PSTAT 131 HW 1

Finn Stack

4/3/2022

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Question 1:

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

Supervised learning uses existing data to train algorithms into classifying or predicting outcomes. Unsupervised used unlabeled data sets in order to analyze and predict outcomes. The main distinction is the use of labeled datasets in supervised and the use of unlabeled datasets in unsupervised datasets.

Question 2:

Explain the difference between a regression model and a classification model, specifically in the context of machine learning. Classification is about predicting a label, whereas regression is about predicting a quantity.

Question 3:

Name two commonly used metrics for regression ML problems. Name two commonly used metrics for classification ML problems. For regression: numerical values such as height and price. Classification model: categorical values such as survived/died and above/below, yes/no.

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: Model is used in order to summarize data.

Inferential models: Model used to interpret the meaning of the statistics.

Predictive models: Used in order to forecast future data or predict data.

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. Empirically driven uses real-world data and analysis in order to predict a theory,

In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice. Empirically driven model is easier to understand because it is based off of real-world data and used to create a theory to predict future values or make theories based off the data.

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

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? predictive

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

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

Exploratory Data Analysis

This section will ask you to complete several exercises. For this homework assignment, we'll be working with the mpg data set that is loaded when you load the tidyverse. Make sure you load the tidyverse and any other packages you need.

Exploratory data analysis (or EDA) is not based on a specific set of rules or formulas. It is more of a state of curiosity about data. It's an iterative process of:

generating questions about data visualize and transform your data as necessary to get answers use what you learned to generate more questions A couple questions are always useful when you start out. These are "what variation occurs within the variables," and "what covariation occurs between the variables."

You should use the tidyverse and ggplot2 for these exercises