Week 10 Overview

This week is about logistic regression. This form of regression is appropriate when the dependent variable is a binary. Like all regression, this binary-base is predictive as well. Logistic regression describes the relationship between one binary variable (we call it the dependent variable) and one or more variables of any kind (these are independent variables).

This form of regression is important for a lot of things. In fact, you may not know it but it is a form of machine learning. More on this next week.

Contents of the Week

Objectives

After completing this week, you should be able to:

- Fit a binary logistic regression model to a dataset
- Create a logistic regression classifier
- Determine accuracy of logistic regression classifiers
- Differences between logistic regression and linear
- Conduct exploratory data analysis

Weekly Resources

- Comprehensive R Archive Network
- R Studio
- Sage Publications. (2021). Discovering Statistics Using R
- RStudio, PBC. (2021). RStudio Documentation
- RStudio, PBC. (2021). R Studio Cheatsheets

Week 10 Readings, Assignments, and Tasks

Here are your tasks for this week:

Read the following:

- *R for Everyone:* Chapter 20
- *Discovering Statistics Using R*: Chapter 8
- Understanding the Bias-Variance Tradeoff
- Calculating UAC: The Area Under a ROC Curve

Complete the following:

- 10.1 Discussion/Participation
- 10.2 Exercise
- 10.3 Final Project Step 2

Helpful Sources

Simplilearn. (2018). Logistic Regression in R.

10.1 Discussion/Participation

Here are optional topics for discussion via Teams this week. Remember, these topics aren't required, but if you are struggling to know what to post about, these can be used to initiate discussion!

1. What is logistic regression? When is it needed?
2. What is survival analysis? How is it done? Does it only apply to human life?
3. What are the principles behind logistic regression?
4. What is log-likelihood? What is deviance?

5. How do r and r^2 play into logistic regression? Is it similar to linear?
6. What is the z statistic?
7. What is the odds ratio?
8. What methods exist for logistic regression in R?
9. What can go wrong? How do you select a method of regression?
10. What is binary logistic regression? When is it used?
11. How should logistic regression be reported?
12. What should you test for with logistic regression?
13. What is multinomial logistic regression?
14. What is exploratory data analysis (EDA)? Where does this step fall in the data science process?

10.2 Exercise

Complete the following exercises using RMarkdown:

1. Fit a Logistic Regression Model to Thoracic Surgery Binary Dataset
    a. For this problem, you will be working with the thoracic surgery data set from the University of California Irvine machine learning repository. This dataset contains information on life expectancy in lung cancer patients after surgery. The underlying thoracic surgery data is in ARFF format. This is a text-based format with information on each of the attributes. You can load this data using a package such as foreign or by cutting and pasting the data section into a CSV file.
    b. Assignment Instructions:
        i. Fit a binary logistic regression model to the data set that predicts whether or not the patient survived for one year (the Risk1Y variable) after the surgery. Use the glm() function to perform the logistic regression. See Generalized Linear Models for an example. Include a summary using the summary() function in your results.
        ii. According to the summary, which variables had the greatest effect on the survival rate?
        iii. To compute the accuracy of your model, use the dataset to predict the outcome variable. The percent of correct predictions is the accuracy of your model. What is the accuracy of your model?

2. Fit a Logistic Regression Model
    a. Fit a logistic regression model to the binary-classifier-data.csv dataset
    b. The dataset (found in binary-classifier-data.csv) contains three variables; label, x, and y. The label variable is either 0 or 1 and is the output we want to predict using the x and y variables.
        i. What is the accuracy of the logistic regression classifier?
        ii. Keep this assignment handy, as you will be comparing your results from this week to next week.

Submission Instructions

For all assignments in this course, you must export the script or Markdown file to PDF. You are welcome to submit your URL to GitHub in addition, but all submissions must include a PDF (no zip files will be accepted either).

The assignment is due by Sunday, 11:59 p.m. CT.

10.3 Final Project Step 2

At this point you should have framed your problem/topic, described the data, and how you plan to solve the problem. Now you need to move on to the next step of analyzing and preparing the data. Adding on to the draft you started in Step 1:

- Data importing and cleaning steps are explained in the text and follow

a logical process. Outline your data preparation and cleansing steps.
- With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.
- What do you not know how to do right now that you need to learn to import and cleanup your dataset?
- Discuss how you plan to uncover new information in the data that is not self-evident.
- What are different ways you could look at this data to answer the questions you want to answer?
- Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.
- How could you summarize your data to answer key questions?
- What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).
- What do you not know how to do right now that you need to learn to answer your questions?
- Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Some additional questions you may want to consider asking yourself as you work through this section of the project:

1. What features could you filter on?
2. How could arranging your data in different ways help?
3. Can you reduce your data by selecting only certain variables?
4. Could creating new variables add new insights?
5. Could summary statistics at different categorical levels tell you more?
6. How can you incorporate the pipe (%>%) operator to make your code more efficient?

You can use the following template for Step 2:

- How to import and clean my data
- What does the final data set look like?
- Questions for future steps.
- What information is not self-evident?
- What are different ways you could look at this data?
- How do you plan to slice and dice the data?
- How could you summarize your data to answer key questions?
- What types of plots and tables will help you to illustrate the findings to your questions?
- Do you plan on incorporating any machine learning techniques to

answer your research questions? Explain.
- Questions for future steps.

Submission Instructions

Submit your draft (with Step 1 & Step 2 now combined) and completed code to date via PDF (of R Markdown file) to the assignment link.

The assignment is due by Sunday, 11:59 p.m. CT.