Introduction to Data Science

Daniel Gutierrez, Data Scientist

Los Angeles, Calif.

Course Outcomes

- Explore supervised machine learning for prediction
- How to use the linear model in R
- How to use R for classification problems using logistic regression and Random Forest

Lesson Objectives

- Be able to employ supervised machine learning using – linear regression and classification
- Using the `lm()` algorithm in R
- Using the `glm()` algorithm in R
- Using the `randomForest()` algorithm in R
- Training the model to obtain coefficients
- Making predictions using the trained model and test set
- Measuring performance of the model

Supervised Machine Learning

- Overview of supervised machine learning
- Linear regression is the work horse of supervised learning
- Classification is the most popular supervised learning technique
- Create a training set (60%) to train the model, and also a test set (40%) to test accuracy of model
- Feature engineering drives process. Can use forward and/or backward stepwise selection

Supervised Machine Learning

- Simple linear regression with a continuous response variable and one continuous predictor
- Use EDA to understand the data: distribution of each feature variable, check correlation using scatterplot
- Use `lm()` to fit a linear model
- Use trained model coefficients to make predictions
- Plot regression line
- Residuals plot – distance between actual points and regression line

Supervised Machine Learning

- Multiple linear regression with a response variable and more than one predictor
- Use EDA to understand the data and detect a trend
- Use `lm()` to fit a linear model
- Use trained model coefficients to make predictions on test set

Supervised Machine Learning

- Classification using logistic regression with a 2 class (binary) categorical response variable and one or more predictor
- Use `glm()` for logistic regression
- Use the trained model to make predictions on test set (probabilities of class membership)

Supervised Machine Learning

- Classification using the Random Forest algorithm with a multi-class categorical response variable and one or more predictor
- Use `randomForest()` for Random Forest
- Use the trained model to make predictions on test set

Supervised Machine Learning

- Evaluating model performance
- Overfitting
- Measuring regression performance
- Measuring classification performance
- Diagnostic plots

Code modules

- WEEK 9-1 Code module – EDA for simple regression
- WEEK 9-2 Code module – Fit a linear model to make a prediction
- WEEK 9-3 Code module – Residuals plot
- WEEK 9-4 Code module – Create training and test set for multiple linear regression
- WEEK 9-5 Code module – Fit a linear model
- WEEK 9-6 Code module – Use trained model with the test set

Summary

- In WEEK 9 of Introduction to Data Science, we continued the data science process by exploring a popular supervised machine learning algorithm – linear regression.
- We used the `lm()` algorithm for both simple and multiple linear regression.
- We also saw useful plots to better understand the regression model.
- We split the data set into a training set and test set to come up with a test set error metric to see how well our model generalizes.