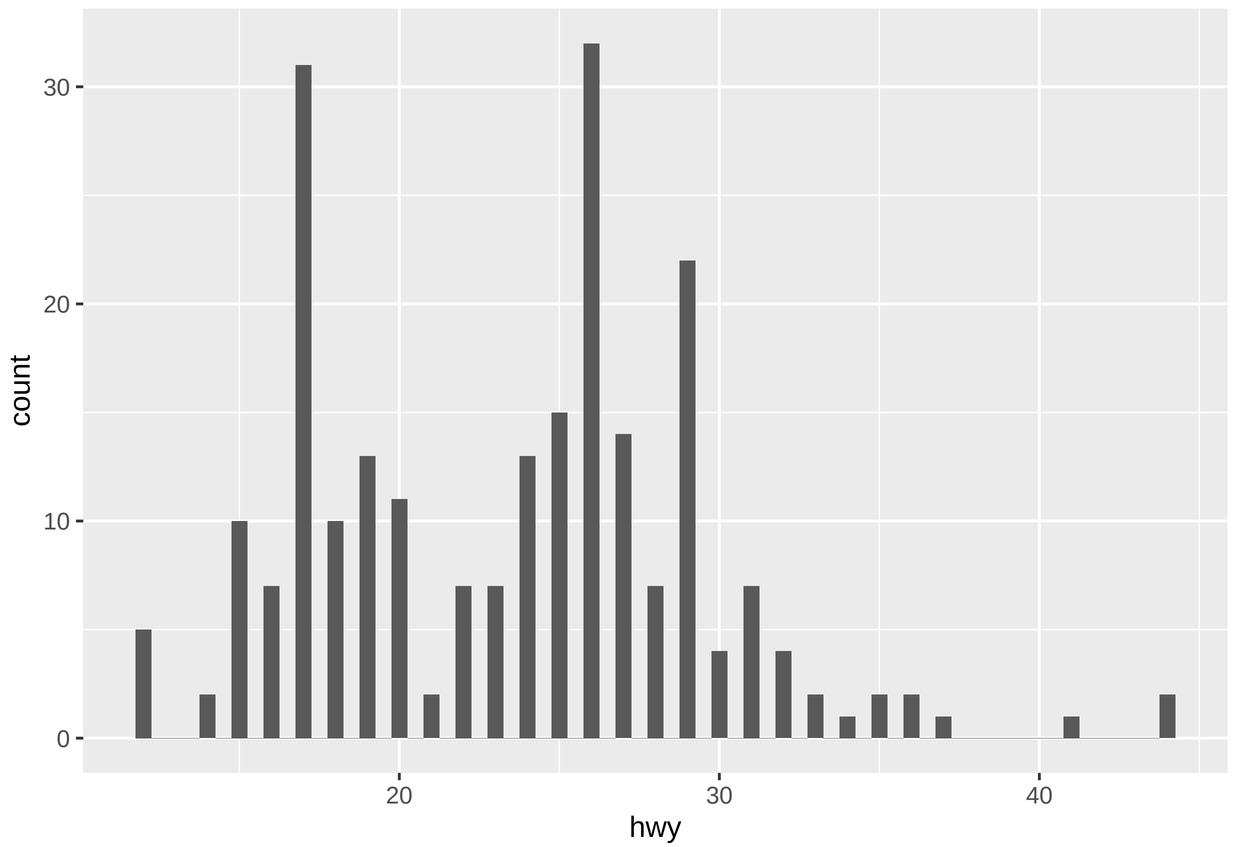
```r
library(tidymodels)
```

```
## -- Attaching packages -------------------------------------- tidymodels 1.0.0 --
## v broom        1.0.1     v rsample      1.1.0
## v dials        1.0.0     v tune         1.0.0
## v infer        1.0.3     v workflows    1.1.0
## v modeldata    1.0.1     v workflowsets 1.0.0
## v parsnip      1.0.2     v yardstick    1.1.0
## v recipes      1.0.1
## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Use tidymodels_prefer() to resolve common conflicts.
```

```r
library(ISLR)
library(ggplot2)
```

Exercise 1:

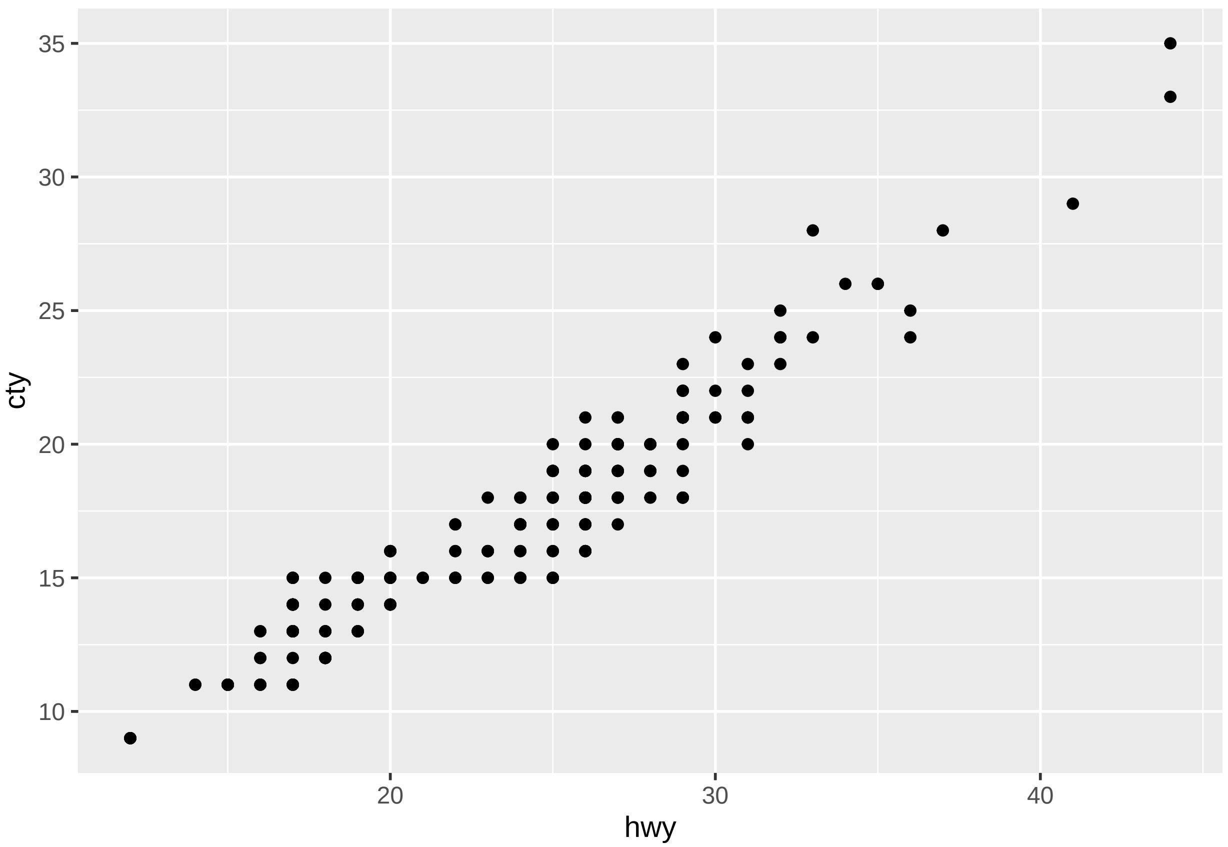
```r
hwy_hist <- ggplot(mpg, aes(x=hwy))+geom_histogram(binwidth = 0.5)
hwy_hist
```



This histogram has a bi-modal distribution(it has double peaks).

Exercise2:

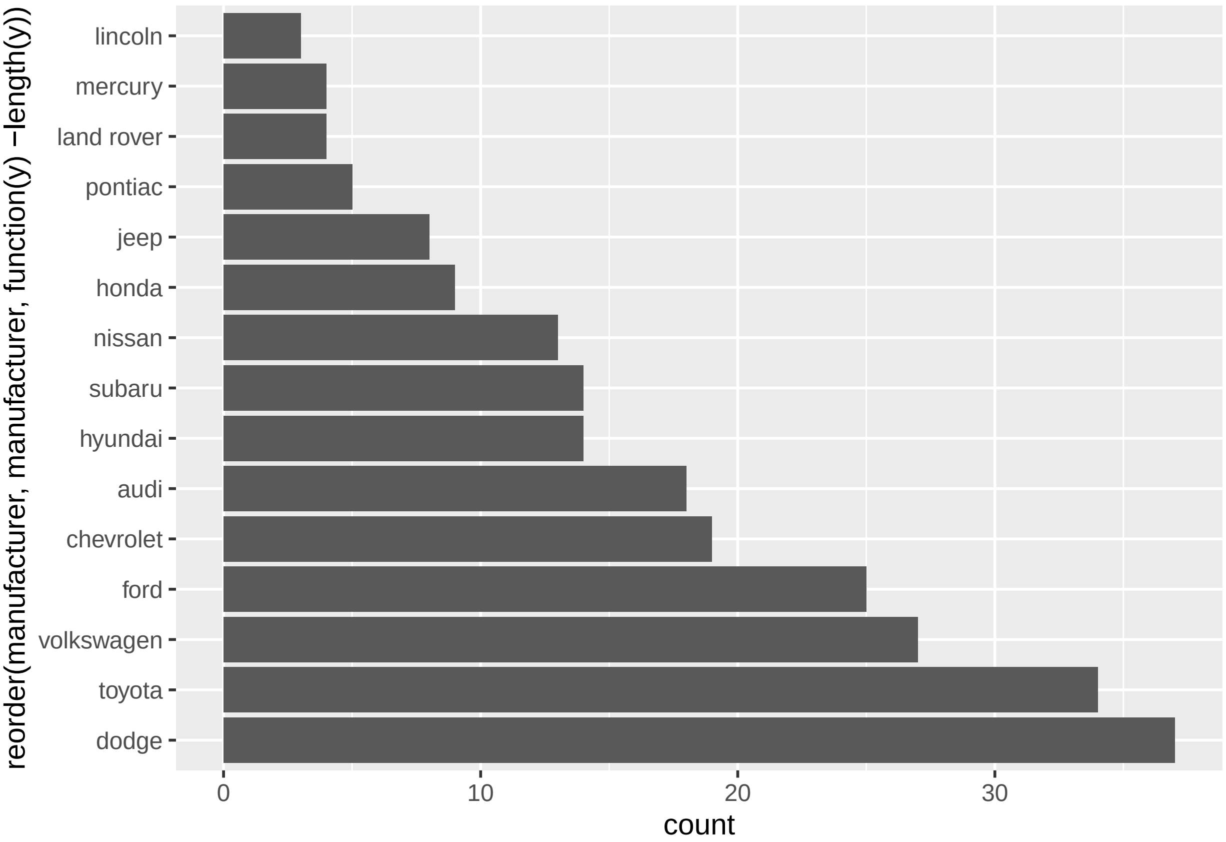
```
ggplot(mpg, aes(x=hwy, y=cty))+geom_point()
```



We can see from the scatterplot that as hwy increases, so does cty, and there do not appear to be any outliers in the data. This scatterplot demonstrates a strong, positive, linear relationship between hwy and cty. This indicates that when the value of hwy increases by one unit, the value of cty also increases, but not by one unit. The same holds true for the scenario of a decrease.

Exercise 3:

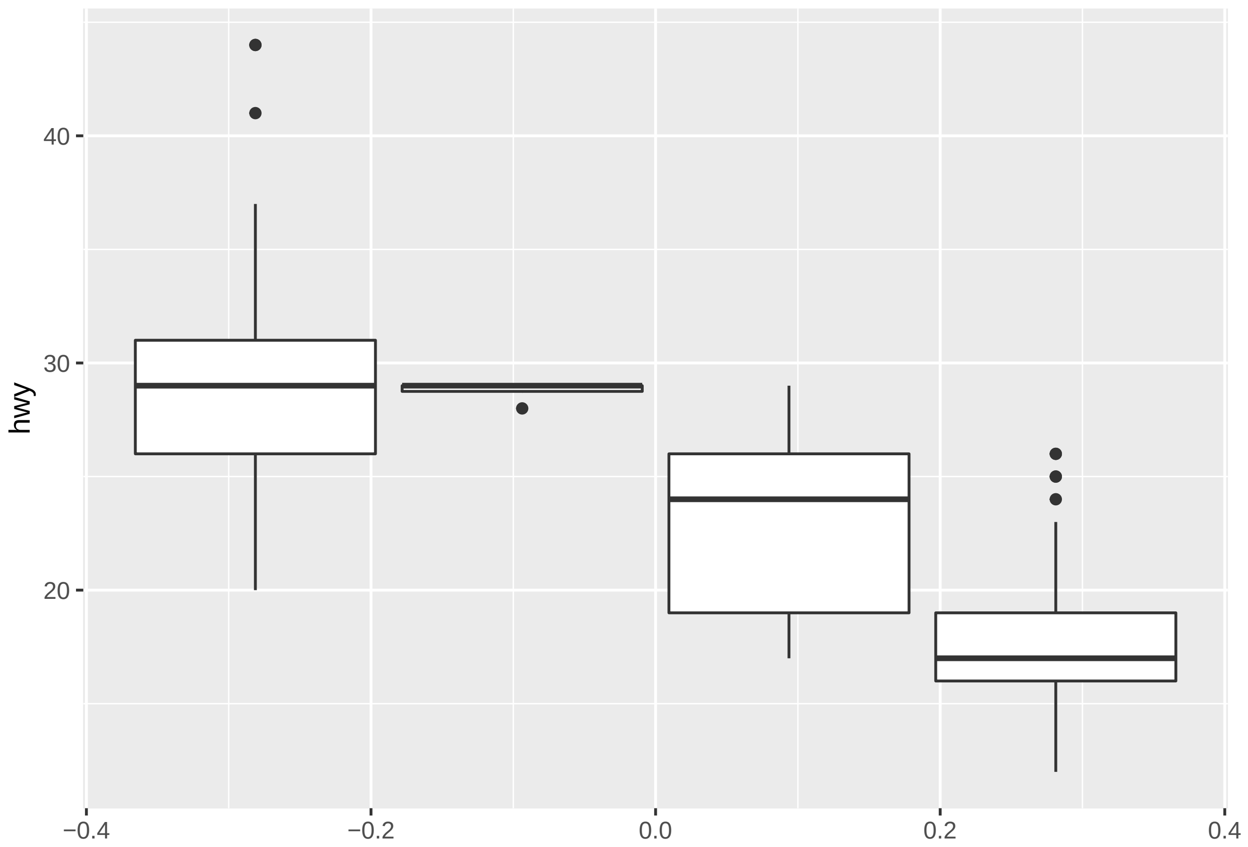
```
ggplot(mpg, aes(y=reorder(manufacturer,manufacturer,function(y)-length(y))))+geom_bar()
```



From the bar plot above, we can conclude that dodge produced the most cars and lincoln produced the least cars.

Exercise 4:

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



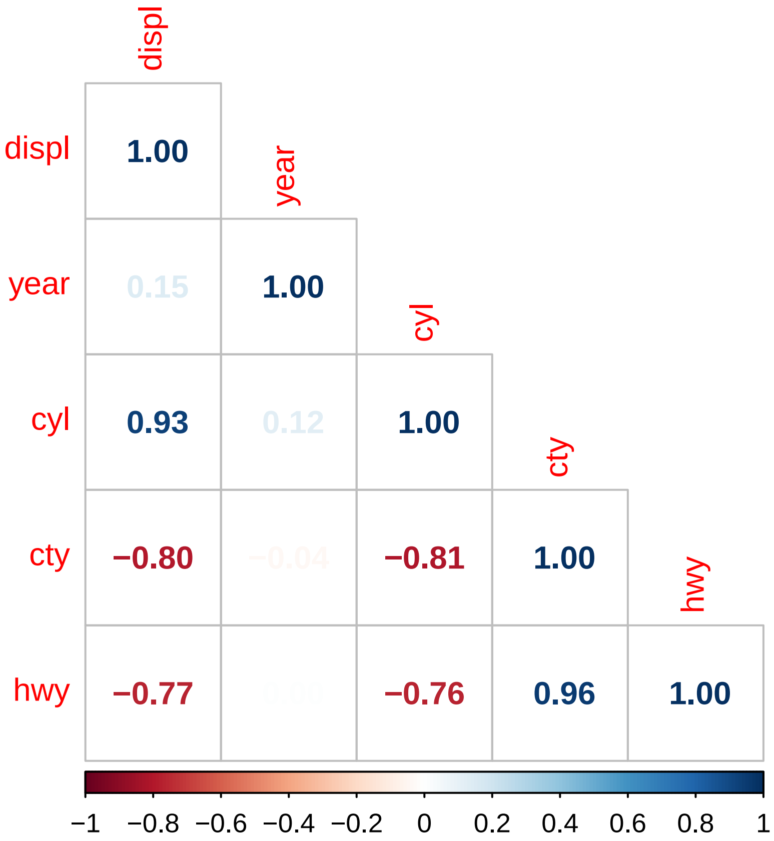
The boxplot demonstrates that cyl and hwy have a negative correlation. As the cyl increases/decreases, hwy decreases/increases. This demonstrates that cars with fewer cylinders tend to have greater highway fuel efficiency than cars with more cylinders.

Exercise 5:

```
#install.packages("corrplot")
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
mpg2 <- cor(select_if(mpg,is.numeric))
corrplot(mpg2,method='number',type='lower')
```



year is positively correlated with displ. cyl is positively correlated with displ and year. cty is negatively correlated with displ, year and cyl. hwy is positively correlated with cty and negatively correlated with displ and cyl.
These relationships make sense.
I was surprised that year is correlated with most other variables, even if the correlation is small.