IFT 511: Analyzing Big Data

Course Syllabus

Instructor: Asmaa Elbadrawy
Email: asmaa.elbadrawy@asu.edu
Phone: 480 727 0550

Office Hours: Virtual over Zoom, by Appointment Only, Time TBA on Canvas Course Page

Course Credits: 3

Course Description

This course covers how data science can be used as tools to analyze large amounts of data for the purpose of extracting business value. Multiple topics are covered with real business examples.

Theoretical Foundation: Foundations of data analytics, converting different problems to data mining tasks & the decision analytic mindset.

Data Mining-Based Data Analytics: Main data mining techniques for analyzing large amounts of data such as predictive & descriptive methods. Understanding & preparing the data for analysis. Proper training of a model & avoid overfitting. Proper evaluation of the analysis results considering issues such as class imbalances, long tail distributions. Systematic implementation and deployment of the data mining process.

Statistical Methods for Data Analytics: Data distribution, standardization, p-value hypothesis testing. Statistical models used for understanding how the data was generated & their applications to business problems.

Prerequisites

- Students are expected to have decent knowledge of a programming language that is used for data analytics; specifically, either Python or Scala.

Textbooks

- **Required:** ***Data Science for Business*** by F. Provost and T. Fawcett, ISBN-13 : 978-1449361327, O'Reilly Media; 1st edition (August 27, 2013)
- **Recommended but not Required:** ***Mathematical Statistics with Applications***, 7th edition, by Wackerly, Mendenhall, and Scheaffer, Brooks/Cole, Cengage Learning, 2008.

Software Tools

- *Anaconda Python 3.+:* It contains multiple libraries that will be used in the course such as pandas, sklearn, numpy & scipy.
- *Databricks community edition*: A free cloud-based cluster environment with an installation of Spark & access to its machine learning library
- *SVM-Light:* A Support Vector Machine Library used for Classification and Regression

Tentative Schedule

Module	Topic	Objectives	Assessments
Module 1 Week 1	Programming Foundation: Python Basics	1.1. Lists 1.2. Dictionaries 1.3. File Handling 1.4. Data Frames in Pandas & Spark 1.5. Difference between using data frames vs simple file handling	Lab 1 Lab 2
Module 2 Week 2-3	Theoretical Foundation: Principles of Data Analytics	2.1. Why Big Data Analysis? • Challenges, Goals & Examples 2.2. The Data Mining Process 2.3. The Data Analyst Mindset 2.4. Converting Business Problems to Data Mining Tasks 2.5. Other Data Analytics Techniques	Assignment 1 Assignment 2 Assignment 3 Assignment 4 Lab 3
Module 3 Week 4	The Data Mining Process: Data Understanding & Preparation	3.1. Types of Datasets 3.2. Types of Attributes 3.3. Missing Values 3.4. Noise 3.5. Conversion Rules for attributes 3.6. Different input formats for different machine learning libraries - Most common formats: CSV, libSVM, JSON	 Assignment 5 Assignment 6 Lab 4 Lab 5
Module 4 Weeks 5-7	The Data Mining Process: Classification Models	4.1. Correlations & Proximities 4.2. Different Classification Models • Decision Trees • Decision Boundary Classifiers & SVM • Using linear boundary models for ranking • Minimizing/Maximizing Objective Functions • ANN 4.3. Interpretability & Prediction Power of each of the models listed above	Assignment 7 Assignment 8 Assignment 9 Assignment 10 Lab 6 Lab 7 Lab 8 Lab 9 Midterm Quiz

Module 5 Week 8	The Data Mining Process: Fitting the Model	5.1. Models ability to generalize 5.2. The problem of overfitting & its reasons 5.3. Proper model training to avoid overfitting • Cross Validation • Parameter Optimization	Assignment 11 Assignment 12 Lab 10
Module 6 Week 9	The Data Mining Process: Evaluating Classification Models	6.1. Problems with simple evaluation metrics such as accuracy 6.2. Confusion Matrix 6.3. Expected Value 6.4. Precision & Recall 6.5. AUC 6.6. Modeling & Evaluation of unbalanced data 6.7. Long-Tail Distributions in data	Assignment 13 Assignment 14 Lab 11
Module 7 Week 10	The Data Mining Process: Regression Models	7.1. Regression Problem 7.2. Linear Regression 7.3. Logistic Regression 7.4. Time Series Regression	Assignment 15 Lab 12
Module 8 Week 11	The Data Mining Process: Clustering Analysis	8.1. Computing Similarities 8.2. Clustering methods for data summarization & exploration 8.3. Analyzing the Clustering Results Generating Cluster Descriptions	Assignment 16 Lab 13
Module 9 Week 12	CRISP: Building the Data Mining Pipeline	9.1. Automation of the data mining task for solving a given business problem	Assignment 17 Lab 14 Lab 15
Module 10 Week 13	Comparing Machine Learning Libraries	10.1. Spark Architecture 10.2. Spark ML vs Sci-Kit Learn	Assignment 18 Extra Credit Lab
Module 11 Weeks 14	Statistical Data Analysis	11.1. Bayes Rule 11.2. Naïve Bayes Classifier	Assignment 19 Lab 16

Book Reading

- Each module will contain some required reading. Make sure to read the required text before attempting the module assignment(s).

Exams

- There is one midterm and one final exam.
- Exams are delivered on Canvas and require Respondus lockdown browser & a webcam.

Assignments & Labs

- Students shall submit at least one lab and one assignment every week.
- Assignments are for evaluating student understanding of the concepts.
- Labs are for evaluating student ability to implement and run scripts for performing a given analytical task.
- All assignments and labs are delivered and submitted on Canvas.
- All submissions should be in a **single** *doc/docx* document. **NO** zip files, or multiple files. Submitting multiple files may result in obtaining a partial grade since graders expect they will be grading one file.
- Assignments must be **neat** and **organized**. If the person grading your homework cannot read and evaluate the answer, points will be deducted, possibly resulting in a zero grade.
- Most labs include code writing. Students MUST include the items below for their code submissions to be accepted:
    1. Code, copied and pasted as text
    2. Screenshots of the code. Screenshots MUST show a username as well as date and time of when the screenshot was taken.

Late Submissions

- Late submissions will be deducted 10% for each extra day. For example, if the assignment was submitted within 1-day after the due date, grade will be deducted by 10%. If it is submitted within the second day after the due date, grade will be deducted 20%, and so on.
- Emergency exceptions: I understand that life happens! If a student needs more time to finish their work due to some emergency, they need to contact the instructor PRIOR to the due date. The instructor is willing to work with the student and provide flexible due dates if the student maintains good communication and the instructor evaluates the provided reasons as valid.

Grading

- All assignments & labs combined are worth *roughly* 80% of the course grade.
- Midterm and Final Exams combined are worth *roughly* 20% of the course grade.
- Final grade is computed by totaling the exam and assignment scores with no weighting.

COURSE GRADING	
Based on Points (**absolute, fixed, no curve**)	
>= 98.0 <= 100.0	**A+**
>= 93.0 < 98.0	**A**
>= 90.0 < 93.0	**A-**
>= 88.0 < 90.0	**B+**

>= 83.0 < 88.0	B
>= 80.0 < 83.0	B-
>= 78.0 < 80.0	C+
>= 70.0 < 78.0	C
>= 60.0 < 70.0	D
< 60.0	E

Course Objectives

1. Understand theoretical foundations of the Principles of Data Analytics
2. Understand how data science is used to address business problems.
3. Understand the data mining process and its main components: Data Understanding & Preparation, Predictive Data Modeling, Fitting the Model, Clustering Analysis and Model Evaluation
4. Understand different predictive methods and algorithms that can be used for classification and regression models. This includes decision trees, SVM models, ANN, Nearest Neighbor methods, linear & logistic regression.
5. Develop analytical thinking skills required for proper model evaluation & apply them to real business problems.
6. Understand the basics of statistical data analysis, and how to apply it to real business problems.
7. Code and run Python scripts for conducting data analytics tasks, properly evaluating their results, and automating the data mining process.

Student Learning Objectives

1. Use data mining to address business problems and to demonstrate ability to convert a given business problem into a set of data mining tasks.
2. Systematically apply the data mining process and its main components to real business problems.
3. Understand the basics of predictive and descriptive data mining methods and how they can be systematically applied to solve business problems.
4. Understand how predictive methods can be used for ranking purposes in certain problems.
5. Understand the problem of overfitting, how to avoid it and how to apply cross validation strategies for proper model training.
6. Understand how different models vary in their interpretability, performance (such as predictive ability) and their tendency to overfit.
7. Use model evaluation tools beyond accuracy such as confusion matrix, expected value and AUC to properly evaluate models.
8. Write python scripts for automating the data mining process for a given problem. This should include data preparation & transformation, proper model training, evaluation & selection.
9. Use statistical data analysis utilizing probabilities, standardization methods and p-value hypothesis testing to analyze real business problems.

10. Demonstrate understanding of advance analytical techniques such as using ensemble methods to improve model performance.
11. Demonstrate understanding of advanced data analytics techniques used for analyzing special types of data such as text & GIS data.

Communicating with the Instructor

- The instructor is reachable by email. However, since there are multiple online & on campus sections of each course, students need to include course & section details in their email subject. Mainly, students must include:
    - Course Title
    - Section type: Online or On campus
    - For on campus students: Include class days (MW vs TT) or the section number.

  If the student does not include this information in the email, the instructor may not be able to identify information necessary to address the student's request, resulting in communication failure. That is, the instructor won't be able to respond to the student's email.

General Course Protocol & Policies

Academic Integrity

Students in this class must adhere to ASU's academic integrity policy, which can be found at https://provost.asu.edu/academic-integrity/policy). Students are responsible for reviewing this policy and understanding each of the areas in which academic dishonesty can occur. In addition, all engineering students are expected to adhere to both the ASU Academic Integrity Honor Code and the Fulton Schools of Engineering Honor Code. All academic integrity violations will be reported to the Fulton Schools of Engineering Academic Integrity Office (AIO). The AIO maintains record of all violations and has access to academic integrity violations committed in all other ASU college/schools.

No Generative AI Use Permitted

In this course, all assignments must be completed by the student. Artificial Intelligence (AI), including ChatGPT and other related tools used for creating of text, images, computer code, audio, or other media, are not permitted for use in any work in this class. Use of these generative AI tools will be considered a violation of the ASU Academic Integrity Policy, and students may be sanctioned for confirmed, non-allowable use in this course.

Plagiarism

Plagiarism is a violation of academic integrity and is not taken lightly. Plagiarism includes, but is not limited to:

1. Copying from other sources without including proper referencing.
2. Copying from other sources with changing a few words here and there, without including references.

3. Copying whole paragraphs as is from external sources, even if they are referenced. If the student must copy a sentence as is, the sentence must be included between quotes with proper referencing to original source(s).
4. Copying from other and previous student's assignments.

If two assignment papers are identical, both will be marked as plagiarized. If one student claims that other students copied his/her work without consent, the claim will not be acceptable since it cannot be validated. It is the student's responsibility to protect their work.
Any submission with a similarity score above 15% is subject to close examination for possible plagiarism.

Plagiarism Penalty

1. First time a student work is marked as plagiarized: the student will receive a zero in the assignment
2. Second time a student work is marked as plagiarized: the student's final grade will be reduced down by one grade letter. That is, an A will be reduced to a B, a B will be reduced to a C, etc.
3. Third time a student work is marked as plagiarized: the student will receive an E grade in the course and will be reported to the dean of students.

Copyright

All course content and materials, including lectures (Zoom recorded lectures included), are copyrighted materials and students may not share outside the class, upload to online websites not approved by the instructor, sell, or distribute course content or notes taken during the conduct of the course (see ACD 304–06, "Commercial Note Taking Services" and ABOR Policy 5-308 F.14 for more information).
You must refrain from uploading to any course shell, discussion board, or website used by the course instructor or other course forum, material that is not the student's original work, unless the students first comply with all applicable copyright laws; faculty members reserve the right to delete materials on the grounds of suspected copyright infringement.

Policy against threatening behavior, per the Student Services Manual, SSM 104–02

Students, faculty, staff, and other individuals do not have an unqualified right of access to university grounds, property, or services. Interfering with the peaceful conduct of university-related business or activities or remaining on campus grounds after a request to leave may be considered a crime. All incidents and allegations of violent or threatening conduct by an ASU student (whether on- or off-campus) must be reported to the ASU Police Department (ASU PD) and the Office of the Dean of Students.

Disability Accommodations

Suitable accommodations will be made for students having disabilities. Students needing accommodations must register with the ASU Disabilities Resource Center and provide documentation of that registration to the instructor. Students should communicate the need for

an accommodation in sufficient time for it to be properly arranged. See ACD 304-08 Classroom and Testing Accommodations for Students with Disabilities.

Harassment and Sexual Discrimination

Arizona State University is committed to providing an environment free of discrimination, harassment, or retaliation for the entire university community, including all students, faculty members, staff employees, and guests. ASU expressly prohibits discrimination, harassment, and retaliation by employees, students, contractors, or agents of the university based on any protected status: race, color, religion, sex, national origin, age, disability, veteran status, sexual orientation, gender identity, and genetic information.

Title IX is a federal law that provides that no person be excluded on the basis of sex from participation in, be denied benefits of, or be subjected to discrimination under any education program or activity. Both Title IX and university policy make clear that sexual violence and harassment based on sex is prohibited. An individual who believes they have been subjected to sexual violence or harassed on the basis of sex can seek support, including counseling and academic support, from the university. If you or someone you know has been harassed on the basis of sex or sexually assaulted, you can find information and resources at https://sexualviolenceprevention.asu.edu/faqs.

Mandated sexual harassment reporter: As a mandated reporter, I am obligated to report any information I become aware of regarding alleged acts of sexual discrimination, including sexual violence and dating violence. ASU Counseling Services, https://eoss.asu.edu/counseling, is available if you wish discuss any concerns confidentially and privately.

Syllabus changes
Any information in this syllabus (other than grading and absence policies) may be subject to change with reasonable advance notice.