DATA SCIENCE II
P8106
2022 Spring

INSTRUCTOR

Yifei Sun
Email: ys3072@cumc.columbia.edu
Office hours: Monday 4-5 pm (virtual)

TEACHING ASSISTANT(S)

Angel Garcia de la Garza
Muhire Kwizera
Madison Stoms
Siquan Wang
Yinjun Zhao

TA OFFICE HOURS

Wednesday and Friday 4-5 pm (virtual)

COURSE DESCRIPTION

With the explosion of "Big Data" problems, statistical learning has become a hot field in many scientific areas. The goal of this course is to provide training in practical statistical learning. It is targeted to Biostatistics MS students with data analysis experience in R.

PREREQUISITES

Working knowledge in Calculus & Linear Algebra
P8105 Data Science I
P8130 Biostatistical Methods I

COURSE LEARNING OBJECTIVES

Students who successfully complete this course will be able to:

- Explain concepts and methods in statistical learning
- Apply classification and regression techniques beyond linear methods
- Conduct exploratory data analysis using methods in unsupervised learning
- Implement various statistical learning methods using R
- Build a pipeline for predictive modeling: data preprocessing, model training, model interpretation

RECOMMENDED REFERENCES

[ISL] *An Introduction to Statistical Learning with Applications in R* (main textbook)

[ESL] *The Elements of Statistical Learning*

[APM] *Applied Predictive Modeling* (available at library.columbia.edu)

ASSESSMENT AND GRADING POLICY

Student grades will be based on:

Homework	40%
Midterm project	40%
Final project	20%

Questions regarding the grading of homework assignments must be raised within a week of the assignment being returned. Collaboration on homework assignments is acceptable, but all submissions must be completed independently and clearly indicate the submitter's understanding of the material. Unclear or disorganized homework may have points removed, even if the content is correct.

The midterm project will consist of analyzing a complex dataset of your choosing using techniques learned in the first half of the semester.

The final project will consist of analyzing a complex dataset and completing a polished report and presentation. This will be a group project. The dataset used in the midterm project can be used in the final project.

SOFTWARE USE

We will use R and R Markdown; R Studio is recommended.
A laptop with R installed is required and should be brought to every class session.

COURSE STRUCTURE

Class sessions will be lectures, delivered using a mix of static content and live demonstrations.

COURSE SCHEDULE

Session 1 – Course Introduction

Learning Objectives:

Discuss the role of statistical learning in data science

Define the terminology in statistical learning

- Training/test/validation
- Supervised/unsupervised learning

Reading:

[ISL] 2.1 What is statistical learning? (Page 24-29)

2.2 Assessing model accuracy (Page 29-36)

Session 2 – An overview of the modeling process

Learning Objectives:

Understand the bias-variance tradeoff

Describe a general predictive modeling workflow

- Data splitting
- Model training
- Model evaluation

Conduct a case study that puts the processes together (using KNN)

Reading:

[APM] Chapter 2: A Short Tour of the Predictive Modeling Process

Session 3 – Resampling methods

Learning Objectives:

Explain how to use resampling methods for model selection/assessment

Describe different resampling procedures

- Validation set
- K-fold CV
- Repeated K-fold CV
- Monte Carlo CV
- 632 Bootstrap

Implement different resampling procedures in R

Reading:

[ISL] 5.1 Cross-validation (Page 176-183)

[APM] 4.4 Resampling techniques

Session 4 – Linear Regression

Learning Objectives:

Review the concepts in linear models, with an emphasize on predictive modeling

Derive the geometric interpretation of least squares (A review of inner product)

Understand potential limitations of the least squares method

Conduct a case study of linear regression when the predictors are high-dimensional and highly correlated

Reading:

[ISL] 3 Linear Regression (Page 59-92)

Session 5 – Linear model selection and Regularization I

Learning Objectives:

Explain subset selection and shrinkage methods for linear models

- Lasso
- Ridge
- Elastic net

Implement regularized linear regression in R

Reading:

[ISL] 6.1 Subset selection (Page 205-210)

6.2 Shrinkage methods (Page 215-220)

Session 6 – Linear model selection and Regularization II

Learning Objectives:

Discuss dimension reduction methods for linear models

- Principle components regression
- Partial least squares

Implement the dimension reduction methods in R

Reading:

[ISL] 6.3 Dimension Reduction Methods (Page 230-238)

Assignment:

Homework 1

Session 7 – Meta engines and data preprocessing

Learning Objectives:

Conduct statistical learning using R package "caret":

- Data preprocessing/feature engineering
- Model tuning and comparison using resampling

Reading:

[APM] Chapter 4.9 (page 80-89)

Building Predictive Models in R Using the caret Package (2008). Kuhn. Journal of Statistical Software, 28, 1-26.
The caret Package: http://topepo.github.io/caret/index.html

Session 8 – Moving beyond linearity I

Learning Objectives:

Define and compare regression splines and smoothing splines

Implement spline methods in R

Reading:

[ISL] 7.4 Regression splines (Page 271-277)

7.5 Smoothing splines (Page 277-280)

Session 9 – Moving beyond linearity II

Learning Objectives:

Understand the generalized additive model (GAM) and multivariate adaptive regression spline (MARS)

Implement GAM and MARS in R

Reading:

[ISL] 7.7 Generalized additive model (Page 283-287)

Assignment:

Homework 2

Session 10 – Classification I

Learning Objectives:

List popular methods for classification

Define metrics for evaluating classification performance

- Confusion matrix
- ROC and AUC
- kappa

Review the use of logistic regression in classification and its potential limitations

Reading:

[ISL] 2.2.3 The classification setting (Page 37-39)

Assignment:

Review logistic regression (e.g., 4.3 in ISL)

Session 11 – Classification II

Learning Objectives:

Explain the idea in linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA)

Implement LDA and QDA in R

Compare LDA with logistic regression

Reading:

[ISL] 4.4 Linear discriminant analysis (Page 138-142)

4.5 A comparison of classification methods (Page 151-154)

Assignment:

Homework 3

Session 12 – Tree-based Methods I

Learning Objectives:

Explain the classification and regression trees (CART)

Define the tree terminology

Explain the Pros and Cons of tree models

Implement CART in R and interpret the results

Reading:

[ISL] 8.1 The basics of decision trees (Page 303-309, 311-312)

Session 13 – Tree-based methods II

Learning Objectives:

Explain ensemble methods

- Bagging
- Random forest
- Gradient descent boosting
- AdaBoost

Reading:

[ISL] 8.2 Bagging, random forest, boosting (Page 316-323)

Session 14 – Tree-based methods III

Learning Objectives:

Implement CART, ctree, bagging, random forest and boosting in R

Assignment:

Homework 4

Session 15 – Support vector machine I

Learning Objectives:

Discuss the idea of maximal margin classifier and support vector classifiers

Reading:

[ISL] 9.1 Maximal margin classifier (Page 338-344)

9.2 Support vector classifier (Page 344-346)

Session 16 – Support vector machine II

Learning Objectives:

Explain support vector machine (SVM) and SVM with more than two classes

Implement SVM in R

Reading:

[ISL] 9.3 Support vector machines (Page 349-355)

9.4 SVMs with more than two classes (Page 355-356)

Assignment:

Homework 5

Session 17 – Visualizing and interpreting black-box models

Learning Objectives:

Global interpretation

- Variable importance
- Partial dependence plot
- Individual conditional expectations

Reading:

Lecture notes

Session 18 – Visualizing and interpreting black-box models

Learning Objectives:

Local interpretation

- Local interpretable model-agnostic explanations (lime)

Reading:

The "lime" paper: https://arxiv.org/pdf/1602.04938.pdf

Session 19 – Unsupervised learning I (PCA)

Learning Objectives:

Explain principle component analysis (PCA)

Implement PCA in R

Reading:

[ISL] 10.2 Principle components analysis (Page 374-385)

Session 20 – Unsupervised learning II (clustering)

Learning Objectives:

Explain clustering methods

- K-Means clustering
- Hierarchical clustering

Reading:

[ISL] 10.3 Clustering methods (Page 385-399)

Assignment:

Homework 6

Session 21 – Unsupervised learning III (case study)

Learning Objectives:

Apply PCA and hierarchical clustering on an example dataset (Pokémon)

Interpret the output from R

Reading:

[ISL] 10.6 NC160 data example (Page 407-413)

Session 22 – Neural networks I

Learning Objectives:

Explain neural networks

Discuss practical issues in training neural networks

Reading: (not required)

[ESL] 11.2 Projection pursuit regression

11.3 Neural networks

Session 23 – Neural networks II

Learning Objectives:

Explain basic concepts in deep learning

Implement deep learning in R on a data example (ZIP code data)

Reading:

Lecture notes

Session 24 – Stacked models

Learning Objectives:

Explain stacking and super learner

Implement model stacking using h2o

Reading:

The “Super Learner” paper by van der Laan et al.

The h2o documentation: http://docs.h2o.ai/h2o-tutorials/latest-stable/tutorials/ensembles-stacking/index.html

Session 25 – Concluding remarks

Learning Objectives:

Review the modeling building process

Discuss modeling strategies for the final project

Session 26 – Other topics

Statistical learning for censored data; missing data in machine learning; …

MAILMAN SCHOOL POLICIES AND EXPECTATIONS

Students and faculty have a shared commitment to the School's mission, values and oath. http://mailman.columbia.edu/about-us/school-mission/

Academic Integrity

Students are required to adhere to the Mailman School Honor Code, available online at http://mailman.columbia.edu/honorcode.

Disability Access

In order to receive disability-related academic accommodations, students must first be registered with the Office of Disability Services (ODS). Students who have, or think they may have a disability are invited to contact ODS for a confidential discussion at 212.854.2388 (V) 212.854.2378 (TTY), or by email at disability@columbia.edu. If you have already registered with ODS, please speak to your instructor to ensure that s/he has been notified of your recommended accommodations by Lillian Morales (lm31@columbia.edu), the School's liaison to the Office of Disability Services.

Student Affairs

The Office of Student Affairs (OSA) supports the needs of students who experience life challenges, which may disrupt their successful completion of a Public Health degree. Students' needs may manifest in such areas as their physical, mental, and/or emotional health; economic, family, and/or social stressors; difficulties resulting from adjustment to graduate-level work and/or transitioning to academia after time away from school; as well as other barriers to students' success. Students in need of support should reach out to OSA by phone (212-342-3128), email, or as a walk-in during office hours (8:00 a.m. – 6:00 p.m.; located on the 10th floor of ARB). Students may also directly access the resources and services of Student Health Services, Mental Health, Services, the Center for Student Wellness, and other supportive offices throughout CUMC directly through the offices' websites, links to which can be found on the Health and Wellness page of the Mailman website.