CSC 558 – Data Mining and Predictive Analytics II, Spring 2023 First Day Handout

https://faculty.kutztown.edu/parson/spring2023/CSC558Spring2023.html

Wednesday 6-8:50 PM in Old Main 158 or remote attendance via Zoom at class time.
http://faculty.kutztown.edu/parson Dr. Dale E. Parson, Old Main 260, parson@kutztown.edu. This course is multi-modal; you can attend in person in Old Main 158 or via Zoom at class time.
Class-time Zoom for CSC558: See D2L Course CSC558 -> Content -> Overview for the link.
IF you don't want to be recorded or are a minor, use PRIVATE ZOOM CHAT to me for questions.
Dr. Dale E. Parson, parson@kutztown.edu, Office hours: https://kutztown.zoom.us/j/94322223872
Office Hours Monday 3-5 PM, Tuesday 3-4 PM, Friday (Zoom only) 3-5 PM, or by appt.
Monday & Tuesday office hours are either Zoom using the above link or at Old Main 260.

This course covers advanced study and practice in data mining and predictive analytics. Topics include understanding, configuring, and applying advanced variants of data association, classification, clustering, and statistical analysis engines, analyzing and applying underlying machine learning algorithms, exploring instance-based, support vector, time-series, ensemble, graphical, and lazy learning algorithms, meta-learning, neural nets, genetic algorithms, and validating results. The course examines topics specific to very large data sets. Data cleaning and formatting require some programming in a modern scripting language. Other course activities include using, extending, and customizing off-the-shelf machine learning software systems to accomplish the tasks of data analysis.

Prerequisite: CSC458 or graduate student status.

Textbook: *Data Mining: Practical Machine Learning Tools and Techniques*, **Fourth Edition**, Witten, et. al., ISBN 978-0128042915. You can buy a discounted copy of the **Third Edition** at the KU Book Store. Either edition is fine.

There are on-line copies of the Third Edition available in Rohrbach Library.

Grading (A = 92:100, A- = 90:91, B+ = 87:89, B = 82:86, C+ = 77:79, C = 70:76, F = 0:59)

http://app.kutztown.edu/policyregister/policy.aspx?policy=ACA-048

Projects 100% divided equally among the project assignments.

Final two projects are student presentations and materials for an individual student project.

Project assignment grading criteria Grading criteria will accompany each assignment handout. Please re-check requirements when you feel ready to turn in an assignment.

The academic integrity policy is at http://cs.kutztown.edu/pdfs/AcademicIntegrityPolicy.pdf

Your first reading assignment is to read the above policy statement.

You may openly discuss ideas, algorithms, pitfalls, and the use of programming tools.

You may not share code, test drivers or test data except within groups for group projects.

Group projects, when assigned, have documented partitioning of student responsibilities.

There will be a 10% per day late penalty for projects that come in after the due date.

Class attendance is not graded, but I will be teaching using data sources and concepts both inside and outside the scope of the textbook. You are responsible for all material covered in class, including technical information, coding standards and conventions, verbal specification of assignments, and your questions about topics that are not clear to you. Please, there should be no classroom conversations, cell phones, text messaging, eating, sleeping, obscenities, listening to music or other disruptions of the class.

If you have already disclosed a disability to the Disability Services Office (215 Stratton Administration Building) and are seeking accommodations, please feel free to speak with me privately so that I may assist you. If you have an injury sustained during military service including PTSD or TBI, you are also eligible for accommodations under the ADA and should contact the Disability Services Office.

If you have preferred pronouns for yourself, or a name that differs from the MyKU roster, please let me know.

Any course work submitted to the instructor (including but not limited to assignments, tests, and projects) may be photocopied and retained for the purpose of assessment, accreditation and quality improvement, after removal of any information identifying the student.

Week	Text chapters	Lecture Topics[1]
1		Intro to the course, Zoom. Weka tool set. Review of csc458[2].
2	Weka[3] 4.7, 6.5	Instance-based, lazy machine learning approaches. (Kotu[4] 4.3; K* paper)
3		Data sonification & machine listening analysis. Assn1 out.
4		Batch scripting with command-line Weka, the Unix shell, and Python.
5	Weka 8	Ensemble machine learning; bagging; boosting. Assn1 due. (Kotu 4.7) Assn2 out.
6		Work session.
7		Supervised filters. Batch script preprocessing. Validating results. Cost. Assn2 due.
8	Kotu Ch. 10	Analyzing time series data relationships. (Weka 7.3, time series filters) Assn3 out.
9		Analyzing time series data relationships. Work session.
10	Weka 6.1,6.4	Neural nets and support vector machines. (Kotu 4.5, 4.6) Assn3 due. Assn4 out.
11	Weka 4.8,6.8,6.7	Clustering and advanced Bayesian techniques. (Kotu 7)
12		Work session. Assn4 due.
13		Work session or sessions.
14		Consolidation and review.
15		**Final project (Assn5) presentations.**

1. Assignment 1 on using instance-based (lazy) learning to evaluate machine listening. Hand out week 3, due end of week 5.
2. Assignment 2 on using ensemble learning (meta-learning) to evaluate machine listening. Hand out week 5, due end of week 7.
3. Assignment 3 on framing and analyzing time series data. Hand out week 8, due end of week 10.
4. Assignments 4 & 5 are student selection and preparation of distinct, individual projects. Each student will select a dataset / problem, organize the approach, and clean & format the data for Assignment 4. Each student will do the analysis, write a report, and make a presentation to the class, the latter during the final exam period. This approach worked well last time.

We will be using Zoom for remote attendance during class time. Recorded archives of class sessions will be available.
I will post video archives at the bottom of the course page.

[1] This is a draft schedule that may need revision as we go along.
[2] Review linear regression, model trees, rules, decision trees, and Bayesian techniques.
[3] *Weka* here refers to the Data Mining textbook by Witten, et. al., chapter numbers from 3rd Edition.
[4] Slides from *Predictive Analytics and Data Mining* by Kotu & Deshpande.