131 Homework 1

Tonia Wu

3/29/2022

Q1. *Define supervised and unsupervised learning. What are the difference(s) between them?*

- Supervised learning involves modelling to predict an output based on input(s). This requires training data from which the model will learn.
- Unsupervised learning involves inputs and no outputs so that we can learn about the data and potentially discover patterns

Q2. *Explain the difference between a regression model and a classification model, specifically in the context of machine learning.*

- Regression models involve continuous, quantitative output
- Classification models involve qualitative output (such as a yes/no resu)

Q3. *Name two commonly used metrics for regression ML problems. Name two commonly used metrics for classification ML problems.* - Regression: - Classification: whether something will increase or decrease

Q4. *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*:
- *Inferential models*:
- *Predictive models*:

Q5. *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 takes a relationship or theory and imposes it on the data (from lecture)
- AAn empirically-driven model looks at the data and sees what best fits it (from lecture)
- ASimilarities: !!!!!!!!!!!!!!!
- ADifferences: !!!!!!!!!!!!!!

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.

Q6. *A political candidate's campaign has collected some detailed voter history data from their constituents. The campaign is interested in two questions:*

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

b. *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.

- a: as we are using past data to predict the future, this involves a predictive model.
- b: since we're interested in the relationship between the inputs and the output, this involves an inferential model.

Exploratory Data Analysis

E1. *We are interested in highway miles per gallon, or the hwy variable. Create a histogram of this variable. Describe what you see/learn.*

E2. *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?*

E3. *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?*

E4. *Make a box plot of hwy, grouped by cyl. Do you see a pattern? If so, what?*

E5. *Use the corrplot package to make a lower triangle correlation matrix of the mpg dataset. (Hint: You can find information on the package here.)*

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