STAT 3260: Statistical Learning with Application in R

2023 Summer Session	
Total Class Sessions: 25 **Class Sessions Per Week: 5** **Total Weeks: 5** **Class Session Length (Minutes): 145** **Credit Hours: 4**	**Instructor: Staff** **Classroom: TBA** **Office Hours: TBA** **Language: English**

Course Description:

This course introduces students to the various statistical learning methods and their applications in statistics modeling. Topics discussed includes review on elementary probability and statistics learning, principles and different types of statistical learning, simple and multiple linear regression, regression models, time series models, principal components analysis, decision trees, and cluster analysis.The statistical programming language R will be introduced.

Textbook:

"An Introduction to Statistical Learning with Applications in R," James, Witten, Hastie, and Tibshirani, 2013, New York: Springer.

Course Format and Requirements:

Class time will be used for a combination of lectures, class discussions, lab assignments and student design project presentations. There will be a series of lab material component and associated assignments in addition to the regular in class lectures.

Attendance:

Attendance at lectures is vital to get a thorough understanding of the material. This course also requires lab attendance and completion of assignments. Students must be present and actively involved in class discussions.

Course Assignments:

Quizzes

Weekly quizzes will usually consist of short answer questions and or short essay questions. No make-up quiz will be given.

Exams (One Midterm and One Final)

Exams may not be taken early, made-up, or turned in late. Students must comply with all applicable instructions to receive credit. The exams will include discussion questions and case problems. During the exams, each student must work individually without consulting others.

Course Project

Students will actively use their knowledge to present a project with the consultation with the instructor. The project might be done individually or in 2-4 persons groups.

Homework Lab and Theory Assignments There will be weekly set of assignments to complete at home. The instructor will select and distribute during each class meeting a selection of mostly quantitative problems from each appropriate chapter.

Course Assessment:

Quizzes	- 15%
Homework Assignments	- 20%
Midterm	- 20%
Course Project	- 15%
Final Exam	- 30%

Grading Scale (percentage):

A+	A	A-	B+	B	B-	C+	C	C-	D+	D	D-	F
98-100	93-97	90-92	88-89	83-87	80-82	78-79	73-77	70-72	68-69	63-67	60-62	<60

Academic Integrity:

Students are encouraged to study together, and to discuss lecture topics with one another, but all other work should be completed independently.

Students are expected to adhere to the standards of academic honesty and integrity that are described in the Chengdu University of Technology's Academic Conduct Code. Any work suspected of violating the standards of the Academic Conduct Code will be reported to the Dean's Office. Penalties for violating the Academic Conduct Code may include dismissal from the program. All students have an individual responsibility to know and understand the provisions of the Academic Conduct Code.

Special Needs or Assistance:

Please contact the Administrative Office immediately if you have a learning disability, a medical issue, or any other type of problem that prevents professors from seeing you have learned the course material. Our goal is to help you learn, not to penalize you for issues which mask your learning.

Tentative Course Schedule:

Class 1: Introduction

Topics: Review of the syllabus, introduction to the scope of the course and the overview of the textbook, including a short refresher of basic statistical terminology. Intro to various statistical data; brief history of statistical learning; statistical learning tools.

Class 2:
Material: Chapter 2. Statistical Learning (Section 2.1) What is statistical learning and section 2.3 (Lab) Introduction to R.

Topics: Why and how to estimate function "f;" the trade-off between prediction accuracy and model interpretability; supervised versus unsupervised learning; regression versus classification problems.
Lab part: introduction to R; basic commands and graphics. See and download: https://cran.r-project.org/

Assignments to do (homework):
Theory exercises: 2.4.1, 2.4.2, and 2.4.5
Lab exercises: 2.4.8

Class 3:
Material: Chapter 2. Statistical Learning (Section 2.2) and continuing with section 2.3 (Lab).

Topics: Discussion of how to assess model accuracy; measuring the quality of Fit; the bias variance trade-off; the classification setting.
Lab part: Indexing and loading data; graphical and numerical summaries

Assignments to do (homework):
Theory exercises: 2.4.3, 2.4.4, and 2.4.7
Lab exercises: 2.4.9 and 2.4.10

Class 4:
QUIZ no. 1 (Chapters 2)

Material: Chapter 3. Linear Regression (Section 3.1 – Simple Linear Regression), and section 3.6 (Lab)

Topics: Understanding simple linear regression; estimating the coefficients; assessing the accuracy of coefficients estimates; assessing the accuracy of the model.

Lab part: 3.6.1 and 3.6.2 - Libraries and Simple linear regression

Assignments to do (homework):

Theory exercises: 3.7.1 and 3.7.2

Lab exercises: 3.7. 8 and 3.7.12

Class 5:

Material - Chapter 3. Linear Regression (Section 3.2 – Multiple Linear Regression), and section 3.6 Lab, continuing)

Topics: Understanding multiple linear regression; estimating regression coefficients; analysis if there is a relationship between the response and predictors; how to decide on important variables; measures of model fit and predictions.

Lab part: 3.6.3 – Multiple linear regression

Assignments to do (homework):

Theory exercises: 3.7.3

Lab exercises: 3.7.9 and 3.7.14

Class 6:

Material: Chapter 3. Linear Regression (Section 3.3 – Other Considerations in Regression Model), and section 3.6 Lab, continuing.

Topics: Qualitative predictors; extensions of linear model; analysis of potential problems.

Lab part: 3.6.4 to 3.6.6; interaction terms, non-linear transformation of the predictors, qualitative predictors.

Assignments to do (homework):

Theory exercises: 3.7.4 and 3.7.6

Lab exercises: 3.7.10-12

Class 7:

Material: Chapter 3. Linear Regression (Sections 3.4 – The Marketing Plan and 3.5 – Comparison of Linear Regression with K-Nearest Neighbors, and section 3.6 Lab, continuing.

Topics: Review of the data and analysis of various regression methods as used in hypothetical marketing plan; parametric and non-parametric methodology.
Lab part: 3.6.7 – Writing functions

Assignments to do (homework):
Theory exercises: 3.7.2, 3.7.5, and 3.7.7
Lab exercises: 3.7.11 and 3.7.15

Class 8:

QUIZ no. 2 (Chapters 3)

Material: Chapter 4. Classification (Sections 4.1 – Classification and 4.2 - Why Not Linear Regression), and section 4.6 Lab.

Topics: Analysis of classification issues and problems; linear regression versus qualitative response.
Lab part: N/A

Assignments to do (homework):
Theory exercise: 4.7.1
Lab exercise: 4.7.12

Class 9:

Material: Chapter 4 Classification (Section 4.3 – Logistic Regression), and section 4.6. Lab, continuing.

Topics: Discussion of the logistic regression; the logistic model; estimating regression coefficients, making predictions; multiple logistic regression; logistic regression for more than two response classes.
Lab part: 4.6.1 and 4.6.2: The stock market data and logistic regression.

Assignments to do (homework):

Theory exercises: 4.7.6, 4.7.7, and 4.7.9

Lab Exercises: 4.7.10 (a-d) and 4.7.11 (a-c)

Class 10:

Material: Chapter 4. Classification (Section 4.4 – Linear Discriminant Analysis LDA and QDA, section 4.5 – Comparison of Classification Methods, and section 4.6 Lab, continuing.

Topics: Concept of linear discriminant analysis; using Bayes' theorem for classification; linear discriminant analysis for p=1 or for p>1; quadratic discriminant analysis.

Lab part: 4.6.3 to 4.6.6: Linear discriminant analysis, quadratic discriminant analysis, K-nearest neighbor, an application to Caravan Insurance data

Assignments to do (homework):

Theory exercises: 4.7.8

Lab exercises: Remaining sections of 4.7.10 and 4.7.11

Class 11:

Material: Chapter 5. Resampling Methods (Section 5.1 Cross-Validation), and section 5.3 Lab.

Topics: Analysis of the validation set approach; leave-one-out validation method; k-Fold cross-validation method; bias-variance trade-off for k-Fold cross validation; using cross validation on classification problems.

Lab part: 5.3.1 to 5.3.3: The validation set approach, leave-one-out cross validation, k-Fold cross validation

Assignments to do (homework):

Theory exercises: 5.4.1, 5.4.3, and 5.4.4

Lab exercises: 5.4.5, 5.4.6, and 5.4.8

Class 12:

Material: Chapter 5. Resampling Methods (Section 5.2 The Bootstrap) and section 5.3 Lab, continuing.

Topics: The understanding and use of powerful statistical tool – the Bootstrap.

Lab part: 5.3.4 – the Bootstrap

Assignments to do (homework):
Theory exercises: 5.4.2
Lab exercises: 5.4.7 and 5.4.9

Class 13:

QUIZ no. 3 (chapters 4-5)

Material: Chapter 6. Linear Model Selection and Regularization, (section 6.1. Subset Selection) and section 6.5, Lab.

Topics: Analysis of the best subset selection methodology; stepwise selection; choosing the optimal model.
Lab part: 6.5.1 to 6.5.3: Best subset selection, forward and backward stepwise selection, choosing among models using the validation set approach and cross validation.

Assignments to do (homework):
Theory exercises: 6.8.1
Lab exercises: 6.8.8

Class 14:
Material: Linear Model Selection and Regularization, (section 6.2, Shrinkage Methods, RSS, and the Lasso), and section 6.5 Lab, continuing.

Topics: The understanding and usage of Ridge Regression and the Lasso method; selecting the tuning parameter.
Lab part: 6.6.1 and 6.6.2: Ridge Regression and the Lasso

Assignments to do (homework):
Theory exercises: 6.8.2, 6.8.3, and 6.8.4
Lab exercises: 6.8.9 (a-d), 6.8.10

Class 15:
Material: Linear Model Selection and Regularization, (section 6.3. Dimension Reduction Methods, section 6.4 Considerations in High Dimensions), and section 6.5 Lab, continuing.

Topics: Understanding dimension reduction methods; principal components regression (PCR); partial least square (PLS); what do we need to consider in high dimensions – data, what could go wrong, regression, and interpretation of results.

Lab part: 6.7.1 and 6.7.2: Principal Components Regression (PCR) and Partial Least Squares (PLS).

Assignments to do (homework):

Theory exercises: 6.8.5 to 6.8.7

Lab exercises: 6.8.12 and 6.8.13

Class 16:

Review for Midterm Exam

Class 17:

MIDTERM (Chapters 1-6)

Class 18:

Material - Chapter 8. Tree-Based Methods (Section 8.1 – The Basics of Decision Trees), and Section 8.3 Lab.

Topics: Understanding decision trees methodology; regression trees; classification trees; trees versus linear methods; advantages and disadvantages of trees methods.

Lab part: 8.3.1 and 8.3.2: Fitting classification and regression trees

Assignments to do (homework):

Theory exercises: 8.4.1 and 8.4.3

Lab exercises: 8.4.7, 8.4.8 (a-c), 8.4.9

Class 19:

Material - Chapter 8. Tree-Based Methods (Section 8.2 – Bagging, Random Forests and Boosting), and Section 8.3 Lab, continuing.

Topics: Understanding and use of bagging, random forests and boosting in tree-based methodology.

Lab part: 8.3.3 and 8.3.4: Bagging and Random Forests and Boosting

Assignments to do (homework):

Theory exercises: 8.4.2., 8.4.4, and 8.4.5

Lab exercises: 8.4.8 remining sections, 8.4.12

Class 20:

Quiz 4 (Chapter 8)

Material - Chapter 9. Support Vector Machines (Section 9.1 – Maximal Margin Classifier, Section 9.2 – Support Vector Classifiers, Section 9.3 – Support Vector Machines, and Section 9.6 Lab.

Topics: Understanding and using hyperplane; the maximal margin classifier and its construction; non-separable cases; overview and details of support vector classifiers; classification with nonlinear decision boundaries; the support vector machines and its application.

Lab part: 9.6.1, 9.6.2: Support vector machines and classifiers.

Optional Assignments:

Theory exercises: 9.7.1 to 9.7.3

Lab exercises: 9.7.4

Class 21:

Material – Chapter 9. Support Vector Machines (Section 9.4 – Support Vector Machines with More than Two Classes, Section 9.5 – Relationship to Logistic Regression, and Section 9.6 Lab, continuing

Topics: Learning about support vector machines one-versus-one classification and one-versus-all classification; SVMs relationship to logistic regression.

Lab part: 9.6.3 to 9.6.5: ROC curves, SVMs with multiple classes, application to gene expression data

Optional Assignments:

Theory exercises: n/a

Lab exercises: 9.7.5 to 9.7.7

Class 22:

Material: Chapter 10. Unsupervised Learning (Section 10.1 – The Challenge of Unsupervised Learning, Section 10.2 – Principal Component Analysis (PCA), and Section 10.4 Lab.

Topics: Principal Components Analysis (PCAs), including various interpretations of PCAs, and diverse uses of them.
Lab part: 10.4.1 – Principal Clustering Analysis

Optional Assignments:
Theory exercises: n/a
Lab exercises: 10.7.8 and 10.7.10 (a-b)

Class 23:
Material: Chapter 10. Unsupervised Learning (Section 10.3 – Clustering Methods), and Section 10.4 Lab, continuing.

Topics: K-means and hierarchical clustering methods and their applications
Lab part: 10.5.1 and 10.5.2: K-means clustering and hierarchical clustering

Optional Assignments:
Theory exercises: 10.7.1, 10.7.2, and 10.7.4
Lab exercises: 10.7.9 and 10.7.11

Class 24:

QUIZ no. 5: (Chapters 9-10)

Presentation of the course project, Course Project Due

Class 25:
Review for Final Exam

FINAL EXAM(Cumulative) TBA