Syllabus

M5364-010–Data Mining 1 Fall 2019 Tarleton State Univ Dr. Scott Cook

Sec (010):	Math Bldg 304, TR 9:25-10:40AM	**My Office**:	Math Bldg 132 (254)968-1958
Instructor:	Dr. Scott Cook	**Office Hours**:	MW 1:00-3:15PM or by appointment
My Email:	scook@tarleton.edu	**Math Clinic**:	Math Bldg 203
Math Dept Office:	Math Bldg 142 (254)968-9168	**Math Clinic Hours**:	MTWR 8:00-5:00, F 8:00-2:00

Materials:

- Required: Anaconda distribution of Python, GitHub account, Google Colab & Drive
- Required: *A Whirlwind Tour of Python*, Jake Vanderplas, github.com/jakevdp/WhirlwindTourOfPython
- Required: *Python Data Science Handbook*, Jake Vanderplas, github.com/jakevdp/PythonDataScienceHandbook AND http://shop.oreilly.com/product/0636920034919.do
- Suggested: *The Elements of Statistical Learning*, Hastie, Tibshirani, Friedman, web.stanford.edu/ hastie/ElemStatLearn
- Suggested: Many, many internet resources, blogs/newsletters, package documentation, tutorials (written & video), etc. This field is huge and changes blindingly fast - most specific things we cover in this class will be out of date soon (if not already). The most important skill is to learn to teach yourself using internet-based resources.

Course Content: This field is enormous. There are dozens of core concepts/algorithms, hundreds of niches that address the specific needs of a vast and diverse set of users/organizations, dozens of platforms/libraries/packages to run the computations, etc. We can not possibly cover everything. (And even if we did, everything would be out of date before the class ends in May anyway).

Our goals for the first semester are

1) Establish the core foundational elements of modern data science - we'll focus primarily on classification tasks using relatively clean, well-structured, and moderately-sized data sets.
2) Start to explore the vast array of other tools used by modern data scientists for data cleaning and engineering, handling big data, regression tasks, many specific niches tasks, etc. We'll continue this in the second semester.

Core topics for the first semester include, but not limited to:

- Packages: Python, Numpy, Pandas, Scikit Learn, Tensor Flow, regex, ...
- Visualization & Presentation: Matplotlib, Seaborn, Jupyter (Notebook & Lab), interactive widgets, ...
- Platforms: Anaconda distribution, Google Colab, GitHub, Google Compute Platform, ...
- Supervised Learning Algorithms: Decision Trees, Naive Bayes, k-nearest neighbors, support vector machines, artificial neural networks & deep learning (feed forward, recurrent, convolutional), random forests/boosting/bagging/other ensemble methods, linear and logistic regression
- Unsupervised Learning Algorithms: dimensionality reduction (principle components analysis, multidimensional scaling, locally linear embedding, UMAP, etc), clustering (k-means, gaussian mixtures, etc)
- Concepts: overfitting & underfitting, curse of dimensionality, cross-validation, eager vs lazy, weak vs strong, feature engineering
- Model tuning & evaluation: confusion matrices, accuracy/precision/recall/F_1, ROC curves, cost sensitive learning, hyper-parameter optimization

Homework: Homework will be assigned regularly and will take many different forms (mathematical exercises, coding projects, critical reading of expository articles and tutorials, data exploration, applications and optimization of ML techniques, in-class presentations, vocab quizzes, designing approaches to data science problems, other tasks that build/reinforce data science skills).

Assignments will typically be distributed via Github and submitted via Google Drive (typically as Jupyter notebooks which I will run and grade in Google Colab).

Collaboration: Collaboration is essential, especially for the coding work. The mathematical ideas are complex; writing them into working code is even more complex. You will have bugs and errors of all sorts which will be difficult to fix without help. Collaborate freely. But write up your own homework independently. You must cite your collaborators to give them credit where it is due. You will NOT lose points for crediting a classmate, but you might lose points if you don't give credit where you should have.

I strongly urge you to do as much work in the lab and grad offices as possible. This will allow you to get help from classmate. In this course, you will get stuck frequently. Sometimes, this is good because you're learning a data science lesson. Other times, you just have a silly bug in your code which is wasting time. Having someone else to talk to really helps minimize the wated time and frustration.

Project: You will do a capstone project using techniques from the course. I will post detailed instructions shortly. I may ask for periodic status reports in class so your classmates and I can help refine your idea, avoid pitfalls, and solve technical challenges.

Exams: There will be a comprehensive final exam Friday, December 6, 8:00-10:00. This is intended to provide a reason to consolidate and clarify the core concepts from the course as preparation for future job interviews and your comprehensive oral exams.

Contributions: This field is huge, technically complex, and dynamic. Our class needs to function as a team, trying to identify and present the best, most up to date resources available. As graduate students, you need to do more than just complete the tasks I assign; you need to contribute to the class in ways that have not been specifically asked for. There are many ways to do this, here are a few examples:

- In class, we struggle to understand some detail of an ML algorithm. You find a great blog post that explains it clearly. You post it on the GitHub repo and tell us a about it in class.
- You make really great suggestions for a classmate's project during their status report.
- You figure out a slick way to handle some IT difficulty.

Grading Policy: The guaranteed grade cutoffs are listed below. At my sole discretion, I may curve the course by relaxing them at the end of the semester.

Homework	50%
Capstone Project	40%
Final exam	10%

A	$[90\%, 100\%]$
B	$[80\%, 90\%)$
C	$[70\%, 80\%)$
D	$[60\%, 70\%)$
F	$[0\%, 60\%)$

Notes

- In the event that the university is closed for a scheduled class time, you should assume that whatever was scheduled/due on that day will be scheduled/due on the next class meeting.
- You are expected to present a valid TSU ID upon request.
- Aside from university and departmental policy, all aspects of course policy are at the discretion of the instructor and subject to change.

University Master Syllabus

http://catalog.tarleton.edu/syllabus

Catalog Description: This course centers on the identification, exploration, and description of new patterns contained within data sets using appropriate software. Selected topics will be chosen from data exploration, classification, cluster analysis, and model evaluation and comparison.

Student Learning Objectives: Upon completing this course, a student should be able to do the following:

a) Examine raw data in order to detect data quality issues and interesting subsets or features contained within the data.

b) Transform raw data into a form appropriate for modeling.

c) Select and train appropriate models using the transformed data.

d) Measure the effectiveness of each model.

e) Draw appropriate conclusions.

University Policy: Students are responsible for knowing and abiding by the policies and information contained in the Tarleton Student Handbook [TSUSH].

Student Responsibilities: The student is solely responsible for:

- Attending class.
- Completing every assignment by the specified due date.
- Utilizing, as needed, all available study-aid options (including meeting with the instructor, attending Supplemental Instruction (SI) sessions, going to the Math Clinic, using tutorial software, purchasing a student solutions manual, hiring a personal tutor, etc.) to resolve any questions that they might have regarding homework, course material, and/or technology projects.
- Reading all relevant material in the course text and lecture.
- Being present and prepared for each exam on the specified date and time, unless the instructor determines that a makeup exam is warranted (see Makeup Policy above).
- Obtaining assignments and other materials for classes from which they are absent.
- Giving as much effort as it takes to pass this course.

Student Success Statement - ADA: It is the policy of Tarleton State University to comply with the Americans with Disabilities Act and other applicable laws. If you are a student with a disability seeking accommodations for this course, please contact the Center for Access and Academic Testing, at 254.968.9400 or caat@tarleton.edu. The office is located in Math 201. More information can be found at www.tarleton.edu/caat or in the University Catalog.

Cell phones: Students are expected to set their cell phone so as to emit no audible noise in the classroom. Except for emergency situations, cell phone use (including texting) during the class period is prohibited. A student who is noticeably (to the instructor) distracted by his/her cell phone and/or distracting others with it may be asked to immediately disable it or to leave the classroom. To compensate for your electronic deprivation, keep your calculator on.

Absence Policy: Class absence policies will be established and enforced by each individual course instructor. The course instructor may recommend to the Dean of Students that a student be dropped from a course if excessive absences prevent satisfactory progress.

Makeup Policy: Each course instructor has the responsibility and authority to determine if work can be made-up because of absences. Students may request make-up considerations for valid and verifiable reasons such as the following:

- Illness
- Death in the immediate family
- Legal proceedings
- Participation in sponsored University activities (It is the responsibility of students who participate in University-sponsored activities to obtain a written explanation for their absence from the faculty/staff member who is responsible for the activity.)

Failing grades Tarleton differentiates between a failed grade in a class because a student never attended (F0 grade), stopped attending at some point in the semester (FX grade), or because the student did not pass the course (F) but attended the entire semester. These grades will be noted on the official transcript. Stopping or never attending class is considered an unofficial withdrawal and can result in the student having to return aid monies received. For more information see the Tarleton Financial Aid website.

Student Safety and Title IX: You are in college to achieve academic success, but you must feel safe and take care of yourself to reach your full potential. You have the right to pursue your education in a safe environment. Title IX makes it clear that violence and harassment based on sex and gender are civil rights offenses subject to accountability. *If you or someone you know has been harassed or assaulted, there is help and support on campus.* You may seek assistance confidentially through the Student Counseling Center or the Student Health Center. You may also make a report to the campus Title IX coordinator, which may trigger a university investigation (not a criminal investigation). Additionally, you may pursue criminal charges through the university police department. If the assault occurred away from campus, UPD can assist you in connecting with the appropriate law enforcement agency.

Student Counseling Center: 254-968-9044 (phone is answered 24 hours a day, 7 days a week), TSC 212

Student Health Services: 254-968-9271, TSC 212

Title IX Coordinator: 254-968-9754, Admin Annex 1, Room 112

University Police Department: 254-968-9002, located on the back side of Wisdom Gym

University Core Values

Academic Integrity Statement: Tarleton State University's core values are integrity, leadership, tradition, civility, excellence, and service. Central to these values is integrity, which is maintaining a high standard of personal and scholarly conduct. Academic integrity represents the choice to uphold ethical responsibility for one's learning within the academic community, regardless of audience or situation.

Academic Civility Statement: Students are expected to interact with professors and peers in a respectful manner that enhances the learning environment. Professors may require a student who deviates from this expectation to leave the face-to-face (or virtual) classroom learning environment for that particular class session (and potentially subsequent class sessions) for a specific amount of time. In addition, the professor might consider the university disciplinary process (for Academic Affairs/Student Life) for egregious or continued disruptive behavior.

Academic Excellence Statement: Tarleton holds high expectations for students to assume responsibility for their own individual learning. Students are also expected to achieve academic excellence by:

- honoring Tarleton's core values.
- upholding high standards of habit and behavior.
- maintaining excellence through class attendance and punctuality.
- preparing for active participation in all learning experiences.
- putting forth their best individual effort.
- continually improving as independent learners.
- engaging in extracurricular opportunities that encourage personal and academic growth.
- reflecting critically upon feedback and applying these lessons to meet future challenges.

Academic Honesty: Tarleton State University expects its students to maintain high standards of personal and scholarly conduct. Students guilty of academic dishonesty are subject to disciplinary action. Academic dishonesty includes, but is not limited to, cheating on an examination or other academic work, plagiarism, collusion, and the abuse of resource materials. The faculty member is responsible for initiating action for each case of academic dishonesty that occurs in his or her class.

Academic dishonesty includes, but is not limited to, cheating on an examination or other academic work, plagiarism, collusion, unauthorized use of technology and the abuse of resource materials.

1) Academic work means the preparation of an essay, thesis, problem, assignment or other projects submitted or completed for course credit and to meet other requirements for noncourse credit.

2) What constitutes an act of academic dishonesty may, in part, depend on the particular course and expectations of academic integrity in the context of the course objectives. This includes, but is not limited to, the following:

   2.1) Copying, without instructor authorization, from another student's test paper, laboratory report, other report, computer files, data listing and/or programs.

   2.2) Using, during a test, materials not authorized by the person giving the test.

   2.3) Collaborating with another person without instructor authorization during an examination or in preparing academic work.

2.4) Knowingly and without instructor authorization, using, buying, selling, stealing, transporting, soliciting, copying, or possessing, in whole or in part, the contents of an unadministered test or other required assignment.

2.5) Substituting for another student or permitting another person to substitute for oneself in taking an examination, preparing academic work, or attending class.

2.6) Bribing another person to obtain an unadministered test or information about an unadministered test.

2.7) Using technological equipment such as calculators, computers or other electronic aids in taking of tests or preparing academic work in ways not authorized by the instructor or the university.

3) Plagiarism means the appropriation of another's work and the unacknowledged incorporation of that work in one's own written work in any academic setting.

4) Collusion means the unauthorized collaboration with another person in preparing written work in any academic setting.

5) Abuse of resource materials means the mutilation, destruction, concealment, theft or alteration of materials provided.

This syllabus subject to change as deemed appropriate by the instructor.