Logistics

CSE 5334 Data Mining

Won Hwa Kim

Department of Computer Science and Engineering, University of Texas at Arlington, Spring 2020

Instructor

Won Hwa Kim

- Assistant Professor, CSE
- Office: SEIR 324
- won.kim@uta.edu
- http://ranger.uta.edu/~wonhwa

Research Interests

- Machine Learning, Computer Vision, Data Science, Neuroimaging

Instructor

#324 SEIR: access required to get to 3rd floor

Basics

Lectures

- Mon/Wed 1:00-2:20pm, NH 111

Office hours

- Mon 3:00pm-4:00pm, SEIR 324

TA

- Tong Feng (Ph.D student in CSE at UTA)

Preparation/Expectation

- Be hands-on and have good programming experience
  - You are expected to use either or both Python and Matlab
- Be comfortable with topics in your math and statistics
  - Must be familiar with **Linear Algebra**, **Calculus** and **Probability Theory**
- Expect heavy workload, challenging assignments, exams
  - Be hard-working; expect to spend many, many hours; likely your heaviest course.
  - Exam is demanding; almost no student can finish all exam questions.
- Equal Opportunities and Subsequent Results
  - No Extra Credit Assignments
  - Same standard for everyone in grading

Academic Integrity

Violations

- Cheating on test/assignment; Plagiarism; Collusion

Can I refer to external materials?

- Yes, but in your homework, source code, and documentation you must explicitly acknowledge the source of information.
- If you copy sentences (completely or partially) from other places, you must enclose them with quotation marks, in addition to provide references to the information source.
- Even if you rephrase, you still need to acknowledge the source.

Academic Integrity

What types of discussions are allowed?

- You can discuss topics related to assignments with your fellow students.
- But you cannot discuss your solutions.
- You must not provide your work (email, hard copy, or in any form) to anyone for any purpose. Following actions are not acceptable:

  "I emailed it to my roommate/friend so that I can submit from their computer, since I couldn't get online from mine."

  "I sent it to my roommate/friend so that I can compile and test my program on their computer, since mine was down."

Academic Integrity

Tutorial: http://library.uta.edu/plagiarism/index.php

More information at http://www.uta.edu/conduct/academic-integrity/index.php

The chance of being caught is large; we use tools to diligently check and compare the documents and source codes that you submit to us.

The consequence is certain:

- I will submit the form of "faculty referral of honor code violation" to the university. No exception!
- Academic penalty in the context of this course: 0 on assignment/exam, reduced grade, failing grade of the course
- Penalty by the university: probation, suspension, expulsion, …

Textbooks

- (Required) Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Introduction to Data Mining. (Sample chapters at http://www-users.cs.umn.edu/~kumar/dmbook/index.php)
- (Required for relevant chapters) Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. Introduction to Information Retrieval. (Free book at http://nlp.stanford.edu/IR-book/)
- (Required for relevant chapters) Tom Mitchell, Machine Learning (Free book at https://www.cs.ubbcluj.ro/~gabis/ml/ml-books/McGrawHill%20-%20Machine%20Learning%20-Tom%20Mitchell.pdf/)
- (Reference) Jure Leskovec, Anand Rajaraman and Jeff Ullman. Mining of Massive Datasets. (Free book at http://www.mmds.org/#ver21)
- (Reference) Jiawei Han, Micheline Kamber and Jian Pei. Data Mining: Concepts and Techniques.
- (Reference) Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. An Introduction to Statistical Learning with Applications in R. (Free book at http://www-bcf.usc.edu/~gareth/ISL/index.html)
- (Reference) I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.

The Slides

- The slides highlight the gist of most important concepts and techniques.
  - It is **not meant to be complete**. Details may not be included.
  - It may be simplified for ease of explanation.
- Only studying the slides is not enough.
- Many lecture notes are adopted from
  - Vipin Kumar (Minnesota)
  - Jiawei Han (Illinois)
  - Mark Craven / David Page (Wisconsin)

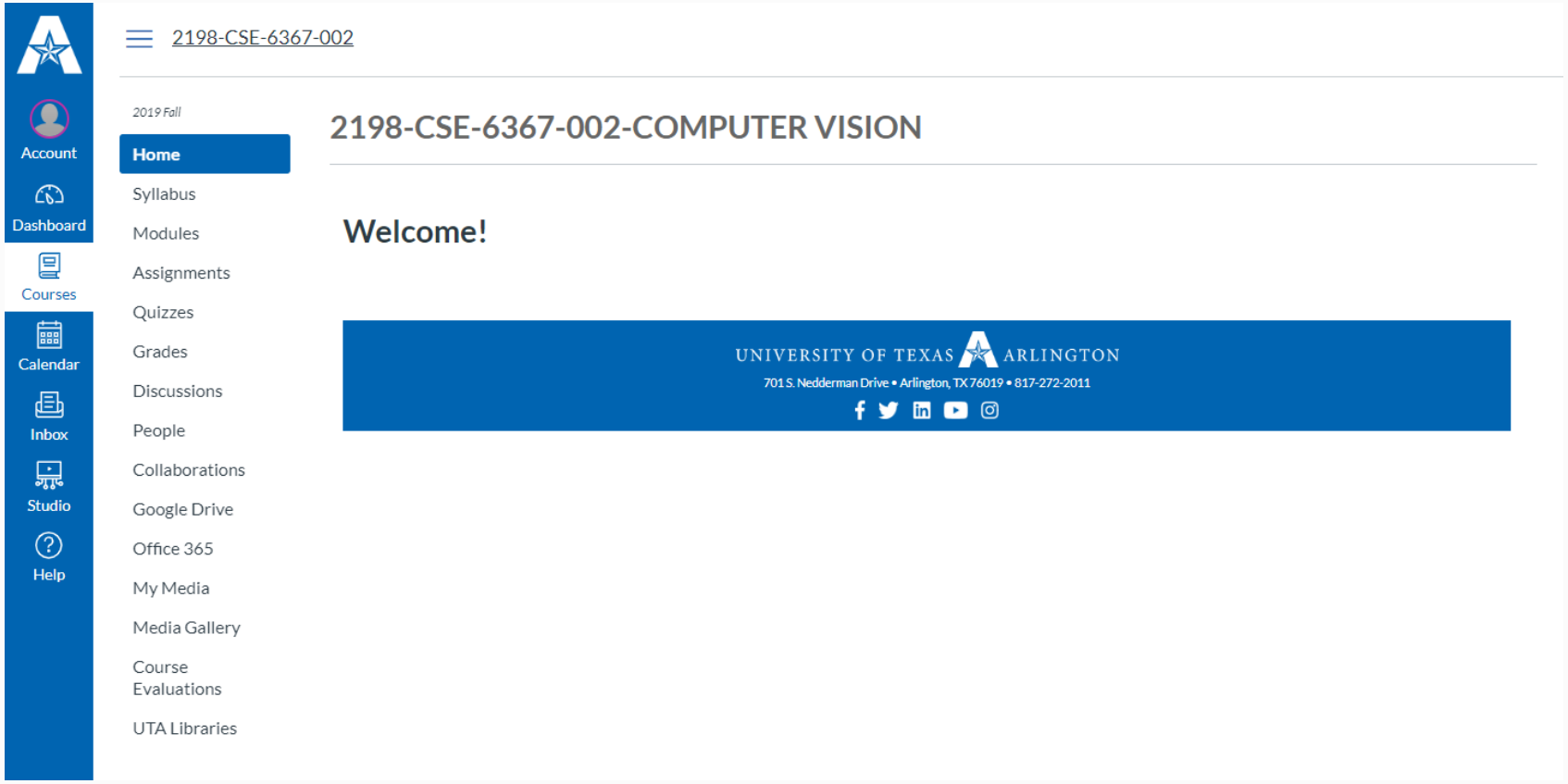
Tentative Grading Scheme

- Assignments (P) 30% (Must be done independently)
- (Pop) quizzes (Q) 40%
- Final 30%
  - 3 Assignments expected.
  - 3 quizzes expected
- You are required to attend classes (consequences for missing classes are yours.)
- All assignments must be electronically prepared (in Word or Latex). We won't accept images of handwritten answers and hand-drawn pictures.
- Final Exam: Time and Location: TBD
- Final Letter Grade:
  - No pre-defined cutoffs. Will be based on curve of your performance.

Canvas

- Assignment instruction (Please READ!)
- Submission (we don't accept email submission or hard-copy)
- Grades
- Questions, Discussion Forum

Canvas



Deadlines

- Everything will be submitted through Canvas.
- Due time: 11:59pm
- Late days: you have 2 late days **throughout the semester**, so use them wisely!
- Late submission: points will be deducted by certain percentages based on the number of late days

Regrading

- Within 7 days we post scores on Blackboard, TA will handle regrade requests. Won't consider it after 7 days.
- If not satisfied with the results, 7 days to request again. Instructor will handle it, and the decision is final.
- No Exception

Your Email

- Make sure your UTA email account works.
- We will only contact you by your UTA email. Check it on a daily basis.

We will cover…

- Basics of Data Science
- Data in different forms
- Data processing methods
- Data analysis methods