Announcements

10/24: First day of class. Welcome!

CLASS MEETS:
Time: Tue & Thu 10:30AM - 11:50AM
Place: HBH 2008

PEOPLE:
Instructor: Leman Akoglu, (lakoglu@andrew)
  Office: HBH 2118C, office ph 412-268-30 four three
  Office hours: Tue 1-2 PM; also, by appointment

Teaching Assistant: Abhinav Maurya, (amaurya@andrew)
  Office: HBH 3034
  Office hours: TBD

COURSE DESCRIPTION:
The rate and amount of data being generated in today's world by both humans and machines are unprecedented. Being able to store, manage, and analyze large-scale data has critical impact on business intelligence, scientific discovery, social and environmental challenges.

The goal of this course is to equip students with the understanding, knowledge, and practical skills to develop big data / machine learning solutions with the state-of-the-art tools, particularly those in the Spark environment, with a focus on programming models in MLib, GraphX, and SparkSQL. See the syllabus for more details. Students will also gain hands-on experience with MapReduce and Apache Spark using real-world datasets.

This course is designed to give a graduate-level student a thorough grounding in the technologies and best practices used in big data machine learning. The course assumes that the students have the understanding of basic data analysis and machine learning concepts as well as basic knowledge of programming (preferably in Python or Java). Previous experience with Hadoop, Spark or distributed computing is NOT required.

Learning Objectives
By the end of this class, students will

- gain understanding of the MapReduce paradigm and Hadoop ecosystem
- understand scalability challenges for common ML tasks
- study distributed machine learning algorithms
- understand details of SparkSQL, GraphX, and MLib (Spark's ML library)
- implement distributed pipelines in Apache Spark using real datasets

RECOMMENDED TEXTBOOKS:
There is no official textbook for the course. I will post all the lecture notes and several readings on course website. Below you can find a list of recommended reading.

- Scaling up Machine Learning: Parallel and Distributed Approaches, Cambridge University Press
  Ron Bekkerman, Mikhail Bilenko, John Langford

- Hadoop in Practice, Manning Publications Co.
  Alex Holmes
- **Learning Spark**, O'Reilly
  Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia

- **Advanced Analytics with Spark**, O'Reilly
  Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills

**BULLETIN BOARD and other info**

- We will use the [Blackboard](http://www.andrew.cmu.edu/user/lakoglu/courses/95869/index.htm) for course materials, homework deposits, announcements, and grades.
- We will use [Piazza](http://www.andrew.cmu.edu/user/lakoglu/courses/95869/index.htm) for questions and discussions.

**MISC - FUN:**

- [Joke-1](http://www.andrew.cmu.edu/user/lakoglu/courses/95869/index.htm)
- [Joke-2](http://www.andrew.cmu.edu/user/lakoglu/courses/95869/index.htm)
- [Joke-3](http://www.andrew.cmu.edu/user/lakoglu/courses/95869/index.htm)
# Tentative Syllabus

<table>
<thead>
<tr>
<th>Date</th>
<th>Lectures and Readings</th>
<th>Out / Due</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/24</td>
<td><strong>(Recitation) Lecture 0: Set up</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Installation of Hadoop and Spark on your local machine</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Setting up AWS clusters</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Review</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Please take this <a href="http://www.andrew.cmu.edu/user/lakoglu/courses/95869/syllabus.htm">Python mini-quiz</a> before the course and take this <a href="http://www.andrew.cmu.edu/user/lakoglu/courses/95869/syllabus.htm">Python mini-course</a> if you need to learn Python or refresh your Python knowledge.</td>
<td></td>
</tr>
<tr>
<td>10/24</td>
<td><strong>Lecture 1: Introduction</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Big Data applications</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Technologies for handling big data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Apache Hadoop and Spark overview</td>
<td></td>
</tr>
<tr>
<td>10/24</td>
<td><strong>Lecture 2: Hadoop Fundamentals</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Hadoop architecture</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- HDFS and the MapReduce paradigm</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Hadoop ecosystem: Mahout, Pig, Hive, HBase, Spark</td>
<td></td>
</tr>
<tr>
<td>10/26</td>
<td><strong>HW1 out</strong></td>
<td></td>
</tr>
<tr>
<td>10/26</td>
<td><strong>Lecture 3: Introduction to Apache Spark</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Big data and hardware trends</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- History of Apache Spark</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Spark's Resilient Distributed Datasets (RDDs)</td>
<td></td>
</tr>
<tr>
<td>11/2</td>
<td>- Transformations and actions</td>
<td></td>
</tr>
<tr>
<td>11/7</td>
<td><strong>Lecture 4: Machine Learning Overview</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Basic machine learning concepts</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Steps of typical supervised learning pipelines</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Linear algebra review</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Computational complexity / Big O notation review</td>
<td></td>
</tr>
<tr>
<td>11/9</td>
<td><strong>Lecture 5: Linear Regression and Distributed ML Principles</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Linear regression</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- formulation and closed-form solution</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- gradient descent</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- grid search</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Distributed machine learning principles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- computation, storage, and communication</td>
<td></td>
</tr>
<tr>
<td>11/14</td>
<td><strong>HW1 due</strong></td>
<td></td>
</tr>
<tr>
<td>11/14</td>
<td><strong>HW2 out</strong></td>
<td></td>
</tr>
<tr>
<td>11/16</td>
<td><strong>Lecture 6: Logistic Regression and Click-through Rate Prediction</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Online advertising</td>
<td></td>
</tr>
</tbody>
</table>
• Linear classification
• Logistic regression
  o working with probabilistic predictions
  o categorical data and one-hot-encoding
  o feature hashing for dimensionality reduction

HW2 due  
HW3 out

11/23
No class: Thanksgiving

Lecture 7: Principal Component Analysis and Neuroimaging
11/21
• Exploratory data analysis
• Principal Component Analysis (PCA)
11/28
• Formulations and solution
• Distributed PCA

Lecture 8: Big Data ML with MLlib
11/30
• k-means Clustering
• Decision Trees and Random Forests
• Recommenders

HW3 due  
HW4 out

Lecture 9: Introduction to SparkSQL
12/5
• Working with tables in Spark
• Higher-level declarative programming

Lecture 10: Analyzing Networks with GraphX
12/7
• Understanding network structure
• Computing graph statistics

HW4 due  
Project out

TBD  Final Exam
Assignments

**COURSEWORK:**

Coursework consist of 4 homework assignments, 1 take-home course project on big data analytics, and 1 final exam (grading in parentheses):

- Homework (40%, 10% each)
- Project (20%)
- Final Exam (30%)
- Class Participation (10%)

**IMPORTANT DATES:**

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Note</th>
<th>Out</th>
<th>Due</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recitation</td>
<td>Installation, Set up (Date TBA)</td>
<td>Oct 24</td>
<td>--</td>
<td>0%</td>
</tr>
<tr>
<td>Homework 1</td>
<td>Programming in MapReduce and Spark</td>
<td>Oct 26</td>
<td>Nov 9</td>
<td>10%</td>
</tr>
<tr>
<td>Homework 2</td>
<td>Regression in Spark</td>
<td>Nov 9</td>
<td>Nov 21</td>
<td>10%</td>
</tr>
<tr>
<td>Homework 3</td>
<td>Classification in Spark</td>
<td>Nov 21</td>
<td>Nov 30</td>
<td>10%</td>
</tr>
<tr>
<td>Homework 4</td>
<td>Data Analysis with PCA in Spark</td>
<td>Nov 30</td>
<td>Dec 7</td>
<td>10%</td>
</tr>
<tr>
<td>Project</td>
<td>Open-ended problem (NO LATE DAYS!)</td>
<td>Dec 7</td>
<td>Dec 9</td>
<td>20%</td>
</tr>
<tr>
<td>Final Exam</td>
<td>TBA</td>
<td>TBA</td>
<td>--</td>
<td>30%</td>
</tr>
<tr>
<td>Class participation</td>
<td>Pop-quizzes</td>
<td>--</td>
<td>--</td>
<td>10%</td>
</tr>
</tbody>
</table>

**HOMEWORK:**

The goal of the homework is to enable the students to practice the concepts learned in class using real-world datasets.

- ASSIGNMENTS ARE DUE AT THE BEGINNING OF LECTURE ON THE DUE DATE.
- All assignments are to be done individually. Please see the collaboration policy.
- To submit:
  - Submit your soft-copy in .pdf as well as all code in .zip on Blackboard.
  - Return a printed hard-copy on due date in class.
  - Make sure to your answers are clear and writing is legible.
  - See course policies for assignment questions, late submissions, graded homework pick-up.

**EXAM:**

There will be a final exam. It will be open book, notes, papers, etc., but you are not allowed to use a computer or any other electronics. The tentative dates are posted above, the finalized dates will be announced during the semester.

**PROJECT:**

Your class project is an opportunity for you to explore a machine learning problem in the context of a real-world data set using big data analysis tools.

For the project, we will provide you with a large dataset as well as a list of machine learning problems possible on the provided data. Your task will be to choose one of those ML problems, or define your own, on the provided dataset and address the problem of your choice with the big data analysis tools you learned during the course as well as others you explore based on the APIs.

By design, the project is open-ended; you are free to decide how you want to approach the problem and what
tools you want to employ. We want to see a best-effort solution that utilizes what you learned in class and also potentially trying new things beyond class.

Important things to note:

- You have to use the data we have provided you. You cannot choose your own dataset.
- You will be given 48 hours to work on the project. Use of late days are not allowed for this submission.
- Project is to be done individually. No collaboration is allowed. Students who use each other's ideas or code will be heavily penalized.

Your project will be worth **20%** of your final class grade.

**Project Writeup:**

Course staff will use the following rubric when grading your final project.

- **Introduction/Motivation/Problem Definition (10%)**
  - Identify, define, and motivate the problem that you are addressing.
  - How (precisely) will a machine learning solution address the problem?

- **Data Understanding and Preparation (15%)**
  - What preliminary analyses have you performed on the data? What observations have you made?
  - How did those observations help shape your approach?
  - Provide the preliminary data analysis results and your observations.
  - Specify how the data will be transformed to the format required for machine learning.

- **Methodology (35%)**
  This is where you give a detailed description of your primary contributions. It is especially important that this part be clear and well written so that we can fully understand what you did.
  - Specify the type of model(s) built and/or information/knowledge extracted.
  - Discuss choices for machine learning algorithm: what are other alternatives, and what are their pros and cons (in the context of the problem and as compared to your proposed solution)?
  - Discuss why and how this model should "solve" the problem (i.e., improve along some dimension of interest).
  - Outline the big data analysis tools and libraries you have used.

It is not so important how well your method performs but rather, (a) how thorough and careful your methodology is, and (b) how interesting and clever the approaches your took and the tools you have used are.

- **Evaluation and Results (30%)**
  We are interested in seeing a clear and conclusive set of experiments which successfully evaluate the problem you set out to solve. Make sure to interpret the results and talk about what we can conclude and learn from your approach.
  - How do you evaluate your machine learning solution to the specific question(s) you have addressed?
  - What do these results tell you about your solution?
  - Present and discuss your evaluation results and findings. You may use tables or figures (e.g. ROC plot) to visualize your results.

- **Style and writing (10%)**
  Overall writing, organization, figures and illustrations.

Please follow the instructions in the IPython notebook that will be handed out to you. You will fill in your answers in the IPython notebook, upload a zipped folder containing the IPython notebook and its HTML output on Blackboard, and submit a printed hardcopy of the HTML output to the TA.
Course Policies

LECTURES
- All devices such as laptops, cell phones, noisy PDAs, etc. should be turned off for the duration of the lectures and the recitations, because they may distract other fellow students.
- Please come to all lectures on time and leave on time, again so that there are no distractions to the classmates.

PREREQUISITES
Students are expected to have the following background:
- Basic knowledge of data analysis and machine learning concepts; having taken:
  - 95-791 Data Mining, OR
  - 95-828 Machine Learning for Problem Solving
- Basic programming skills at a level sufficient to write a reasonably non-trivial computer program in a language of preference (preferably Python)

ASSIGNMENTS
- Assignments are due at the *beginning of lecture* on the due date.
- The due date of assignments are posted at the assignments page.
- Assignments will be posted on Blackboard.
- Students should submit the programming part of assignments electronically via Blackboard (no print outs).

Important Note: As we reuse problem set questions, covered by papers and webpages, we expect the students not to copy, refer to, or look at the solutions in preparing their answers. Since this is a graduate-level class, we expect students to want to learn and not google for answers. The purpose of problem sets in this class is to help you think about the material, not just give us the right answers. Therefore, please restrict attention to the books mentioned on the front page when solving problems on the problem set. If you do happen to use other material, it must be acknowledged clearly with a citation on the submitted solution.

Academic integrity
All students are expected to comply with CMU's policy on academic integrity. Please read the policy and make sure you have a complete understanding of it.

Collaboration
You are encouraged to discuss homework problems with your fellow students. However, the work you submit must be your own. You must acknowledge in your submission any help received on your assignments. That is, you must include a comment in your homework submission that clearly states the name of the student, book, or online reference from which you received assistance.

Submissions that fail to properly acknowledge any help from other students or non-class sources will receive NO credit. Copied work will receive NO credit. Any and all violations will be reported to the Heinz College administration and may appear in the student's transcript.

Questions and requests
- You should use Piazza for all your questions about the assignments and the course material. Instructor and TA(s) will do their best to answer your questions timely.
- Regrade requests should be done in writing/email,
  - within 2 days after graded assignments are distributed
  - to the TA that graded the question, and specifying
    - the question under dispute (e.g., 'HW1-Q.2.b')
    - the extra points requested (e.g., '2 points out of 5')
    - and the justification (e.g., 'I forgot to divide by variance, but the rest of my answer was correct')
  - In the remote case there is no satisfactory resolution, please contact the instructor.
Homework pick-up information

- You may pick up graded homeworks etc., from the course admin
  - Mrs. Adrienne McCorkle, HBH 2250
  - 9:00-11:30am, 1:30-4:30pm every weekday
  - with photo-id (for your privacy protection)

Late policy

- No delay penalties, for medical/family/etc. emergencies (bring written documentation, like doctor's note).
- Each student is granted '3 slip days' total for the whole course duration, to accommodate for coinciding deadlines/interviews/etc. That is, no questions asked, if the total delay is 3 days or less.
  - You can use the extension on any assignment during the course. For instance, you can hand in one assignment 3 days late, or 3 different assignments 1 day late each.
  - Late days are rounded up to the nearest integer. For example, a submission that is 4 hours late will count as 1 day late.
  - After you have used up your slip days, any assignment handed in late will be marked off 25% per day of delay.
- To use slip days:
  - upload your homework on Blackboard to mark the time of submission
  - make sure to return hard copy next time in class
  - note on front page of hard copy submission: count of slip days you used, as well as the count of slip days left