Boston University Metropolitan College

Foundations of Machine Learning

MET - CS555

Faculty: Farshid Alizadeh-Shabdiz, PhD, MBA
alizadeh@bu.edu
Office hours: by appointment

Course Description

This course covers foundations of machine learning, regression, and classification. Topics include how to describe data, statistical inference, 1 and 2 sample tests of means and proportions, simple linear regression, multiple linear regression, multinomial regression, logistic regression, analysis of variance, and regression diagnostics. These topics are explored using the statistical package R, with a focus on understanding how to use these methods and interpret their outputs and how to visualize the results. In each topic area, the methodology, including underlying assumptions and the mechanics of how it all works along with appropriate interpretation of the results are discussed. Concepts are presented in context of real world examples in order to help students to learn when and how to deploy different methods.

Learning Objectives

By successfully completing this course you will be able to:

- You will learn basic blocks of regression and classification
- Understand the basics of machine learning
- Summarize and present data in meaningful ways
- Select the appropriate analysis depending on research questions at hand
- Form testable hypotheses that can be evaluated using statistical analyses
- Understand and verify the underlying assumptions of a particular analysis
- Effectively and clearly communicate results from analyses performed with others
- Conduct, present, and interpret data analyses using R

Course Material

Course material will be available on Blackboard.

Class lectures, notes, slides, short write-ups, and video tutorials provide you with all needed information (theory, concepts, practical examples, R code basics, R code examples and all the details) that you will need to complete the quizzes, homework assignments, and final exam.

Blackboard hosts class information, including slides, short write-ups, video tutorials, quizzes, assignments, etc.

Text Book

The following book is the text book for the course. It is highly recommended to use the text book along side of the course material.

- **An Introduction to Statistical Learning with Applications in R, second edition (2021)** ; by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Published by Springer.
  The book has been made available online for free at https://statlearning.com. You can also purchase a hard copy.

There will be no specific reading assignments from the Text book.

Reference books

The following books are not required for the course. They can be used as a reference and help with the assignments and term project.

The following books are excellent supplemental texts for R that you may want to review as we go through the course.

- Teetor, P. (2019). **R cookbook.** Sebastopol, CA: O'Reilly. ISBN -13: 978-1492040682
  The book has been made available online at https://rc2e.com and
  the code at https://github.com/CerebralMastication/R-Cookbook

- Chang, W. (2021). R graphics cookbook. Sebastopol, CA: O'Reilly. ISBN 9781491978573
  The book has been made available online at https://r-graphics.org
  And the code at https://github.com/wch/rgcookbook

Additional Reference Books

- **Andy Field, Jeremy Miles and Zoe Field. (2012) Discovering Statistics Using R.**
  Publisher: SAGE Publications Ltd. ISBN-13: 978-1446200469

- https://www.openintro.org/stat/ Free PDF for download & R tutorials and codes.

- "Using R for Introductory Statistics, 2nd edition", by John Verzani, CRC Press, 2014. ISBN13: 978- 1466590731. (Reference book)

- "R for Everyone: Advanced Analytics and Graphics, 2nd Edition", by Jared P. Lander, Addison-Wesley Professional, 2017. ISBN13: 978-0134546926. (Reference book)

Class Policies

1) **Assignment Completion & Late Work** – all the assignment has to be submitted in person or electronically on Blackboard. No late work will be acceptable.

2) **Laptop Requirement –** Students should have a personal laptop. We will use laptops in classroom to write R programs. You will need a Laptop in the final exam as well.

3) **Academic Conduct Code** – Cheating and plagiarism will not be tolerated in any Metropolitan College course. They will result in no credit for the assignment or examination and may lead to disciplinary actions. Please take the time to review the Student Academic Conduct Code: http://www.bu.edu/met/metropolitan_college_people/student/resources/conduct/code.html.

Grading Criteria

The course grade will be based on

- In class check ins (15%)
- Quizzes (take home and in-class) (30%)
- Assignments submission (10%)
- Final project (15%)
- Final exam (30%)

Assignments are expected to be submitted by their respective due dates. Late submissions are not accepted.

Homework Assignments

There will be homework assignments focused on applying theory learned in the class to analyze a data set in R. Assignment submissions should be in a single **Microsoft Word**

or PDF file. The R code used to generate your results should be appended to the end of your assignment.

In class check ins

There will be in-class check ins to assess students understanding of concepts presented in the class. Students should ensure adequate preparation before the check ins. The check ins attendance is sufficient to get the full grade.

Quizzes

There will be take home and in-class quizzes to assess students understanding of concepts presented in the class. Students should ensure adequate preparation before the quiz. Please note that it won't be possible to do well on the quizzes without reviewing the course materials.

Final Examination

The final exam will be comprehensive and will cover material from the entire course.

The final exam will be closed notes and closed book.

Study Guide

Introduction to the science of statistics

- Fundamental Elements of Statistics
- Qualitative and Quantitative Data Summaries
- Normal distribution
- Sampling
- The Central Limit Theorem

Confidence intervals and hypothesis tests

- Statistical Inference
- Stating Hypotheses

- Test Statistics and p-Values
- Evaluating Hypotheses
- Significance Test “Recipe”
- Significance Tests and Confidence Intervals
- Inference about a Population Mean
- Two-Sample Problems

Understanding the association between two continuous or quantitative factors

- Scatterplots
- Correlation

Linear Regression

- Simple Linear Regression
- F-test for Simple Linear Regression
- T-test for Simple Linear Regression

Regression diagnostics

- Residual Plots
- Outliers and Influence Points
- Assumptions of least-square regression

Multiple linear regression

- Equation of multiple linear regression
- Interpretation of multiple linear regression
- F-test for Multiple Linear Regression
- T-test in Multiple Linear Regression
- Cautions about Regression
- Piecewise multiple linear regression
- Spline

Sampling methods

- Sampling methods
- Random sampling
- Hold out
- Training / Test / Validation
- Cross Validation

- Bootstrapping

Analysis of Variance (ANOVA)

- One-Way Analysis of Variance
- F-test for ANOVA
- Evaluating Group Differences
- Type I and Type II Errors
- Issues with Multiple Comparisons
- Assumptions of Analysis of Variance
- Relationship between One-Way Analysis of Variance and Regression
- One-Way Analysis of Covariance
- Two-Way Analysis of Variance
- Two-Way Analysis of Covariance

Analysis for proportions

- One-Sample Tests for Proportions
- Significance Tests for a Proportion
- Confidence Intervals for a Proportion
- Two-Sample Tests for Proportions
- Confidence Intervals for Differences in Proportions
- Significance Tests for Differences in Proportions
- Effect Measures

Classification algorithms assessment

Logistic Regression

- Logistic Regression
- Multiple Logistic Regression

Clustering algorithms assessment

Parameter selection methods

- Forward and backward parameter selection
- Regularization

Generalized linear model (GLM)

- Assumptions
- Maximum likelihood estimation
- MLE and MSE relationship