CSC 6515

Machine Learning and Big Data

Goals of the course:

- Exposure to the basic components of the Big Data science
- Hands-on experience with some of the Machine Learning tools used for Big Data
- Enabling self-study of advanced concepts in the Big Data context

Course description:

In this course, we will focus on Big Data and the pillars of that emerging discipline: machine learning/data mining, and elements of high-performance computing, data visualization, and data privacy. Significant part of the course will be devoted to selected, efficient methods for building models from data using machine learning techniques. We will start with decision trees as an initial algorithm, and we will focus on decision forests, linear methods and Bayesian methods. We will also present selected Deep Learning methods used for Big Data, esp. Convolutional Neural Networks and Autoencoders. We will address issues specific to text data, such as embeddings for text. We will discuss the fundamentals of data visualization as a new paradigm for data exploration. We will round up the material with discussion of the basic legal and technical issued related to protection of data privacy in the Big Data context.

Relationship to other courses:

There is some overlap with CSCI6505 Machine Learning. However, the focus here is on select methods that can deal with large data sets, hence some classical approaches are omitted.

Marking scheme

Assignments: 60%

Final exam: 40%

Course plan:

1	Sep.5	Big Data, intro to ML	
2	Sep. 12	Linear models	
3	Sep. 19	Bayesian models	
4	Sep. 26	Ensembles and Random Forest	X

5	Oct. 3	Learning theory	
6	Oct 10	Clustering	
7	Oct 17	Neural Networks and Intro to Deep Learning	
8	Oct 24	Convolutional Neural Networks	
9	Oct 31	Project	X
10	Nov 7	Reading week	No class
11	Nov 21	Viz	X
12	Nov 28	Text/embeddings	
		Project; course summary	X

Textbooks and material:

There is no specific textbook for this material. We will refer to online resources (specific papers and tutorials) for different classes.

We will use scikit and/or AWS for assignments.

Expectations about the students taking this class for credit:

- General familiarity with databases
- Analytical and probability skills at the level of 4th yr CS students
- Due to the use of the tool (Scikit-python), familiarity with Python is expected

Instructor:

Dr. Stan Matwin, Professor and CRC
FCS, Dalhousie
stan@cs.dal.ca

TA/instructor's Helper:
Farshid Varno, email fcsci6515@gmail.com