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 data tools used for Big Data
- Enabling self-study of advanced concepts in the Big Data context
- Sensitization and exposure to data privacy issues

Calendar 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 introduce decision trees as an initial algorithm, and we will focus on decision forests, linear methods and Bayesian methods. In high-performance computing, we will discuss the cloud from the perspective of the Map-Reduce paradigm, and how to engineer Machine Learning algorithms for Map-reduce. 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 (eg decsion trees) are omitted.

Use of Amazon cloud.

Marking scheme (TBD)

Assignments: 60%

Final exam: 40%

Course plan:

Sep. 19	Big Data, intro to ML	
Sep. 26	Linear models	
Oct. 3	Bayesian models	
Oct. 10	Txgvng no class	
Oct 17	Learning theory	
Oct 24	Deep learning	
Oct 31	Cloud - AWS	
Nov 7	Cloud - AWS	
Nov 14	Amazon Machine Learning	
Nov 21	Clustering	
Nov 28	Visualization	
Dec 5	Current research; course summary	

Textbooks and material:

We will use parts of the following books:

1. Flach, P. Machine Learning, Cambridge University Press, 2013
2. Jiawei Han, Micheline Kamber, and Jian Pei, Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann, 2011
3. Zaki, M., Meira, W., Data Mining and Analysis (free pdf avail. online)

Course website will also contain a number of articles and links on which class material will be based.

We will use scikit and 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
- We will limit our practical exercises and the project to the use of the tool (R). No particular programming skills are assumed, but familiarity with and practice in at least one programming language may be needed for data preprocessing.

Instructor:

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