Data mining

Krzysztof Choromanski
Google Research
New York, NY, USA

November 11, 2016
1 Course Description

The goal of the course is to present some fundamental techniques used in data mining and machine learning nowadays. The course will show the students those techniques in a way that will enable them to quickly start working on many problems on the frontier of modern data mining and machine learning. The course will cover many important real-life applications of the presented tools, as well as propose problems that can be solved by using a mixture of these techniques.

We will focus on key concepts that built modern big data analysis, i.e. dimensionality reduction mechanisms, linear models, neural networks, etc. I will also show some more recent techniques such as sketching algorithms for graph sparsification. We will also discuss applications of machine learning in robotics.

I will also show how do deal with large datasets in practice when it is not clear in advance which algorithm to use and which data mining technique might be particularly useful.

During this class you will learn what to do with data that cannot fit into your computer’s memory, how to effectively use subsampling techniques, what to do with datapoints consisting of millions of dimensions. Finally, we will also talk about very recent advances in machine learning.

2 Prerequisites

The course does not require its attendees to have any prior knowledge in data mining or machine learning. All concepts will be explained during lectures. It is assumed that participants have some basic background in linear algebra, calculus and probability theory.

3 Grading policy

There will be one midterm, one final exam and a set of homeworks, contributing equally to the final grade.

4 Content

- Dimensionality reduction mechanisms and embeddings (2 lectures)
  - Lecture 1: Linear and nonlinear embeddings (Johnson-Lindenstrauss Transform and extensions).
  - Lecture 2: When you need to learn how to reduce: PCA.

- Kernels (5 lectures)
  - Lecture 3: Introduction to kernels, random feature maps and structured random feature maps.
  - Lecture 4: From linear to nonlinear classifiers - SVM & kernel trick -part I
  - Lecture 5: From linear to nonlinear classifiers - SVM & kernel trick -part II
  - Lecture 6: Hilbert embeddings of distributions with applications in robotics - part I
  - Lecture 7: Hilbert embeddings of distributions with applications in robotics - part II

- Lecture 8: Midterm.
• Neural networks (2 lectures)
  – Lecture 9: Introduction to neural networks - part I
  – Lecture 10: Introduction to neural networks - part II
  – Convolutional neural networks

• Sketching algorithms (2 lectures)
  – Lecture 11: Introduction to sketches
  – Lecture 12: Applications - graph sparsification techniques

• Lecture 13: Model Predictive Control method with applications in robotics.

5 Literature:

• James, G., Witten, D. Hastie, T. and Tibshirani, R. An Introduction to Statistical Learning Springer, 2014.

• Torgo, L. Data Mining with R. CRC Press, 2011.


6 Office hours

Office hours will be hosted once a week in 318 Mudd, exact date will be discussed with students. Students having any questions regarding the course may send emails to: kme2178@columbia.edu.
References