Foundations of Data Science (CSCI 4966/6967)  
Syllabus

Course Web Page:  http://www.cs.rpi.edu/~drinep/FoDS/
Instructor:  Prof. Petros Drineas, Lally Hall 317  
Lecture:  Monday and Thursday 10:00-11:50  
Office hours:  Monday 12:00-14:00, Thursday after class, and by appointment

There will be no required textbook. We will cover material from various sources (e.g. conference papers, journal papers, instructor’s notes, etc.), that will be made available at the course web page.

Course Description

Modern scientific, engineering, and business applications are increasingly dependent on data, yet traditional data analysis technologies were not designed for the complexity of the modern world. Data Science has emerged as a new, exciting, and fast-paced discipline that explores novel statistical, algorithmic, and implementation challenges that emerge in processing, storing, and extracting knowledge from Big Data.

In this course, we will cover foundational aspects of data science, building upon popular models of Big Data using matrices (and tensors) and graphs (and hypergraphs). Numerous matrix and graph algorithms have been developed to analyze such data, taking into account the time complexity and inherent structure of data matrices and graphs. We will discuss popular approaches for data analysis tasks such as dimensionality reduction, clustering, classification, as well as a number of exciting special topics such as streaming algorithms, recommendation systems, ranking algorithms, etc. The course will combine theoretical foundations with implementation and evaluation of data analysis algorithms on real data.

Prerequisites

CSCI 4020 or equivalent; basic probability theory; basic linear algebra; experience in MatLab programming or some equivalent programming language (e.g., Python). Programming assignments will assume that the students are able to at least process data in MatLab format (.mat files). This will be the format that the instructor will use in order to make the data files available to the students.

Course Learning Outcomes

At the end of the course, the student

1. is able to apply fundamental algorithmic ideas to process data,
2. is able to leverage algorithmic insights in order to design novel data science algorithms,
3. has developed a solid background on foundations of data science algorithms,
4. is able to determine running times for common algorithms that process large-scale data,
5. is able to understand and use terminology related to data science,
6. is able to critique material from published research in the area of foundations of data science.
Requirements and Grading
Semester requirements will include bi-weekly homework assignments and a final project that will involve open-ended data analysis tasks on real data, a final report, and a brief presentation. The homework assignments will typically have both a theoretical and a programming component and must be typed in LaTeX (absolutely no exceptions, handwritten homeworks will not be graded). In the theoretical component, the students will be asked to design/analyze algorithms to process (big) data, whereas in the programming component the students will be asked to perform data analysis tasks on medium-scale datasets.

As part of the final project, we will provide real data and a number of open-ended data analysis tasks on these data, as high-level objectives. The students will work towards the completion of these goals in the last four weeks of the semester and will report their findings in a report that will be submitted at the last day of classes (minimum seven pages of text, not including plots and bibliography). The students will also present their findings in a short five minute presentation with a two-minute Q and A session per student.

The weights in determining the semester average are:

<table>
<thead>
<tr>
<th>Component</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homeworks:</td>
<td>60%</td>
</tr>
<tr>
<td>Final project (including report and presentation)</td>
<td>40%</td>
</tr>
</tbody>
</table>

Letter grades will be computed from the semester average. Maximum lower bound cutoffs for A, B, C and D grades are 90%, 80%, 70%, and 60%, respectively. These bounds may be moved lower at the instructor’s discretion.

Late Policy
Homework assignments must be submitted in class at the due date. Late assignments will not be accepted without a written excuse from Student Experience office (4th floor of Academy Hall, x8022, se@rpi.edu).

Academic Integrity
Unless the instructor has given explicit permission, copying and/or communicating solutions to homework assignments is cheating. Any student caught cheating in the homework assignments, midterm, or final, will receive an immediate F in the whole course and will be reported to the appropriate academic authorities.

Preliminary Schedule
Here is a description of the material that will be covered in this course. The instructor might change the order and contents of the syllabus depending on the background of the students in the class and other considerations.

Lecture 1: Intro to Foundations of Data Science.

Lectures 2, 3, and 4: Dimensionality reduction via matrix factorizations (Principal Components Analysis (PCA), Non-Negative Matrix Factorization (NMF), the Semi-Discrete Factorization (SDD), the Maximum Margin Matrix Factorization (M MMF), and the CX and CUR decompositions. Applications to data analysis, from eigenfaces and Latent Semantic Indexing (LSI) to microarray and population genetics data.
Lectures 5 and 6: Dimensionality reduction via random projections, ranging from Gaussian-based random projections to the use of random sign matrices, the Fast Hadamard Transform, and the recent sparse random projection of Clarkson and Woodruff.

Lectures 7 and 8: Multi-linear dimensionality reduction (tensors) and non-linear dimensionality reduction (Locally Linear Embeddings (LLE), Semidefinite Embeddings (SDE), IsoMAP, etc.).

Lectures 9 and 10: Clustering fundamentals: the $k$-means formulation, Lloyd’s iteration to approximately solve the $k$-means clustering problem, connections with PCA, $k$-means++ and other extensions.

Lectures 11 and 12: More on clustering: spectral clustering, hierarchical clustering, agglomerative clustering, etc.

Lectures 13, 14, 15, and 16: Classification fundamentals: problem formulation, nearest neighbors approaches, Linear Discriminant Analysis (LDA), regression, and Support Vector Machines (SVMs). Other classifiers: decision trees, random forests, etc.

Lectures 17 and 18: The web graph: power laws on vertex degrees and other properties of the web graph. Basic generative models of the web graph (preferential attachment models), hubs and authorities, and the page-rank algorithm.

Lectures 19 and 20: The streaming model: variants of the model, elementary algorithmic results.

Lectures 21 and 22: Sparse approximations and compressed sensing: basic formulations and algorithms.

Lectures 23 and 24: Recommendation systems and matrix completion.