Statistics 422/722 Predictive Analytics
Spring 2017 Course Syllabus

The Wharton School at the University of Pennsylvania

document last updated Wednesday 11th January, 2017 6:31pm

Professor	Adam Kapelner
Contact	`kapelner@wharton.upenn.edu`
Section A (401) Time / Loc	Tuesday 6–9PM / JMHH 250
Section B (403) Time / Loc	Wednesday 4:30–7:30PM / JMHH 240
Office Hours / Loc	TBA
Teaching Assistant	Gemma Moran
Course Homepage	https://github.com/kapelner/Wharton_Stat_422_722

Course Overview

Statistics 422/722 is an introduction to predictive analytics with an emphasis on applications, especially those in business. This seven-week course introduces students to the statistical techniques that extend the ideas of regression analysis introduced in STAT 102/613. Digressing from traditional approaches that focus on carefully modeling how one or two chosen measurements relate to a response, we will take a “modern” approach applicable to managerial decision making in the presence of large data sets.

We will first introduce modeling from a philosophical point of view and define the main terms and main concepts. We will then review least squares regression from 102/613 and place it into this new context, we will round out our regression toolbox by learning how to build models for predicting categorical responses. Equipped with a solid foundation, we will switch our approach to the point of view of predictive modeling using automatic tools with a focus on predicting a response under new data. If, for example, we can show a bank how to predict who will default on a loan better than their existing system, the bank can increase profits. Similarly, if we help a company identify those in the market most interested in its products, then it can construct a much more focused product launch.

We will end the class with so called “black-box” models which feature better predictive accuracy but less interpretability. We will talk about this tradeoff and hopefully try to get a glimpse into how the black-box models work.

As the business world rapidly progresses towards a paradigm of data-driven decision making, the *primary goal* of this course is on understanding both the power and limitations of regression analysis. The course is designed to allow future managers–both data scientists and not– to communicate effectively with the data science team within an organization.

We are going to let software do the number crunching for us – our value-add comes from how to choose to tackle the problem and what insights we can draw from the model results.

A tentative list of topics is below

- Modeling, Mathematical Modeling, Statistical Modeling, Causal Models
- An introduction to: AI, Machine Learning, Deep Learning and Linear Regression
- Framework for assessing model fit, p values, R^2, t-tests, AIC
- Classification, the Logistic Regression Models
- Automatic Model Selection and High Dimensional Linear Regression
- Assessing prediction quality
- Classification and Regression Trees
- Basic machine learning

Course Materials

Textbook: There is no required textbook. However there will be readings from Silver (2012) and there is a rather lengthy optional reading list at the end of this syllabus

Lecture Slides: The lecture slides and notes I put on the board will be primary for learning material in this course.

Computer Software: We will also be using `JMP` version 12 pro, software that you either already own from 102/613 or have access to by using the Wharton computer labs. While most of the features we will need for this course are available in the standard edition of JMP, a few features require the Pro edition. JMP Pro can be downloaded from Canvas. Mac and Windows versions are available. Instructions are provided on Canvas. Manuals for JMP can be found here. I will reference appropriate chapters in each slide deck. Spending some quality time with the manuals is part of learning any software tool.

We will also make use of `R` which is a free, open source statistical programming language and console. It is the de factor choice for statistical programming today. You can download it from: `http://cran.mirrors.hoobly.com/`. I do not expect you to do *any* programming. I will be giving you `R` code to run and expect you to interpret the results based on concepts explained during the course.

`Python` would be my next recommendation as it is also very popular among data scientists. But we have limited time so I will likely not get to any Python examples.

Also, please note that Microsoft Excel *will not be sufficient* for executing many of the analyses in the course, but you may use it as you wish to help explore and clean data or generate output.

Calculator: You can use a TI-84, 85, 89 or any calculator which you wish for exams.

Announcements

Announcements will be made via email. Course grades will be posted on

Lectures

I recommend not using your computer / tablet / phone during lectures — only pen / pencil and paper. Classes are 180 minutes and run from Tuesday, January 17 until Wednesday, March 1. There will be 7 lecture periods with the last two hours of the last class being the final exam (the first hour will be a Q&A session).

Assignments

Homework

There will be 3 or 4 homework assignments plus one project.

Homeworks will be assigned and placed on the course homepage and will usually be due a week later in class. Homework will be **graded** out of 100. There may be extra credit and thus scores can be $>$ 100. I and the TA and the grader will be doing the grading. We reserve the right to grade an *arbitrary subset of the assignment* which is determined after the homework is handed in. But you will still be penalized for leaving questions blank regardless of whichever we grade.

Homework must be printed, neat and stapled (**it cannot be emailed to me**). Homework can be given to me in class or delivered to my cardboard drop box in the Statistics department (4th floor of Huntsman Hall).

Graded homework will be returned in class. Regrades are handled during office hours or right after class is over. Scores for homeworks are finalized one week after the graded copies are handed back. Thereafter there will be no changes and no re-grading.

Do not delay checking your graded homeworks. I (and the TAs) are not perfect and we do make mistakes. It is your obligation to find our mistakes and report them.

You are highly recommended to work with each other and help each other. You must, however, submit your own solutions, *with your own write-up* and in *your own words*. There can be no collaboration on the actual *writing*. The university honor code is something I take very seriously.

Forecasting Competition

Kaggle, a startup based in San Francisco, offers a solution to clients requiring predictive models for their businesses by holding forecasting competitions. Any data scientist is allowed to build and submit forecasting models and the winning model receives a monetary prize in addition to being implemented by the sponsoring company.

We will have our own forecasting competition. You will be required to submit nothing but your predictions on a data set. This is meant to be a chance to get creative in predictive modeling without needing to worry about model interpretability.

The Project

There will also be a project with a writeup. It will likely be related to the forecasting competition.

Philosophy of Homework

Homework is an important part of this course. Success in Statistics and Mathematics courses comes from experience in working with and thinking about the concepts. It's kind of like weightlifting; you have to lift weights to build muscles. My job as an instructor is to provide assistance through your zone of proximal development. You can grow more in a class than you can alone but the growth comes from you. To this effect, homework problems are color coded **green** for easy, yellow for harder, **red** for challenging and purple for extra credit. You need to know how to do all the greens by yourself. If you've been to class and took notes, they are a joke. Yellows and reds: feel free to work with others. Only do extra credits if you have already finished the assignment.

Late Homework

Late homework will be penalized 10 points per day for a maximum of five days. Do not ask for extensions; just hand in the homework late. After five days, **you can hand it in whenever you want** until the last day of class, Wednesday, March 1. I realize things come up. Do not abuse this policy; you will fall behind.

Homework and Project LaTeX Bonus Points

Beautiful presentation counts. Thus, **there will be abonus** added to your homework grade for typesetting your assignment writeup using the LaTeX system. The bonus will be 1–10 points based on the elegance of your presentation and thus the bonus points are arbitrarily determined by me. LaTeX is a good skill to know no matter what kind of career you are in.

If you would like to use LaTeX, I recommend using overleaf to write up your homeworks (make sure you upload both the hw#.tex and the preamble.tex file Pprovided on the course homepage for each homework assignment). Overleaf has the advantage of (a) not having to install anything on your computer and not having to maintain your LaTeX installation (b) allowing easy collaboration with others (c) **always having a backup of your work since it's always on the cloud** but it can be quite slow. For speed you may insist on having LaTeX running on your computer, you can download it for Windows here and for MAC here. For editing and producing PDF's, I recommend TeXworks which can be downloaded here.

If you are handing in homework this way, read the comments in the code; there are two lines to comment out and you should replace my name with yours and write your section. The easiest way to use overleaf is to copy the raw text from hwxx.tex and preamble.tex into two new overleaf tex files with the same name. If you are asked to make drawings, you can take a picture of your handwritten drawing and insert them as figures or leave space using the "\vspace" command and draw them in after printing or attach them stapled.

Since this is completely extra credit and independent of the course material and goals, do not ask me or the TA for help in setting up your computer with LaTeX in class or in office hours. Also, **never share your LaTeX code with other students** — it is cheating.

Final Examination

The final examination will be mostly in the style of the homeworks. If you can do all the green and yellow problems on the homeworks, the exam should not present any challenge. I will *never* give you exam problems on concepts which you have not seen at home on one of the weekly homework assignments.

The final exam will be in class during the last two hours of the last lecture period (i.e. 2/28/17 7-9PM for Section A and 5:30-7:30PM 3/1/17 for Section B). The first hour will be a general Q&A session or material that is not covered on the exam.

Exam Materials

I allow you to bring any calculator you wish but it cannot be your phone. The only other items allowed are pencil and eraser. I do not recommend using pen but it is allowed

I also allow “cheat sheets” for the final exam: you are allowed to bring three 8.5” × 11” sheet of paper (front and back). On these sheets you can write or print anything you would like which you believe will help you on the exam. You must hand in your cheat sheets if you are in Section A.

Special Services

If you are a student who takes exams at the special services center, I need to see your blue slip one week before the exam to make proper arrangements with the center.

Class Participation (and attendance)

Given that this course is only 7 lectures with 150 students, I cannot learn your names. Thus, I do not think it is fair to use class participation or attendance into the grade calculation.

Grading and Grading Policy

Your course grade will be calculated based on the percentages as follows:

Homework	30%
Final Examination	40%
Forecasting Competition Accuracy	15%
Project	15%

Course grades are given on an approximate Wharton school curve.

Checking your grade and class standing

You can always check your grades in real-time using the grading site. You will enter in your email and the password I will provide to you via email.

Use of Canvas

I will not be using canvas to post assignments, instead, see the course homepage. However, you are encouraged to use the discussion board to post questions and comments about the course. I encourage you all to reply to your fellow classmates inquiries, as one of the best ways to learn is through teaching another. The course staff will do its best to provide answers as well. I may also post pertinent questions I receive via email in addition to my reply, leaving the original student’s identity anonymous.

If you are not a Wharton student and need access to Canvas, you can create an account by visiting `accounts.wharton.upenn.edu` once the semester begins.

Auditing

Auditors are welcome only with a permit from the professor as there are limited seats in the rooms and people sitting on the steps poses a fire risk. If you are auditing, you are encouraged to do all homework assignments and we will even grade them.

Course Acknowledgments

Special thanks to Justin Bleich who shared his course materials with me and discussed my curriculum.

References

Abbott, D. (2014). *Applied predictive analytics: principles and techniques for the professional data analyst.* John Wiley & Sons.

Abu-Mostafa, Y. S., Magdon-Ismail, M., and Lin, H.-T. (2012). *Learning from data*, volume 4. AMLBook Singapore.

Alpaydin, E. (2014). *Introduction to machine learning.* MIT press.

Bell, J. (2014). *Machine Learning: Hands-on for developers and technical professionals.* John Wiley & Sons.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Information Science and Statistics. Springer.

Box, G. E., Draper, N. R., et al. (1987). *Empirical model-building and response surfaces*, volume 424. Wiley New York.

Burnham, K. P. and Anderson, D. (2003). Model selection and multi-model inference. *A Pratical informatio-theoric approch. Sringer.*

Covington, D. (2016). *Analytics: Data Science, Data Analysis and Predictive Analytics for Business.* CreateSpace Independent Publishing Platform.

Finlay, S. (2014). *Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods.* Business in the Digital Economy. Palgrave Macmillan UK.

Freund, R. J., Wilson, W. J., and Sa, P. (2006). *Regression analysis.* Academic Press.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer (freely available here).

Gershenfeld, N. A. (1999). *The nature of mathematical modeling.* Cambridge university press.

Glymour, C., Scheines, R., Spirtes, P., and Kelly, K. T. (1987). Discovering Causal Structure: Artifical Intelligence, Philosophy of Science and Statistical Modeling.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press (freely available here).

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 6. Springer (freely available here).

Konishi, S. and Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer Science & Business Media.

Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*. Springer.

Miller, T. W. (2014). *Modeling techniques in predictive analytics with Python and R: a guide to data science*. FT Press.

Mitchell, T. M. (1997). *Machine learning*. McGraw Hill.

Provost, F. and Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. "O'Reilly Media, Inc.".

Ratner, B. (2004). *Statistical modeling and analysis for database marketing: effective techniques for mining big data*. CRC Press.

Silver, N. (2012). *The signal and the noise: Why so many predictions fail-but some don't*. Penguin.

Wu, J. and Coggeshall, S. (2012). *Foundations of predictive analytics*. CRC Press.