UNIVERSITY OF TEXAS ARLINGTON

CSE4334/5334 Data Mining

Logistics

Fall 2020

Chengkai Li

Self Introduction

Chengkai Li

- Professor and Associate Chair, CSE
- Office: ERB 628
- cli@uta.edu
- https://idir.uta.edu/cli.html

Research Interests

- Big Data Intelligence and Data Science (data management, data mining, natural language processing, machine learning)

Basics

Microsoft Teams

- Class team on Microsoft Teams: use code 46aqhj8 to join if you are not already in

Lectures

- Mon/Wed 2:30-3:50pm, Online synchronous lecturing through Microsoft Teams

Chengkai Li Office hours

- Mon/Wed 4-5pm, Instructor's Office hours through Microsoft Teams

TA

- Mohammed Samiul Saeef, mohammedsamiul.saeef@mavs.uta.edu

TA Office hours

- Tue/Fri 12-1pm, TA's Office hours through Microsoft Teams

Teaching Modality

- The course modality is Hybrid for 4334-003 and 5334-002 and Online for 5334-902. The instructor intends to use online as much as possible. It is possible that all learning activities, quizzes, and exams will be online. However, if situation changes, the instructor may conduct exams, quizzes, and even lectures in person.

- We will use Respondus Monitor and Lockdown Browser to administer quizzes and exams. and webcam for quizzes and exams. Students will need a webcam, a microphone, and Internet access.

Preparation/Expectation

- Solid coding skills
  - Multiple significant programming assignments
  - You are expected to use Python
- Be comfortable with topics in your math, statistics, probability courses
- Expect heavy workload, challenging assignments, exams
  - Be hard-working; expect to spend many hours; likely one of your heaviest courses.
  - Exam is demanding and comprehensive.
- Plagiarism is absolutely not tolerated. No excuse or second chance.

Academic Integrity

Violations

- Cheating on exams/quizzes; 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 providing references to the information source.
- Even if you paraphrase, you still need to acknowledge the source.
- If you copy source codes (completely or partially) from other places, you must provide references to the information 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) [TSKK] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. Introduction to Data Mining, 2nd ed., Pearson, 2019. (Sample chapters at http://www-users.cs.umn.edu/~kumar/dmbook/index.php)
- (Required for relevant chapters) [MRS] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. Introduction to Information Retrieval, Cambridge University Press. 2008. (Free book at http://nlp.stanford.edu/IR-book/)
- (Reference) Jure Leskovec, Anand Rajaraman and Jeff Ullman. Mining of Massive Datasets, 3rd ed., Cambridge University Press, 2020. (Free book at http://www.mmds.org/)
- (Reference) Jiawei Han, Micheline Kamber and Jian Pei. Data Mining: Concepts and Techniques, 3rd ed. (2nd edition is also fine), Morgan Kaufmann Publishers, June 2011.
- (Reference) Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. An Introduction to Statistical Learning with Applications in R, 1st ed., Springer, 2013. (Free book at http://faculty.marshall.usc.edu/gareth-james/ISL/)
- (Reference) I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 4th ed. 2016.

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

- (Required) [TSKK] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. Introduction to Data Mining, 2nd ed., Pearson, 2019. (Sample chapters at http://www-users.cs.umn.edu/~kumar/dmbook/index.php)
- (Required for relevant chapters) [MRS] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. Introduction to Information Retrieval, Cambridge University Press. 2008. (Free book at http://nlp.stanford.edu/IR-book/)
- (Reference) Jiawei Han, Micheline Kamber and Jian Pei. Data Mining: Concepts and Techniques, 3rd ed. (2nd edition is also fine), Morgan Kaufmann Publishers, June 2011.

Tentative Grading Scheme

- Pop quizzes (30%): best 6 out of 8 quizzes
- Programming Assignments (30%): must be done independently
- Midterm Exam (15%): October 14th, Wednesday, 2:30pm-3:50pm
- Final Exam (25%): December 14th, Monday, 2pm-4:30pm

- You are required to attend classes and actively participate in discussions.

- All assignments must be electronically prepared. We won't accept images of handwritten answers and hand-drawn pictures.

- Final Letter Grade:
  - No pre-defined cutoffs. Will be based on bell curve of your performance.
  - Undergraduate and graduate students are compared in separate groups.

Final Exam

- Time: December 14th, Monday, 2pm-4:30pm
- We will use Respondus Monitor and Lockdown Browser to administer quizzes and exams. and webcam for quizzes and exams. Students will need a webcam, a microphone, and Internet access.
- Comprehensive (25% from first-half of semester; 75% from second-half of semester)

Canvas

- Video recordings of lectures
- Assignment instruction and files
- Submission (we don't accept email submission or hard-copy)
- Grades
- Questions, discussion forum
- Pop quizzes and exams (Respondus Monitor and Lockdown Browser)

- Mobile app of Canvas

Deadlines

- Everything will be submitted through Canvas.
- Due time: 11:59pm
- Late submission: 5-point deduction per hour, till you get 0. 12:01am –5, 1:01am -10, … (The raw score of each assignment is 100. So there is no point to submit it after 19 hours).

Regrading

- 7 days after we post scores on Canvas. 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.

Your Email

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

A few interesting Data Mining related questions to consider …

(1) During World War II, researchers at the Center for Naval Analysis faced a critical problem. Many bombers were getting shot down on runs over Germany. The naval researchers knew they needed hard data to solve this problem and went to work. After each mission, the bullet holes and damage from each bomber was painstakingly reviewed and recorded. The researchers poured over the data looking for vulnerabilities.(Syed, Matthew. 2015. Black Box Thinking. New York: Penguin Random House. pp 33-37) The data began to show a clear pattern. Most damage was to the wings and body of the plane. The solution to their problem was clear. Increase the armor on the plane's wings and body. **But there was a problem. The analysis was completely wrong.**
(https://www.trevorbragdon.com/blog/when-data-gives-the-wrong-solution)

The analysis didn't consider planes that were shot down. Most surviving bombers didn't get damage in areas such as cockpit, engine, and part of the tail. Those are actually the most vulnerable areas.

This counter-intuitive phenomenon is called by data sampling bias.

(2) In May 2018 the CSE department graduated 95 undergraduate students and 50 graduate students. 50 of the undergraduate students had at least one job offer before graduation. 40 of the graduate students had at least one job offer before graduation.

Another department, X, graduated 20 undergraduate students of which 10 landed jobs before graduation, and it graduated 60 graduate students of which 45 had job offers before graduation.

Which department is more successful in training students ready for the job market?

(2) In May 2018 the CSE department graduated 95 undergraduate students and 50 graduate students. 50 of the undergraduate students had at least one job offer before graduation. 40 of the graduate students had at least one job offer before graduation.

Another department, X, graduated 20 undergraduate students of which 10 landed jobs before graduation, and it graduated 60 graduate students of which 45 had job offers before graduation.

Which department is more successful in training students ready for the job market?

CSE department:

	undergrad.	grad.	Total
Job	50	40.	90
No job	45	10	55
Percentage	52.63%	80%	62.07%

Department X:

	undergrad.	grad.	Total
Job	10	45	55
No job	10	15	25
Percentage	50%	75%	68%

(3) 1314 women took part in a study of heart disease and smoking that was conducted in 1972-1974 in Newcastle, United Kingdom. A follow-up study of the same subjects was conducted thirty years later. Of the 582 women who were smokers in the original study 76.2% were still alive in the follow-up study. Of the 732 non-smokers 68.6% were still alive thirty years later.

Does this show a beneficial effect of smoking?

(3) 1314 women took part in a study of heart disease and smoking that was conducted in 1972-1974 in Newcastle, United Kingdom. A follow-up study of the same subjects was conducted thirty years later. Of the 582 women who were smokers in the original study 76.2% were still alive in the follow-up study. Of the 732 non-smokers 68.6% were still alive thirty years later.

Does this show a beneficial effect of smoking?

Smokers:

	????	????	Total
alive			443
deceased			139
Survival rate			76.2%

Non-smokers:

	????	????	Total
alive			502
deceased			230
Survival rate			68.6%

(3) 1314 women took part in a study of heart disease and smoking that was conducted in 1972-1974 in Newcastle, United Kingdom. A follow-up study of the same subjects was conducted thirty years later. Of the 582 women who were smokers in the original study 76.2% were still alive in the follow-up study. Of the 732 non-smokers 68.6% were still alive thirty years later.

Does this show a beneficial effect of smoking?

Smokers:

	older than 40	others	Total
alive	83	350	443
deceased	80	59	139
Survival rate	50.92%	85.57%	76.2%

Non-smokers:

	older than 40	others	Total
alive	350	152	502
deceased	215	15	230
Survival rate	61.95%	92.02%	68.6%

These counter-intuitive observations are called Simpson's Paradox.

(4) Correlation

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Observation: P(Coffee | Tea) = 0.75

Interpretation: If a person enjoys Tea → they likely enjoys Coffee too

Implication: recommendation, marketing, shelf arrangement, ….

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

$P(\text{Coffee} \mid \text{Tea}) = 0.75$

but $P(\text{Coffee}) = 0.9$, and $P(\text{Coffee} \mid \overline{\text{Tea}}) = 0.9375$

The chance of someone enjoying coffee is even greater among those who don't like tea. There is a negative correlation between the two variables.