Week 4: Regression Models

Introduction

Contents of the Week

- Introduction
- Readings
- Supplemental Materials
- 4.1 Discussion/Participation
- 4.2 Exercise: Predicting Fuel Efficiency

Topics

- A Data preparation/cleaning
- B Training regression models
- C Evaluating regression models
- D Interpreting regression models

Readings

Read the following:

- Chapters 11-13 of *Machine Learning with Python Cookbook*

Supplemental Materials

All of the materials below are from external sources. Authorship and ownership are indicated within the sources themselves.

Readings

- Regression Metrics for Machine Learning
- How To Interpret R-squared and Goodness-of-Fit in Regression Analysis
- R-squared Cautions
- Interpreting Regression Coefficients
- Regression with Categorical Variables
- A comprehensive beginners guide for Linear, Ridge and Lasso Regression in Python and R
- 7 types of regression techniques you should know in Machine Learning
- How to Interpret P-values and Coefficients in Regression Analysis

Videos

4.1 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 a regression model?
2. What are some common types of regression models?
3. What is the difference between a regression model and a classification model?
4. Do you consider logistic regression as a regression model?
5. What are some common metrics for evaluating regression models?
6. How do we handle a categorical feature when we are building a regression model?

7 What are lasso and ridge regression?

4.2 Exercise: Predicting Fuel Efficiency

In the Week 4 Exercise, you will build a linear regression model to predict fuel efficiency (miles per gallon) of automobiles. Download the auto-mpg.csv dataset from: Auto-mpg dataset.

1. Load the data as a Pandas data frame and ensure that it imported correctly.
2. Begin by prepping the data for modeling:
    - Remove the car name column.
    - The horsepower column values likely imported as a string data type. Figure out why and replace any strings with the column mean.
    - Create dummy variables for the origin column.
3. Create a correlation coefficient matrix and/or visualization. Are there features highly correlated with mpg?
4. Plot mpg versus weight. Analyze this graph and explain how it relates to the corresponding correlation coefficient.
5. Randomly split the data into 80% training data and 20% test data, where your target is mpg.
6. Train an ordinary linear regression on the training data.
7. Calculate R2, RMSE, and MAE on both the training and test sets and interpret your results.
8. Pick another regression model and repeat the previous two steps. Note: Do NOT choose logistic regression as it is more like a classification model.

Submission Instructions

Click the title above to submit your assignment.

This exercise is due by Sunday 11:59 PM.

Submit your code, output, and answers at the link above. Comment all your code and answer any questions that are asked in the instructions. It is perfectly fine to answer a question by displaying output from your code, but make sure you are displaying the appropriate output to answer the question. I would recommend using and submitting a Jupyter Notebook, but this is not required.

View the rubric for this Assignment by clicking on the link below:

Exercise Rubric