UNIVERSITÉ BISHOP'S UNIVERSITY

Computer Science Department

CS 405/505 – Data Mining

Course Outline

Fall 2019

Professor:

Name: Russell Butler, Ph.D.
Email: rbutler@ubishops.ca
Office: Johnson 114A
Extension: 2420
Office Hours: Monday, Wednesday, Friday 9:00 am – 10:30 am
Tuesday and Thursday by appointment (email)

Schedule:

Lecture:	Monday	11:30 am to 12:30 pm	Nicolls 002
	Wednesday	11:30 am to 12:30 pm	Nicolls 002

Objective:

Learn how to use a variety machine learning algorithms to extract useful information from data and make predictions.

Content:

Data mining is the extraction of implicit, previously unknown and potentially useful information from data. Machine learning provides the technical basis for data mining. This course will focus on the theory and practice underlying machine learning algorithms used in data mining. We will cover decision trees, rules, linear models, instance-based learning, clustering, blind source separation, and denoising algorithms. Pre-processing (preparing input) and post processing (cross validation, bootstrap, etc.) will also be covered. Assignments and examples will be given in python, students will gain hands-on experience working with datasets used in practical applications (such as predicting loan default probability from bank records). Students will emerge from the course equipped to hand real-world data mining problems from both a theoretical and practical standpoint and generate actionable insights for companies as a data scientist.

Credits: 3

Organization

3 hours of lecture per week
4-8 hours of personal time per week (assignments)

Specific Objectives:

By the end of the course, the student should be able to:

1) Write clean and concise vectorized numpy (python) code to manipulate large multidimensional arrays.
2) Efficiently load and visualize small to medium sized datasets in python.
3) Differentiate strengths and weaknesses of specific machine learning algorithms and understand in which cases a given algorithm is appropriate.
4) Validate the output of machine learning algorithms and give confidence intervals for future predictions.
5) Apply appropriate pre-processing and data transformation strategies for manipulating the input to machine learning algorithms.
6) Generate an interpretable structural description of data (when appropriate).
7) Understand the algorithms and math behind methods such as k-means, naïve Bayes, ICA, and decision trees.

Organization:

There will be three lectures per week, each lecture lasting 50 minutes. Lectures will be given in Powerpoint to illustrate theory/concepts, and examples will be given using the scientific python development environment 'Spyder'. Assignments will involve either implementing a specific machine learning algorithm in python, or using existing libraries (scikit-learn) to solve a data mining problem.

Evaluation:

Assignments:	45%
Midterm:	25%
Final Project:	30%

There will be three assignments (15% each), a midterm, and a final project. Assignments and final project may be completed individually or in pairs (2 people max). Failure to submit an assignment/project before the deadline will result in a loss of 10% on that assignment/project for each day late (including weekends). Plagiarized and undelivered assignments/projects will be given 0%.

Resources:

Textbook:

Data Mining – Practical Machine Learning Tools and Techniques (4^{th} edition)

Software:

https://www.spyder-ide.org/	(scientific python development environment)
https://scikit-learn.org/stable/	(machine learning library for python)
https://numpy.org/	(python support for large multi-dimensional arrays)
https://www.anaconda.com/	(simplified python package management)
https://matplotlib.org/	(plotting library for python)