[STA 6714] – [Data Preparation]
[Spring] – [Statistical Computing] – [3 Credit Hours]

INSTRUCTOR:	Morgan C. Wang
PHONE:	407-823-2818
OFFICE LOCATION:	TC II 203
OFFICE HOURS:	Monday and Wednesday 1:30 AM to 4:00 PM
E-MAIL:	Chung-Ching.Wang@ucf.edu

WITHDRAW DATELINE:	March 23, 2016
HOLIDAYS:	January 18 and March 7 to March 12
SPECIAL NOTES:	**Students who are not officially registered in the class will not have exams graded or returned.**

LEARNING OBJECTIVES:

At the end of the course, students will be able to:

- Learning Objective 1: Use SAS/Enterprise Miner, Text Miner, and SAS/STAT software effectively
- Learning Objective 2: Preparing Structure Data for Modeling Tools such as Regression, Support Vector Machine, and Neural Network
- Learning Objective 3: Preparing Non-structure Text Data
- Learning Objective 4: Preparing High Dimensional Categorical Data using Linking Technique
- Learning Objective 5: Preparing Time Series Data (Optional)
- Learning Objective 6: Combine Structural and Non-Structural Data for Modeling

COURSE DESCRIPTION:

Variable reduction, variable clustering, missing value imputation, and data survey. Additional data preparation topics associated with data mining techniques. Prerequisite: STA 5104

PURPOSE OF THE COURSE:

Data mining is a process of selecting, exploring, and modeling large amounts of data to uncover previously unknown patterns hiding in the data. However, data used in this mining process are typically observed without thinking to mine them later. Therefore, the data collected are not ready for any meaningful mining exercise. The purpose of this course is to introduce techniques used to prepare data to ensure that after this process that data is ready to be mined. Well-prepared data can ensure any data mining techniques to work much more efficient. Consequently, the mining process can obtain better results. This course will address data preparation issues with some working projects of the instructor and basic concepts and principles from case studies of other professionals in data mining field. We will use case study approach to deal with several center problems on data preparation. In addition, we will cover

some data mining techniques briefly because data preparation task depends on the data mining techniques as well.

REQUIRED MATERIALS:

Required Text

- Lecture Notes from the instructor
- Mamdouh Refaat (2007) Data Preparation for Data Mining Using SAS, Morgan Kaufmann Publishing: San Francisco, CA.
- David Hand, Heikki Mannila, and Smyth (2001) Principles of Data Mining, Massachusetts Institute of Technology, London, England.

Supplemental Materials

- Selected articles by the instructor
- Thomas H. Davenport and Jeanne G. Harris (2007) Competing on Analytics, Harvard Business School Press, Boston.
- David J. Hand (1999) Construct and Assessment of Classification Rules, John Wiley & Son, New York.
- William H. Inmon and Anthony Nesavich (2008) Tapping into Unstructured Data, Prentice Hall, Boston, MA.

ATTENDANCE POLICY:

There are pop up quizzes every week and students who miss pop up quiz without advanced permission from the instructor will get zero score of that quiz. Students who miss two or more quizzes without adequate excuse and advance permission from the instructor will get 25 additional points deduction from their final grade. It is your responsibility to attend all classes and to notify instructor all absences in advance and provide instructor the related documents.

MAKE-UP EXAM POLICY:

Make-up exams will be allowed only in extreme instances and with advanced permission of the instructor. It is the student's responsibility to work with faculty to notify them of an excused absence (e.g. work, illness) and to coordinate a make-up exam. EDC staff are available to administer and proctor make-up exams during the EDC's regular operating hours.

ACADEMIC INTEGRITY:

All students are expected to abide by the University's Code of Student Conduct.

Plagiarism and Cheating of any kind on an examination, quiz, or assignment will result at least in an "F" for that assignment (and may, depending on the severity of the case, lead to an "F" for the entire course) and may be subject to appropriate referral to the Office of Student Conduct for further action. See the UCF Golden Rule for further information.

GRADING SCALE:

LETTER GRADE	PERCENTAGE
A	92.5-100 %
A-	90.0-92.4 %
B+	87.0-89.9%
B	83.0-86.9%
B-	80.0-82.9%
C+	77.0-79.9%
C	73.0-76.9%
C-	70.0-72.9%
D+	67.0-69.9%
D	63.0-66.9%
D-	60.0-62.9%
F	Less than 59.9%

GRADING PROCEDURES:

All quizzes in this semester worth 30% of the final grade. Each quiz worth approximately 30 points. All assignments in this semester worth 20% of the final grade. Each assignment worth approximately 30 points as well.

ASSIGNMENT	% OF GRADE
Homework Assignments	20%
Exam 1 / Midterm Project (March 16, 2015 6:00 PM to 7:15 PM)	25%
Weekly Pop Quizzes	30%
Exam 2 / Final Project (May 4, 2016 4:00 PM to 6:50 PM)	25%
Total	100%

ASSIGNMENT DESCRIPTIONS:

Pop up quiz will be given at least once each week. After the completion of each lecture, the lecture material will be covered in the weekly pop quiz. Exam I covers first six lectures. Exam II covers all lectures. Assignments will be given six times during the semester after week #3. All assignments are data preparation projects using techniques presented in class.

COURSE OUTLINE:

- Lecture 1: Using Enterprise Miner/R
- Lecture 2: Data Visualization and Presentation
- Lecture 3: Identify Data Problems
- Lecture 4: Using Categorical Data with Many Levels (Less than 1000 levels)
- Lecture 5: Variables with Missing Values
- Lecture 6: Transformation of Numerical Variables
- Lecture 7: Outliers Detection and Treatment
- Lecture 8: Variables Selection
- Lecture 9: Text Miner Nodes Overview
- Lecture 10: Text Data Preparation using Text Miner
- Lecture 11: Model Assessment and Optimal Decision
- Lecture 12: Data Mining Flow – A Case Study
- Lecture 13: Transactional Data Preparation and Making Usage of Geographic Variables
- Lecture 14 (Optional): Use Categorical Variables with More than 1000 Levels
- Lecture 15 (Optional): Time Series Data Preparation

Note: Syllabus subject to change based on needs of students, University, and instructor. ***All material covered in class, regardless of whether material is listed, is fair to be tested***.

DISABILITY STATEMENT

The University of Central Florida is committed to providing reasonable accommodations for all persons with disabilities. This syllabus is available in alternate formats upon request. Students with disabilities who need accommodations in this course must contact the professor at the beginning of the semester to discuss needed accommodations. No accommodations will be provided until the student has met with the professor to request accommodations. Students who need accommodations must be registered with Student Disability Services, Ferrell Commons, 7F, Room 185, phone (407) 823-2371, TTY/TDD only phone (407) 823-2116, before requesting accommodations from the professor.

COPYRIGHT

This course may contain copyright protected materials such as audio or video clips, images, text materials, etc. These items are being used with regard to the Fair Use doctrine in order to enhance the learning environment. Please do not copy, duplicate, download or distribute these items. The use of these materials is strictly reserved for this online classroom environment and your use only. All copyright materials are credited to the copyright holder.

***NOTE**: For additional sample syllabi information, including copy ready syllabi clauses related to using Webcourses, Turnit in, etc., please visit UCF's Faculty Center for Teaching & Learning: http://www.fctl.ucf.edu/TeachingAndLearningResources/CourseDesign/Syllabus/statements.php#ethics