Applications of Machine Learning on Big Data

26:198:685:01

Spring 2016

Course outline

Instructor: Periklis A. Papakonstantinou, e-mail: periklis.research@gmail.com, 1WP-1072

Lectures: Monday 1-3.50pm @ 1WP-202; office hours: Monday 11.30am

Emailing: Make sure that your emails contain in their subject “big data” or “machine learning” or “26:198:685:01”

Textbooks and readings: For Machine Learning we will be using “Machine Learning: A Probabilistic Perspective”, by Murphy. We will be also using survey articles about (i) MapReduce (and Hadoop), (ii) subsampling methods, (iii) Machine Learning over Big Data from the recent workshops held together with the ICML and NIPS conferences. We will also study elements of Fourier Analysis over the Boolean Cube (the instructor will provide lecture notes) and for Computing over Big Datastreams we will rely on the excellent lecture notes by Amit Chakrabarti https://ssl.cs.dartmouth.edu/~ac/Teach/CS35-Fall15/.

Course description, objectives & prerequisites

This is a very exciting, modern, and very useful knowledge to acquire. It is common for “Big Data” courses to just introduce students to good-old-school statistics or computer science techniques. The claim is that these old techniques were developed for what today we call Big Data. To some extent this is true. However, there is so much more in analytics, computation, and inference techniques when one faces modern Big Data challenges. In this class, the students have the chance to acquire knowledge and and learn how to use a set of tools that it is very rarely taught.

Modern Machine Learning (ML) techniques have shown certain success these days. The vast majority of classic and popular ML techniques suffer when used on big amounts of data. NL techniques for Big Data is a very new topic – there is no single place where one can find reasonable teaching material. We will discuss material the instructor acquired through his own research work. This class consists of three parts:

1. A crash-course with the ML prerequisites

2. State-of-the-art techniques for computation over Big Data. These include MapReduce, Property testing algorithms, Streaming Computation, and if time permits Finite Metric Embeddings and their applications.
3. Combine (1) and (2): what can be done with ML on Big Data. The last part of the class will come only in the last 2-3 weeks.

This class is a very modern and rigorous introduction to the topic. It puts forth both the basics and the state-of-the-art in the field, and is designed for a diverse audience. Ideally, the students should have introductory courses in (i) probability and statistics, (ii) linear algebra, and (iii) algorithms or data-structures or a programming course that focuses on such topics. If any of these topics is missing the prerequisites should be covered during this class. The design of the course assumes only minimal prerequisites. The only actual requirement is a basic level of mathematical maturity and basic algorithmic thinking. The assumption is that the student will put effort by self-studying and covering prerequisites when necessary.

**Topics**

- Elements of Machine Learning
- Basic Machine Learning applications (not necessarily over big data)
- The MapReduce paradigm
- Testing properties of Very Big Objects and an introduction to Fourier Analysis over the Boolean Cube
- Introduction to computation over Very Big Datastreams
- Machine Learning techniques over Big Data

**Grading**

Emphasis is given in taking and preparing notes during the class (the fastest and most creative way to acquire very advanced material) and to doing a term project on the intersection of Machine Learning and Big Data Computation or Analytics.

Here is the breakdown of the final grade.

- **Scribe notes**: 40% – also occasional, brief (∼5 min each) presentations in each class
- **Term project**: 40% – at the level of a published research paper
- **2 Quizzes**: 20% (10% each)

**Remark**: The class is designed in a way that the student can choose either (i) to acquire industry-oriented, unique toolbox or (ii) to go deeper in doing state-of-the-art research in the topic. Interaction and participation is strongly encouraged.