Syllabus for MATH 191
MATH 191 Topics in Data Science: Algorithms and Mathematical Foundations
Department of Mathematics, UCLA
Fall Quarter 2015

Lecture: MWF: 1:00-1:50pm, GEOLOGY 4645

	Office:	MS 7310
Instructor: Mihai Cucuringu	Email:	mcucurin@ucla.edu
	Office Hours:	Monday 5-6:30 pm, Wednesday 5-6:30 pm

Course Description: This is a project-based interdisciplinary course covering a number of topics in *Data Science*, that will combine both theoretical and practical approaches. The goal of the course is, on one hand, to understand (at least at a high level) the mathematical foundations behind some of the state-of-the-art algorithms for a wide range of tasks including organization and visualization of data clouds, dimensionality reduction, network analysis, clustering, classification, regression, and ranking. On the other hand, students will be exposed to numerous practical examples drawn from a wide range of topics including social network analysis, finance, statistics, etc. There will be a strong emphasis on research opportunities.

Textbook: Readings and publicly available material/textbooks will be made available on the course website. The class textbook is

- *An Introduction to Statistical Learning* by James, Witten, Hastie, and Tibshirani,

freely available at `http://www-bcf.usc.edu/~gareth/ISL/`. Lecture notes and/or slides will also be posted on the course website. More advanced material, and also a very good reference book, is *The Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman, which contains a lot more material than what we plan to cover throughout the course. Note that this book is also freely available online at `http://www-stat.stanford.edu/ElemStatLearn`.
Another popular reference in the machine learning and data mining community is *Pattern Recognition and Machine Learning Book*, by Christopher Bishop.

Prerequisites: linear algebra (MATH 33A), elementary probability (MATH 170A) and/or statistics, and computer programming. Familiarity with a programming language such as R or MATLAB is desirable, but students are free to choose their own favorite programming language. Please keep in mind that this course will require that you write a significant amount of code (homework, midterms and final). Please contact the instructor if unsure about the prerequisites.

Programming: For the experimental/empirical parts of the course, I will mostly be using R (and perhaps occasionally MATLAB), and share with you any scripts that I put together. For projects and homework problems, I advise that you use R or MATLAB. I can help you with R or MATLAB, but for any other languages you are on your own.

Exams:

- Midterm: take home Nov 13-16

- Final exam: consists of a project for which you will deliver an in-class presentation and write up a research report. The presentations will take place in the same classroom during the last two (possibly three) lectures in the last week of classes (the week of November 30). You may work either individually or in teams of size at most 3. More details on this will follow as we get closer to the second half of the quarter. You are encouraged to come up with your own project, but I will also design a number of projects you can choose from.

Final letter grade:

- Your final grade is calculated from the average homework grade (scaled to the range 0-100), scribe notes for a lecture, midterm and final project, as a weighted average with the following weights:

Homework	25%
Scribe notes	10%
Midterm	15%
Final Project	50%

- The numeric cut-offs for computing the final letter grade will take into account the overall performance of the class.

Appeals:

- As a rule of thumb, you should only appeal on correctness, and not on the amount of partial credit you received.
- Appeals for the midterm and final projects must be submitted to the instructor within one week of the exam grading.

Homework: The following rules apply to homework:

- I will occasionally assign homework throughout the course. It is important that you do the homework if you want to understand the material taught in class, both the theoretical and the experimental problems. I will announce in class when homework is posted online and when it is due.
- **Write-up:** You must always justify your solution to each homework problem. Correct final answers without a correct or incomplete justification will receive zero or very few points.
- **Group work:** It is OK (as a matter of fact, encouraged) to work together on the homework, **in groups of size at most 5**. However, when it comes time for you to write up the solutions, I expect you to do this on your own. It would be best for your own understanding if you put aside your notes from the discussions with your classmates and write up the solutions entirely from scratch.
- For the experimental problems, it goes without saying that you should write your own code. It is OK to ask others for help with programming, but please make sure you understand what you code.
- **Submission format:** Please attach the cover page (found at the end of this syllabus) as the first page to each and every homework assignment you hand in. It is fine if you handwrite you own version of this cover page.

A list of tentative topics:

1. Introduction & syllabus
2. Review of basic statistics and probability
3. Statistical learning, and introduction to R and several data sets
4. Bias-variance decomposition

   Measures of correlation in data:

5. Pearson (sample and population versions), Spearman, Hoeffding's D
6. Maximal correlation, and review of characteristic functions; Distance correlation
7. Information theory (entropy, mutual information), and Maximal Information Coefficient (MIC) (*Detecting Novel Associations in Large Data Sets*, Reshef et al., Science 2011)
8. Simple/multiple linear regression, proof that OLS is BLUE
9. Linear regression - practical considerations
10. Singular Value Decomposition (SVD), rank-k approximation, Principal Component Analysis (PCA)
11. PCA derivation (best d-dimensional affine fit/projection that preserves the most variance)
12. PCA in high dimensions and random matrix theory (Marcenko-Pastur); applications to finance
13. Basics of spectral graph theory

    Nonlinear dimensionality reduction methods:

14. Diffusion Maps
15. Multidimensional scaling and ISOMAP
16. Locally Linear Embedding (LLE)
17. Kernel PCA
18. Ranking with pairwise incomplete noisy measurements, and applications; Page-Rank
19. Overview of several ranking algorithms: Serial-Rank, Rank-Centrality, SVD ranking
20. The Angular Synchronization problem and an application to ranking

    Clustering:

21. Clustering: K-means and K-medoids, Hierarchical clustering
22. Spectral clustering, isoperimetry, conductance

    Modern regression:

23. Ridge regression
24. The LASSO

Scribe Notes

For each lecture, one or two students will be responsible for scribing the lecture notes. Here are some useful tips:

1. Come to class!
2. Take careful notes during class of what I write on the blackboard, of any slides I may use, of what I explain (sometimes without writing down), and of other students' questions and comments.
3. Put together your notes into a latex document (using the template provided), where you make a clear description in complete full sentences of what has been covered in class. Somebody who missed class or was late for class that day should be able to make full sense of all your notes.
4. Use the textbook or associated readings as a reference if you feel something is not clear in your notes, or you want to add in more details.
5. Make sure you check for spelling mistakes, and read over the document a few times!!
6. Use latex to prepare your documents, I attached a latex template that you should use *Template-Scribe-Notes-191*
7. Include figures where appropriate. Sometimes it may be easier to scan/print screen a figure from the textbook, or find a similar one online.
8. Here are some websites that you may find useful if you are a beginner in LaTex:

   `http://www.latex-project.org/`

   `http://www.personal.ceu.hu/tex/cookbook.html`

   `http://www.thestudentroom.co.uk/wiki/LaTex`
9. Email me both the latex file (the **.tex** file), together with any figures (png, jpg or pdf format), and the compiled final PDF, within one week of the class for which you volunteered to scribe.
10. I will post your scribe notes online, possibly after making some small revisions myself. Usually, you will be asked to make some more revisions based on my comments; once you do so, I will post the polished version online.

MATH 191: Topics in Data Science: Algorithms and Mathematical Foundations

Cover Page

Fall 2015

Homework #	
Last name	
First name	
Student ID	

Worked with (list at most 4 full names):

1.	
2.	
3.	
4.	