IEOR 142: Introduction to Machine Learning and Data Analytics

Fall 2018

Description:
This course introduces students to key techniques in machine learning and data analytics through a diverse set of examples using real datasets from domains such as e-commerce, healthcare, social media, sports, the Internet, and more. Through these examples, exercises in R, and a comprehensive team project, students will gain experience understanding and applying techniques such as linear regression, logistic regression, classification and regression trees, random forests, boosting, text mining, data cleaning and manipulation, data visualization, network analysis, time series modeling, clustering, principal component analysis, regularization, and large-scale learning.

Note: Students cannot receive credit for both IEOR 142 Introduction to Machine Learning and Data Analytics and IEOR 242 Applications in Data Analysis.

Instructor:
Prof. Paul Grigas
Industrial Engineering and Operations Research
Email: pgrigas (at) berkeley (dot) edu
Office Hours: Tuesday and Thursday 5:15 – 6:15pm, 4177 Etcheverry

Graduate Student Instructors (GSIs):
Meng Qi
PhD Student, Industrial Engineering and Operations Research
Email: meng_qi (at) berkeley (dot) edu
Office Hours: TBD

Nathan Vermeersch
MS Student, Industrial Engineering and Operations Research
Email: nathan.vermeersch (at) Berkeley (dot) edu
Office Hours: TBD

Lecture:
Tuesday and Thursday 3:30 – 5:00pm, 3 LeConte Hall

Discussion:
Friday 3:00 – 4:00pm, 150 GSPP

Course Website and Communication Policy:
Announcements, lecture materials, homework assignments, and all other course materials will be posted on the bCourses site. Additionally, we will use a Piazza forum as the **main electronic communication method** for the course. If you have any questions regarding the course, please post them on Piazza rather than emailing

the course staff. If you have a question or concern that is private in nature (i.e., something you would normally send as an email to the course staff), please use a private post on Piazza so that only the course instructor and GSIs can see your message. You are encouraged to use public posts in situations where other students may benefit from the discussion. In rare exceptional circumstances where your message should be kept confidential from the GSIs, please email the course instructor and begin the subject line with "[IEOR 142]". In summary, you should observe the following priority list for course related communications:

1. Make a public post on Piazza
2. Make a private post on Piazza that only the course instructor and GSIs can see
3. In exceptional circumstances, send an email to the course instructor using "[IEOR 142]" to start the subject line.

We ask that you also please observe the following etiquette on Piazza:

1. **Do not post answers:** Please do not post any answers or your current results on Piazza. Instead, you should explain the key points of your question in a way that allows other students to figure out the essence of the problem on their own. Post problem spoilers after the due date. If you think that your post might give out too much information about the problem solution, then make it private and let the course staff know.
2. **No pre-grading:** We will not answer any questions of the form "Is this the correct way to solve Homework X, Problem Y?"
3. **Aim for public posts:** Other students may have the same question, so please try to make your posts public.
4. **Formatting:** Please format code using the code button and format mathematical equations using the fx button or $$math_equation$$.
5. **Piazza is not office hours:** Please do not ask questions that are too broad, would require a long time to explain in person, etc. These types of questions should be reserved for office hours.
6. **Discussion and collaboration:** We encourage you to answer or comment on your fellow students' posts if you know the answer or would like to discuss.

Prerequisites:
IEOR 165 or equivalent course in statistics. Prior exposure to optimization is helpful but not strictly necessary. Some programming experience/literacy is expected.

Readings/Resources:
The required textbook for this course is:

- *An Introduction to Statistical Learning: with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, Springer, 2013.

A PDF version of this textbook is available at http://www-bcf.usc.edu/~gareth/ISL/.

All readings are recommended (not required) and will complement the lecture material. Most of the time, the readings will cover more material than we will be able to cover in lecture. Depending on your learning style, you may find it helpful to complete the readings either before or after the corresponding lecture.

Some other supplementary texts are:

- *The Analytics Edge* by Dimitris Bertsimas, Allison K. O'Hair and William R. Pulleyblank, Dynamic Ideas LLC, 2016.
- *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* by Trevor Hastie, Robert Tibshirani and Jerome Friedman, Springer, 2009. A PDF version of this textbook is available at https://statweb.stanford.edu/~tibs/ElemStatLearn/. This is an advanced textbook and goes far beyond the material that we will be covering, but this could be a valuable resource for students with a strong mathematical background.
- *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* by Hadley Wickham and Garrett Grolemund, O'Reilly Media, 2017. An online version of this book is available at http://r4ds.had.co.nz/.

Software:
The course will primarily use R, both during class and for the homework assignments. R is a free, open-source programming language and environment that has excellent support for statistical computing and graphics. We recommend that you download both R and RStudio:

https://cran.r-project.org
https://www.rstudio.com

Prior experience in R is not assumed. If and when you run into an issue with R, it is highly likely that someone else has run into the same issue in the past. Therefore, we recommend first searching for your problem on Google, Stack Overflow, etc. If you cannot in earnest find a solution online, then we suggest making a post on Piazza. For the course project, you are free to use another language with good statistical packages (e.g. Python) if you prefer, but the course staff may not be able to provide the same level of support with regard to your code.

Course Objectives:

1. To expose students to a variety of statistical learning methods, all of which are relevant in useful in wide range of disciplines and applications.
2. To carefully present the statistical and computational assumptions, trade-offs, and intuition underlying each method discussed so that students will be

trained to determine which techniques are most appropriate for a given problem.
3. Through a series of real-world examples, students will learn to identify opportunities to leverage the capabilities of data analytics and will see how data analytics can provide a competitive edge for companies.
4. To train students in how to actually apply each method that is discussed in class, through a series of labs and programming exercises.
5. For students to gain some project-based practical data science experience, which involves identifying a relevant problem to be solved or question to be answered, gathering and cleaning data, and applying analytical techniques.
6. To introduce students to advanced topics that are important to the successful application of machine learning methods in practice, include how methods for prediction are integrated with optimization models and modern optimization techniques for large-scale learning problems.

Grading:
There will be a final team project, approximately 5-7 homework assignments, and a midterm during class. Grades for the course will be composed as follows:

1. Individual homework assignments: 40%
2. Final project: 35%
3. Midterm exam: 25%

Assignments:
There will be about five or six individual homework assignments assigned during the semester. The tentative homework schedule will be posted on bCourses, and all assignments will be turned in using Gradescope. We **will not accept late homework submissions**; however, we will also drop your lowest homework score. You are encouraged to begin the homework assignments early as they typically involve a significant amount of coding and data analysis in R.

All homework assignments are *individual* work assignments. However, some collaboration is allowed and even encouraged. You may find it helpful to discuss broad concepts and general solution procedures with others. If this is the case, then you are enthusiastically encouraged to do so. The objective here is to learn. However, the final product that you turn in must be done individually – it must be your own product, written in your handwriting or typed up in a computer file of which you are the sole author. Copying another's work or code is not acceptable. For each exercise, you should be able to explain your solution approach after turning in the assignment – if this is not the case, then the learning objectives of the assignment have not been met and you will be at a disadvantage for the midterm and the project.

You are expected to adhere to the UC Berkeley Code of Student Conduct at all times. In particular, please give credit to outside sources that you find helpful in completing the assignments. These include your peers or other people who you discuss your work with, other textbooks, material from other courses, etc. (There is

no need to cite the course textbooks, slides, or other materials distributed on bCourses.)

Midterm:
There will be a midterm exam during class on **Thursday, October 18**. The purpose of the exam is to gauge your understanding of the material taught so far. If you have been properly completing all of the homework assignments prior to the exam, then you will already be quite well prepared. More details will be given in class as the exam date approaches.

Final Project:
In lieu of a final exam, there will be a final project that should be done in **teams of three students**. The final project provides an opportunity for students to apply analytical methods to a problem in a domain of their choosing.

You will gather (and clean up) data relating to your chosen problem, and use the data analytics techniques discussed in class to solve/answer one or more substantive problems or questions. The project will give students some experience in the kind of work that a data scientist might perform in practice. The requirements for the project are outlined below:

- By **Friday, October 26,** each team must submit a one-page proposal that outlines a plan to apply analytical methods to a problem you identify using some of the concepts and tools discussed in the course. The proposal should include a description of: (1) the problem, (2) the data that you have or plan to collect to solve the problem, (3) which techniques you plan to use, and (4) the impact or overall goal of the project (if you could build a perfect model, what would it be able to do?). The teaching staff will be available to answer questions, and will provide all students with electronic feedback.
- The final project submission will consist of a written report of at most **4 pages** (not including appendices) that describes the analysis, as well as a **5-minute** presentation (in powerpoint or pdf format) of your project. Project presentations will be given on the last day of class, **Thursday, November 29**.
- In order to promote reproducibility and also to give you a chance to advertise your project to prospective employers (or other people), you are encouraged to upload your final report and code to GitHub, and to use GitHub to manage your code and files throughout the project process.
- The four-page report (not including appendices) that describes your analysis is due on the nominal day of the final exam, **Friday, December 14.**

Class Format and Participation:
Lecture slides (and any necessary code and data sets) will be posted on the bCourses site prior to lecture if you wish to print them ahead of time. The class schedule (including readings, assignment due dates, and other information) will be periodically updated on bCourses.

Although parts of lecture may be didactic, we will rely upon interactive discussion within the class. In general, questions and comments are encouraged (as long as they are not disruptive). Furthermore, pointing out mistakes and asking "dumb questions" are also encouraged (very often, a large percentage of the class has the same "dumb question").

Discussion sections, run by the GSIs, will consist of interactive sessions that will cover additional examples of the methods presented in the lectures, and – most importantly – discussion sections will be used to show how to create models in R. Please bring a personal laptop to the discussion sections. Discussion attendance is optional but highly recommended.

Tentative List of Topics (weekly):

1. Introduction, motivating examples, and introduction to R; basics of statistical learning.
2. Linear regression – model fitting, analyzing output, and model validation. In-depth application: Predicting the quality of wine.
3. Logistic regression, linear discriminant analysis, and ROC analysis. In-depth applications: Predicting loan defaults and customer churn rates.
4. Classification and regression trees (CART); cross-validation. In-depth application: Making parole decisions.
5. Advanced tree-based methods: Random forests and boosting. In-depth application: Click-through rate (CTR) prediction.
6. Introduction to text mining; introduction to time series analysis. In-depth applications: Twitter sentiment analysis; predicting sales volume.
7. Clustering; community detection in networks. In-depth applications: Customer segmentation; analyzing email networks.
8. Midterm exam.
9. Data wrangling, visualization, and the data science life cycle.
10. Principal component analysis; collaborative filtering. In-depth application: Recommender systems and Netflix.
11. Integrating predictive models with optimization models. In-depth applications: Internet advertising, kidney allocation.
12. Feature selection and engineering: ridge and LASSO regularization methods; exact In-depth application: Predicting housing prices.
13. Introduction to gradient methods and large-scale learning. In-depth example: Predicting high-quality questions on Stack Overflow.
14. Project presentations.

Notes: The "in-depth application(s)" associated with each method are intended to show how each method is used to solve real-world problems. The order of topics has also been carefully chosen so that students will be first equipped with the most valuable tools for the project. Topics and schedule are subject to changes.