**USC**Dornsife

**Psych 625: Advanced Big Data Methods**

**Fall—TH—2-6**

**Location:** BCI 266

**Instructor: Morteza Dehghani**

**Office:** SGM 607

**Office Hours:** Hours Wed 10-12, or by appt.

**Contact Info:** mdehghan@usc.edu

|  |
| --- |
|  |

**Course Description**

In the past several years, social scientists have been facing a quantitative change in technology. This change can be summarized in two main points: 1. availability of vast and seemingly insurmountable volumes of human-related data, and 2. constantly increasing computational power. These have provided an unprecedented opportunity to study and model human cognition with range and detail previously not imaginable. Moreover, there is growing interest (e.g. in marketing) to use such data for predicting a variety of human behavior, for detecting different types of activities, or for more intelligent targeted advertising. *Advanced Big Data Methods* focuses on methods in computer science, specifically in machine learning, which can help us achieve these outcomes. This course is followed by *Topics in Computational Social Sciences* which focuses on the applications natural language processing and network analysis, guided by psychological theories, for identifying various social and cognitive properties evident in human related big data.

The intended audience for this course is psychology graduate students, and more broadly graduate students in social sciences, who are interested in using machine learning techniques for analysis of data, Also, this course may be of interest to PhD students in communications, computer science and the business school.

**Learning Objectives**

This course is designed to be hands-on and students are expected to learn how to apply different machine learning techniques for analyzing different types of data. In order to achieve this objective, each discussed topic is accompanied by a lab session in which we examine how to use that technique on a data set. Lab clinic sessions are used for helping students troubleshoot their code and also for going over the homework.

**Prerequisite(s):** Instructor permission

**Recommended Preparation**: Psych 501 or a similar introductory statistics course

**Course Notes**

Lecture notes and homework assignments will be posted on Blackboard. Students are also highly encouraged to use the course forum on Blackboard.

**Technological Proficiency and Hardware/Software Required**

This class includes lab sessions. Students are required to bring a laptop to class. Homework assignments are programming problems that need to be written in *R*.

**Required Readings and Supplementary Materials**

Required:

* James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with Applications in R*. Springer.

Supplementary:

* Venables, W. N., Smith, D. M., & R Development Core Team. (2002). *An introduction to R*. Available at: <http://cran.r-project.org/doc/manuals/R-intro.pdf>
* Darwiche, A. (2009). *Modeling and reasoning with Bayesian networks*. Cambridge University Press.
* Nagarajan, R., Scutari, M., & Lèbre, S. (2013). *Bayesian Networks in R*. Springer.
* Friedman, J., Hastie, T., & Tibshirani, R. (2009). *The elements of statistical learning* (Second Edition). Springer, Berlin: Springer series in statistics.
* Alpaydin, E. (2014). *Introduction to machine learning*. MIT press.

**Description and Assessment of Assignments**

* 1. Homework assignments. Each week students will complete programming problems from the required book. The assignments will be graded based on both output and style of the code. The homework material will be reviewed during lab clinics.
  2. Lab presentation. Each student will do a lab presentation in which a particular lab module is taught to others using a different dataset than the one used in the book. This dataset will be made available to the class by the second week of class. By having enrolled in the class all student acknowledge the copyright information regarding this dataset.
  3. Class Projects. Students will complete four class projects. These projects will be relatively heavy programming assignments requiring students to use R to implement some specific statistical technique. The first project will be relatively easy and not time consuming. The other projects, however, will take substantial time (~40 hours each). The projects will be assigned at the beginning of the semester.

**Grading Breakdown**



**Assignment Submission Policy**

Homework will be assigned on Thursdays and will be due the following Thursday at 11am, before the start of class submitted on Blackboard. All homework turned in any later than 11:10am will be considered late. Students will be allowed a total of seven late days that can be used on the assignments. In exceptional circumstances, arrangements must be made in advance of the due date to obtain an extension. Once you have used up your seven late days, one additional day late will result in a 25% reduction in the total score, two additional days late will yield a 50% reduction, and no credit will be given for three or more additional days late.  Late days are in units of days, not hours, so using up part of a day uses up the whole day. The final project report, plus the R code used, will be due on the day of the final exam. All assignments, including the projects, need to be written using *sweave* or *knitR*. Copied and pasted code/results will not be accepted.

**Course Schedule:**

The following schedule is tentative and may change during the semester.

1. Week 1 August 23: Introduction, Statistical Learning & Linear Regression
   1. What is Statistical Learning (ISLR 2.1)
   2. Assessing Model Accuracy (ISLR 2.2)
   3. Simple & Multiple Linear Regression (ISLR 3.1, 3.2 & 3.3)
   4. Intro to *kintR* package
   5. HW 1 assigned
2. Week 2 August 30: Classification, Lab 1 & 2, Lab Clinic 1
   1. Overview of Classification (ISLR 4.1, 4.2)
   2. Logistic Regression (ISLR 4.3)
   3. Linear Discriminant Analysis (ISLR 4.4)
   4. Comparison of Classification Methods (ISLR 4.5)
   5. Bayesian Classifiers
   6. Lab Clinic 1
   7. Lab: Linear Regression (ISLR 3.6)
   8. HW 1 due, HW 2 assigned
3. Week 3 September 6: Resampling Methods, Linear Model Selection, Lab 2, Lab Clinic 2
   1. Cross Validation (ISLR 5.1)
   2. The Boot Strap (ISLR 5.2)
   3. Subset Selection (ISLR 6.1)
   4. Shrinkage Methods (ISLR 6.2)
   5. Dimension Reduction Methods (ISLR 6.3)
   6. Considerations in High Dimensions (ISLR 6.4)
   7. Lab Clinic 2
   8. Lab: Logistic Regression, LDA, QDA and KNN (ISLR 4.6)
   9. HW 2 due, HW 3 & 4 assigned
   10. Project 1 due
4. Week 4 September 13: USC Computational Social Science Confrence
5. Week 5 September 20: Moving Beyond Linearity, Lab 3 & 4
   1. Polynomial Regression (ISLR 7.1)
   2. Step Functions (ISLR 7.2)
   3. Basis Functions (ISLR 7.3)
   4. Regression Splines (ISLR 7.4)
   5. Smoothing Splines (ISLR 7.5)
   6. Local Regression (ISLR 7.6)
   7. Generalized Additive Models (ISLR 7.7)
   8. Lab: Resampling Methods (ISLR 5.3)
   9. Lab: Regularization (ISLR 6.5, 6.6 & 6.7)
   10. HW 4 due, HW 5 assigned
6. Week 6 September 27: Tree-Based Methods, Support Vector Machines, Lab 5, Lab Clinic 3
   1. Decision Trees (ISLR 8.1)
   2. Bagging, Random Forests, Boosting (ISLR 8.2)
   3. Maximal Margin Classifier (ISLR 9.1)
   4. Support Vector Classifiers (ISLR 9.2)
   5. Support Vector Machines (ISLR 9.3 & 9.4)
   6. Support Vector Regression (Handouts)
   7. Lab: Moving Beyond Linearity (ISLR 7.8)
   8. Lab Clinic 3
   9. HW 5 due, HW 6, 7 assigned
7. Week 7 October 4: No Class
8. Week 8 October 11: Neural Networks, Lab 6, 7, Lab Clinic 4
   1. The Perceptron (ITML 11.2-11.4)
   2. Multilayer Perceptron (ITML 11.5-11.6)
   3. Backpropagation Algorithm (ITML 11.7)
   4. Lab: Decision Trees (ISLR 8.3)
   5. Lab: Support Vector Machines (ISLR 9.6)
   6. Lab Clinic 4
   7. HW 6,7 due
   8. Project 2 due
9. Week 9 October 18: Neural Networks II
   1. Recursive Neural Networks (Handouts)
   2. Long Short-Term Memory (Handouts)
   3. Attention Networks (Handouts)
   4. Auto-Encoders (Handouts)
   5. Darwiche, A. (2017). Human-Level Intelligence or Animal-Like Abilities? arXiv preprint arXiv:1707.04327.
   6. Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 1-101.
   7. Jordan, M. (2018). Artificial Intelligence – The Revolution Hasn’t Happened Yet. *Medim.com.*
10. Week 10 October 25: Unsupervised Learning, Lab 8, Lab Clinic 5
    1. Principle Component Analysis (ISLR 10.2)
    2. Clustering Methods (ISLR 10.3)
    3. Lab: Neural Networks in R (Handouts)
    4. HW 8 assigned
    5. Lab Clinic 5
    6. Read: Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1), 3133-3181.
    7. Read: Darwiche, A. (2010) Bayesian Networks. *Communications of the ACM*, 53 (12), 80-90, doi :10.1145/1859204.1859227
11. Week 11 November 1: Probability calculus, Bayesian networks: syntax and semantics, Lab Clinic 6
    1. From Propositional to Graded Beliefs (Handouts)
    2. Updating Beliefs (Handouts)
    3. Independence (Handouts)
    4. Capturing Independence Graphically (Handouts)
    5. Parameterizing the Independence Structure (Handouts)
    6. Lab: Unsupervised Learning (ISLR 10.4 – 10.6)
    7. Lab Clinic 6
    8. HW 8 due
    9. Project 3 due
12. Week 12 November 8: Bayesian networks: structure learning & Inference, Lab 7
    1. Reasoning with Bayesian Networks (Handouts)
    2. Learning Bayesian Networks (Handouts)
    3. Inference in Bayesian Networks (Handouts)
    4. Lab 7: Bayesian Network in R
13. Week 13 November 15: Introduction to Search and Network Analysis & Lab Clinic 7
    1. Handouts
    2. Lab Clinic 7
14. Week 14 Nov 22: Thanksgiving
15. Week 15 Nov 29: Final Assignment and Project Reports Due

**Statement on Academic Conduct and Support Systems**

**Academic Conduct**

Plagiarism – presenting someone else’s ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences.  Please familiarize yourself with the discussion of plagiarism in *SCampus* in Section 11, *Behavior Violating University Standards*<https://scampus.usc.edu/1100-behavior-violating-university-standards-and-appropriate-sanctions/>.  Other forms of academic dishonesty are equally unacceptable.  See additional information in *SCampus* and university policies on scientific misconduct, <http://policy.usc.edu/scientific-misconduct/>.

Discrimination, sexual assault, and harassment are not tolerated by the university.  You are encouraged to report any incidents to the *Office of Equity and Diversity* <http://equity.usc.edu/> or to the *Department of Public Safety* <http://capsnet.usc.edu/department/department-public-safety/online-forms/contact-us>.  This is important for the safety whole USC community.  Another member of the university community – such as a friend, classmate, advisor, or faculty member – can help initiate the report, or can initiate the report on behalf of another person.  *The Center for Women and Men* <http://www.usc.edu/student-affairs/cwm/> provides 24/7 confidential support, and the sexual assault resource center webpage [sarc@usc.edu](mailto:sarc@usc.edu) describes reporting options and other resources.

## **Support Systems**

A number of USC’s schools provide support for students who need help with scholarly writing.  Check with your advisor or program staff to find out more.  Students whose primary language is not English should check with the *American Language Institute* <http://dornsife.usc.edu/ali>, which sponsors courses and workshops specifically for international graduate students.  *The Office of Disability Services and Programs* <http://sait.usc.edu/academicsupport/centerprograms/dsp/home_index.html>provides certification for students with disabilities and helps arrange the relevant accommodations.  If an officially  declared emergency makes travel to campus infeasible, *USC Emergency Information* [*http://emergency.usc.edu/*](http://emergency.usc.edu/)will provide safety and other updates, including ways in which instruction will be continued by means of blackboard, teleconferencing, and other technology.