American Public University System
American Military University | American Public University

DATS201

Course Summary

Course : DATS201 **Title :** Analytical Methods I
Length of Course : 8 Weeks
Prerequisites: MATH 302 and MATH 220 **Credit Hours :** 3

Description

This course provides students with the basic toolkit of statistical methods and models that practitioners use for regression, analysis of variance, and linear models. This toolkit could be based on Python or on R. Topics include descriptive statistics/data summaries, inference in simple and multiple linear regression, residual analysis, estimation and testing of hypothesis, transformations, polynomial regression, model building with real data, nonlinear regression and linear models. The course is not mathematically advanced but covers a large volume of material.

Course Scope:

This course is intended for undergraduate students with an interest in data science. It will provide students with the basic, foundational tools required to conduct analyses to find solutions for a wide variety of problems. To conduct analyses, students will learn how to run typical programs such as those listed in the course description. The specific selection of a programming language or code may change from year to year based on current industry standards, which today is R or Python.

Objectives

At the conclusion of this course students will be able to:

- Discuss the basic strategy and application of the tools and methods used in data science for the analysis of data.
- Apply foundational statistical techniques to research data, including testing hypotheses and building regression models.
- Describe the requirements for more advanced analyses in research studies.

Outline

Week 1: Introduction and Statistical learning

Learning Outcomes:

At completion of the modules this week students will be able to:

1. Describe why we need to estimate functions of relationships.
2. Understand the trade-off between prediction accuracy and interpretability.
3. Describe differences between a) supervised & unsupervised as well as
   b) Regression & classification.
4. Determine model adequacy through bias-variance trade-off, classification setting and model fitness.
5. Demonstrate proficiency in fundamental applications in R.

Module 1 – Unit 2.1: Statistical learning

Module 2 – Unit 2.2: Assessing Model Accuracy

Module 3 – Unit 2.3: Introduction to R

Homework I: On pages 52-57, Questions: 1, 3, 7, 9 & 10 Due at the end of the 2nd week, 8/9/2020 @ 11.59pm

Week 2: Linear Regression

Learning Outcomes:

At completion of the modules this week students will be able to:

1. Describe and distinguish the concept of linear and multiple regression clearly identifying the predictors and dependent variables.
2. Distinguish between linear and non-linear relationships.
3. Interpret the coefficients in simple terms and construct confidence intervals.
4. Perform hypotheses testing of regression coefficients and interpret the outcome.
5. Incorporate interaction in models with justification and understand the concept of dummy variables.
6. Use forward and backward selection to select models.
7. Use metrics to assess model's accuracy.
8. Interpret simple R code snippets and functions.
9. Understand the difference between parametric and non-parametric models.

Module 4a – Unit 3.1.1: Estimating coefficients: simple linear regression

Module 4b – Unit 3.1.2: Assessing Accuracy of coefficients estimates in a simple linear regression

Module 5a – Unit 3.2.1: Estimating coefficients: multiple linear regression

Module 5b – Unit 3.2.2: Assessing accuracy of coefficients estimates in a multiple linear regression

Module 6 – Unit 6.1.1-6.1.2: Selecting subset of predictors for use in a multiple linear regression

Module 7 – Unit 3.3.3: Assess Model Accuracy of linear regression

Homework 2: On pages 120-124, Questions: 3, 5, 7, 8 & 10 Due at the end of the 3rd week, 8/16/2020 @ 11.59pm

Week 3: Classification

At completion of the modules this week students will be able to:

1. Articulate the why underlying the use of classification models.
2. Estimate the coefficients in simple terms
3. Understand Bayes' theorem and its applications to discriminant analyses.
4. Distinguish between different classifications methods and their strengths.
5. Understand the assumptions underlying each of the models.

Module 8 – Unit 4.1-4.3: Fit and interpret logistic regression coefficients

Module 9– Unit 4.4.1-4.4.3: Fit and interpret Linear discriminant analyses

Module 10– Unit 4.4.4-4.5: Fit and interpret quadratic discriminant analyses

Module 11– Unit 3.5 & 4.6.5: Fit and interpret KNN models.

Homework 3: On pages 168-172, Questions: 3, 4, 9, 10 & 11 Due at the end of the 4th week, 8/23/2020 @ 11.59pm

Week 4: Resampling Methods

At completion of the modules this week students will be able to:

1. Articulate the need for cross validation and demonstrate its usage.
2. Articulate the need for bootstrapping and demonstrate its usage.

Module 12 – Unit 5.1.1-5.1.2: Leave-one-out cross validation method.

Module 13 – Unit 5.1.3-5.1.4: K-fold cross validation method.

Module 14 – Unit 5.1.5: Application to classification problems.

Module 15 – Unit 5.2: Bootstrapping

Homework 4: On pages 197-201, Questions: 1, 2, 5, 6 & 9, Due at the end of the 5th week, 8/30/2020 @ 11.59pm

Week 5: Linear Model Selection and Regularization

At completion of the modules this week students will be able to:

1. Implement feature selection techniques such including stepwise.
2. Implement feature selection approaches using AIC, BIC, Cp and R squared.
3. Identify when shrinkage methods such as lasso and ridge regression are appropriate and how to apply them.
4. Perform feature engineering using Principal Component Analysis.

Module 16 – Unit 6.1.1: Subset (feature) selection and choosing the optimal model

Module 17 – Unit 6.2, 6.5.3: Validation, cross validation and shrinkage (ridge and lasso regression) methods

Module 18 – Unit 6.3.1: Principal Component Analyses

Module 19 – Unit 6.3.2: Partial least squares

Homework 5: On pages 259-264, Questions: 1, 4, 5, 9 & 10, Due at the end of the 6th week, 9/6/ 2020 @ 11.59pm

Week 6: Tree-Based Methods

At completion of the modules this week students will be able to:

1. Explain the concept of decision trees and the need for pruning.
2. Understand the concept of bagging and its application in random forest models.
3. Understand boosting and how it differs from bagging.
4. Use variable importance as a decision making tool.

Module 20 – Unit 8.1: Introduction to decision trees and pruning

Module 21 – Unit 8.2.1-8.2.2: Bagging and Random forests

Module 22 – Unit 8.2.3: Boosting and variable importance

Homework 6: On pages 332-335, Questions: 3, 5,8, 9 & 11, Due at the end of the 7th week, 9/13/ 2020 @ 11.59pm

Week 7: Unsupervised Learning

At completion of the modules this week students will be able to:

1. When to favor unsupervised over supervise learning.
2. Understand the use of PCA as an unsupervised model.
3. Understand and apply K-means clustering.
4. Understand and apply hierarchical clustering.

Module 23 – Unit 10.1-10.2: Exploring Principal Component Analyses, proportion of variance and interpretations.

Module 24 – Unit 10.3.1: Exploring K-means and interpretations.

Module 25 – Unit 10.3.2: Exploring Hierarchical models, dendrograms and interpretations.

Module 26 – Unit 10.3.3: Practical Issues in Clustering.

Homework 7: On pages 416-418, Questions: 8, 9, 10 & Review, Due at the end of the 8th week, 9/20/2020 @ 11.59pm

Week 8: Wrapping it all up

At completion of the modules this week students will be able to:

1. Describe different data types, outcomes and models that can be applied to them.
2. Distinguish between different types of models, interpret coefficients where they exist and correctly perform diagnostics using appropriate tools.
3. State and test hypotheses.

Module 27: Discuss modeling strategies, testing hypotheses and requirements for advanced analytics in research studies.

Module 28: Course review

Module 29: Practice Final Exam

Module 30: Final Exam

Evaluation

A variety of weekly exercises (Knowledge Checks) will be used to reinforce the material covered in this course. This course will include a final exam, weekly homework, quizzes, discussions and labs (that are not required but highly recommended). The grade distribution of the course is shown below.

- Homework 25%
- Laboratories 0%
- Discussions 15%
- Quizzes 30%
- Final Exam 30%

Materials

Textbook (Required)

Free open source textbook will be used throughout this course. This includes videos, course notes and a variety of articles. http://faculty.marshall.usc.edu/gareth-james/ISL/ . Extra materials will be made available to students for contents outside the prescribed textbook.

Software (Required)

1. Install R and RStudio on windows from: https://cran.r-project.org/bin/windows/base/ and https://rstudio.com/products/rstudio/download/ (scroll down to All installers and select windows version)
2. Install R and Rstudio on mac from: https://cran.r-project.org/bin/macosx/ and https://rstudio.com/products/rstudio/download/ (scroll down to All installers and select mac version)

Software (Optional)

1. Python version 3 on mac and windows available at: https://www.python.org/downloads/
2. Other software usage will require approval from the instructor.

Late Assignments

1. Students are expected to submit classroom assignments by the posted due date and to complete the course according to the published class schedule. The due date for each assignment is listed under each Assignment.
2. Generally speaking, late work may result in a deduction up to 15% of the grade for each day late, not to exceed 5 days.
3. As a working adult I know your time is limited and often out of your control. Faculty may be more flexible if they know ahead of time of any potential late assignments.

Course Guidelines

[Leave this section blank. It is a standard script used for all courses in our department listing the citation style, late policy, DSA and other APUS links, etc.]