Course Information

DSC478 - Programming Machine Learning Applications
Logistics: https://bit.ly/2FdCgEH
Course Management System http://d2l.depaul.edu

Instructor Information

Instructor: Aleksandar Velkoski
Office: CDM Center, Room 522
Office Hours: By appointment only
Phone: (312) 362-1279
Email: avelkosk@cdm.depaul.edu
Homepage: bit.ly/avelkoski

Course Description

The course will focus on the implementations of various data mining and machine learning techniques and their applications in various domains. The primary tools used in the class are the Python programming language and several associated libraries. Additional open source machine learning and data mining tools may also be used as part of the class material and assignments. Students will develop hands on experience developing supervised and unsupervised machine learning algorithms and will learn how to employ these techniques in the context of popular applications such as automatic classification, recommender systems, searching and ranking, text mining, group and community discovery, and social media analytics.

Prerequisites

CSC 401 and DSC 441

Required Textbook

Machine Learning in Action, by Peter Harrington, Manning Publications, 2012. Option Textbooks

Other Textbooks

Python Data Science Essentials - Learn the fundamentals of Data Science with Python, by Alberto Boschetti and Luca Massaron, Packt Publishing, 2015

Python for Data Analysis, by Wes McKinney, O'Reilly, 2012.

Data Mining: Practical Machine Learning Tools and Techniques, by Ian Witten and Eibe Frank, 3rd Ed., Morgan Kaufmann, 2011. Tentative List of Topics

Tentative List of Topics

The following issues and topics will be covered throughout the course. Many of these topics will be revisited several times during the course in a variety of contexts.

Data Mining and Knowledge Discovery

- The KDD process and methodology
- Data preparation for knowledge discovery
- Overview of data mining and Machine Learning techniques
- Review of Python and overview of Python tools for Data Analysis

Supervised Techniques

- Classification and Prediction using K-Nearest-Neighbor
- Classifying with Probability Theory; Naive Bayes
- Building Decision Trees
- Forecasting and Regression models
- Evaluating predictive models

Unsupervised Learning

- Clustering using K-Means
- Association Rule discovery
- Sequential Pattern Analysis
- Principal Component Analysis and Dimensionality Reduction

Possible Applications (covered throughout the course)

- Collaborative Recommender Systems Content Based personalization Predictive User Modeling Concept Discovery from Documents, Blogs, Social Annotations Finding groups using social or behavioral data Building predictive models for target marketing
- Customer or user segmentation

Advance Topics (if time permits)

- SVD and Matrix Factorization
- Search and Optimization Techniques
- Markov Models
- Dealing with Big Data and MapReduce

Grading & Course Requirements

The grading in class will be centered around assignments and a final project.

The assignments will involve Python implementations of selected data mining techniques and their applications in various domains. The assignments will typically involve both programming components and problems related to the course material. Some assignments may also involve the use of other open source data mining tools. The final project, on the other hand, will be a more complex programming and implementation assignment that will involve integrating multiple concepts and techniques. Student will be able to choose from among several possible projects ideas or propose their own. More details on the final project are available below. The final grade will be determined (tentatively) based on the following distribution:

Assignments = 65% Final Project = 35%

The general grading scheme will be based on a curve. At the end of the quarter, some adjustments may be made based on overall class performance as well as signs of individual effort. Plusses and minuses will be given at the high/low ends of each grade range.

Assignment Details

All assignments for the class will be available in electronic form from in the Assignments section of the DSC478 homepage. These assignments must be done individually, unless otherwise specified. Late assignments will be penalized 10% per day (with weekends counting as one day). Assignments are due by 11:59 PM on the due date provided.

You must submit the (documented) code for your scripts, any output files, as well as your interactive Python sessions showing the results of your computations. The preferred option would be to submit your IPython Notebook (similar to examples in class). Be sure to record your interactive session, your code segments, as well as your comments/answers to various questions (along with any auxiliary files). Another option is to copy and paste into a Word document and then add your discussion and answers as necessary.

Final Project Details

Final projects can be done individually or in groups of up to three students. Each group or individual must submit a project proposal for approval by the submission deadline in the tentative schedule. The final project for the class may involve one or a combination of the following:

- ***Data Analysis***: The application of the knowledge discovery process to one or more real-world data sets (see Online

Resources for pointers to various data sets). The tasks must include preprocessing and preparation of the data, data explorations (using statistical approaches to provide an overview of data characteristics), data visualization, and the application of two or more machine learning techniques on the data (e.g., classification, estimation, clustering, association rule discovery, etc.). At least one of the machine learning techniques used must involve building and evaluating a predictive model. Unless otherwise approved (as part of the project proposal), the project must involve the use of Python scripts to perform various data analysis or mining tasks (including available modules or libraries such as NumPy, Scipy, Mathplotlib, Pandas, scikit-learn, and others. In addition to Python tools, you may also use other third-party tools (preferably open-source) to assist with tasks such as preprocessing, data storage and management, and visualization. The deliverables for the project must include a detailed data analysis report, including relevant findings an conclusions about the data, as well as documented code used as part of the project.

- ***Application Development***: The development and evaluation of an original application using machine learning and data mining techniques. The goal of this type of project is not to perform a full analysis of a given data set, but rather to perform useful tasks in a given application domain. The application must be tested and evaluated using a specific data set. The application must also involve the use of one or more of the modeling techniques relevant to the course topics. Your application may also include a significant extension of an existing application discussed in class materials or other sources (in this case, the application must be extended to include additional or more sophisticated types of modeling and analysis). The deliverable for the project must include the fully documented code, distribution files, including any third party sources, installation/deployment documents (including examples, screen shots of test runs, etc.), data used for the application, and a project report providing a description of the components of the application and the results of any evaluation. Many different types of applications are possible, but some examples of such applications include (but are not limited to):
    - **Recommender Systems**: applications that learn from user profiles to provide personalized recommendations for items in a given domain such as movies, books, products, documents, stocks, twitter feeds, etc.
    - **Social Computing Applications**: applications that analyze social network data, including social connections, social

annotations (such as tags), microblog feeds (e.g., on Twitter, Facebook, etc.), blog posts or customer review, and other sources in order to aggregate and present users with useful information or predictions.

- **Business Analytics**: applications involving the use of machine learning and statistical analysis in order to derive business intelligence and assist in business decision making, including tasks such as customer segmentation, predicting customer behavior, market analysis, price prediction, inventory management, Web site analytics, etc.
- **Document Filtering and Analysis**: applications involving the use of machine learning and text mining techniques to identify or filter relevant documents, analyze the content of documents to discover interesting patterns (such as identifying topics or eventsin news stories, analyzing features of items based on customer reviews, spam filtering, etc.), recommending news items, tweets, etc., based on predictive models of users, etc.

Final Project Check List - This document, which can be found on D2L in the course information section, includes a description of the evaluation criteria and deliverables for each type of project. Please review this checklist carefully before the final submission.

Software

Python will be taught in class.

Attendance

It is expected that you will attend every class; it is the single most important action you can take in mastering the course objectives. You are responsible for all material covered, assignments delivered or received, and announcements made in class sessions that you miss. For distance learning students, this means viewing the classes in a timely manner, participate in the discussion forum, and being sure to email or call in any questions that you have.

Recordings of each lecture will be available a few hours after the “live” class, and can be found at the course website https://d2l.depaul.edu. Online students are expected to watch the lectures every week and to keep up with the course information posted on the course website.

Email

Email is the primary means of communication between faculty and students enrolled in this course outside of class time. Students should be sure their email listed under "demographic information" at http://campusconnect.depaul.edu is correct.

Changes to Syllabus

This syllabus is subject to change as necessary to better meet the needs of the students. Significant changes are unlikely, and will be thoroughly addressed in class. Minor changes, especially to the weekly agenda, are possible at any time. If a change occurs, it will be thoroughly addressed during class and posted under Announcements in D2L.

Class Cancellation

In the rare event that class must be canceled, a notification will be posted under Announcements in D2L.

Online Course Evaluations

Instructor and course evaluations provide valuable feedback that can improve teaching and learning. The greater the level of participation, the more useful the results. As students, you are in the unique position to view the instructor over time. Your comments about what works and what doesn’t can help faculty build on the elements of the course that are strong and improve those that are weak. Isolated comments from students and instructors’ peers may also be helpful, but evaluation results based on high response rates may be statistically reliable (believable). As you experience this course and material, think about how your learning is impacted. Your honest opinions about your experience in and commitment to the course and your learning may help improve some components of the course for the next group of students. Positive comments also show the department chairs and college deans the commitment of instructors to the university and teaching evaluation results are one component used in annual performance reviews (including salary raises and promotion/tenure). The evaluation of the instructor and course provides you an opportunity to make your voice heard on an important issue — the quality of teaching at DePaul. Don’t miss this opportunity to provide feedback!

Academic Integrity and Plagiarism

This course will be subject to the academic integrity policy passed by faculty. More information can be found at http://academicintegrity.depaul.edu/. The university and school policy

on plagiarism can be summarized as follows: Students in this course should be aware of the strong sanctions that can be imposed against someone guilty of plagiarism. If proven, a charge of plagiarism could result in an automatic F in the course and possible expulsion. The strongest of sanctions will be imposed on anyone who submits as his/her own work any assignment which has been prepared by someone else. If you have any questions or doubts about what plagiarism entails or how to properly acknowledge source materials be sure to consult the instructor.

Withdrawal

Students who withdraw from the course do so by using the Campus Connection system (http://campusconnect.depaul.edu). Withdrawals processed via this system are effective the day on which they are made. Simply ceasing to attend, or notifying the instructor, or nonpayment of tuition, does not constitute an official withdrawal from class and will result in academic as well as financial penalty.

Retroactive Withdrawal

This policy exists to assist students for whom extenuating circumstances prevented them from meeting the withdrawal deadline. During their college career students may be allowed one medical/personal administrative withdrawal and one college office administrative withdrawal, each for one or more courses in a single term. Repeated requests will not be considered. Submitting an appeal for retroactive withdrawal does not guarantee approval.

College office appeals for CDM students must be submitted online via MyCDM. The deadlines for submitting appeals are as follows:

Autumn Quarter: Last day of the last final exam of the subsequent winter quarter

Winter Quarter: Last day of the last final exam of the subsequent spring quarter

Spring Quarter: Last day of the last final exam of the subsequent autumn quarter

Summer Terms: Last day of the last final exam of the subsequent autumn quarter

Excused Absence

In order to petition for an excused absence, students who miss class due to illness or significant personal circumstances should complete the Absence Notification process through the Dean of Students office. The form can be accessed at

http://studentaffairs.depaul.edu/dos/forms.html. Students must submit supporting documentation alongside the form. The professor reserves the sole right whether to offer an excused absence and/or academic accommodations for an excused absence.

Incomplete

An incomplete grade is a special, temporary grade that may be assigned by an instructor when unforeseeable circumstances prevent a student from completing course requirements by the end of the term and when otherwise the student had a record of satisfactory progress in the course. CDM policy requires the student to initiate the request for incomplete grade before the end of the term in which the course is taken. Prior to submitting the incomplete request, the student must discuss the circumstances with the instructor. Students may initiate the incomplete request process in MyCDM.

- All incomplete requests must be approved by the instructor of the course and a CDM Associate Dean. Only exceptions cases will receive such approval.
- If approved, students are required to complete all remaining course requirement independently in consultation with the instructor by the deadline indicated on the incomplete request form.
- By default, an incomplete grade will automatically change to a grade of F after two quarters have elapsed (excluding summer) unless another grade is recorded by the instructor. o An incomplete grade does NOT grant the student permission to attend the same course in a future quarter.

Students with Disabilities

Students who feel they may need an accommodation based on the impact of a disability should contact the instructor privately to discuss their specific needs. All discussions will remain confidential. To ensure that you receive the most appropriate accommodation based on your needs, contact the instructor as early as possible in the quarter (preferably within the first week of class), and make sure that you have contacted the Center for Students with Disabilities (CSD) at:

Student Center, LPC, Suite #370
Phone number: (773) 325-1677
Fax: (773) 325-3720
TTY:(773) 325-7296