LSE-PKU Summer School 2020

LPS-MY201 | Big Data: Data Analytics for Business and Beyond

Instructor

Qiwei Yao, Professor of Statistics at London School of Economics and Political Science, Distinguished Visiting Professor at Guanghua School of Management of Peking University.

Qiwei Yao is a leading expert in high-dimensional time series analysis and nonlinear time series analysis. He is Fellow of the Institute of Mathematics Statistics, Fellow of the American Statistical Association, and Elected Member of the International Statistical Institute. His current research focuses on modelling and forecasting with vast time series data.

Qiwei Yao has undertaken extensive data analytics consultancy projects from major industry companies including Barclays Bank, Electricite de France (EDF), and Winton Capital Management Ltd.

Course Summary

In this modern information age, the broad availability of Big Data (i.e. data of unprecedented sizes and complexities) brings opportunities with challenges to business and beyond. Companies are focused on exploiting data for competitive advantages. Cyberspace communication reveals complex social interactions. Big Data surveillance is an effective way to detect actionable security threats. Data analytics is a subject of learning from data, of measuring, controlling, and communicating uncertainty, and of data-driven decision-makings (DDD). It will become ever more critical as businesses, governments and also academia rely increasingly on DDD, expanding the demand for data analytics expertise.

The primary goal of this course is to help you view various problems from business, science and social domains from a data perspective and understand the principles of extracting useful information and knowledge from data. You will also gain the hands-on experience using R– a programming language and software environment for data analysis and graphics. (R is free and available from http://www.r- project.org/.) The focus is on and

basic principles and concepts of data analytic methodology. We will also point out the limitation of data analysis: one should not be carried away by the findings from data and models. Common sense, intuition, domain knowledge and creativity often play roles in good data analytics.

To achieve this primary goal, inevitably we will introduce some basic data analytic methods and illustrate them with real-life examples (some from China). This is the second goal of the course. Data analytics has multiple facets and approaches, encompassing diverse statistical techniques under a variety of names such as data mining, machine learning. The methods to be covered include:

- Classification. Among all customers of EDF, who are likely to switch to another energy supplier?
- Regression (i.e. value estimation.) How much will a given customer use the service?
- Similarity matching. Identify individuals who are similar to your most royal customer group.
- Clustering. How should our customer care teams be structured?
- Market-basket analysis. Should beers be placed next to baby napkins in a supermarket.
- Link prediction. As you and John share 10 friends, maybe you would like to be John's friend?
- Causal modelling. Is the increase of sales caused by a particular advertisement?
  This is not a course on algorithms and IT technologies required for handling massive data, which deserve separate courses. The focus is on the fundamental principles and concepts of data analytics or data science. It becomes ever-increasingly important in this information age to gain adequate understanding of data science even if you never intend to apply it yourself.

Course Overview

1. Introduction: data-analytic thinking, data mining for knowledge discovery, data science solution for business problems.

2. Predictive modelling: correlation and supervised learning, regression and classification, support- vector machines, overfitting and its avoidance.

3. Clustering data: similarity, nearest neighbours, unsupervised learning methods.

4. Decision analytic thinking: what is a good model, visualizing model performance, evidence and probabilities.

5. Additional topics: text mining and network data

Prerequisites

Knowledge of calculus and statistics at the undergraduate freshman level. Participants should also bring a laptop and a calculator (calculators will be needed in the final examination).

Assessment

Coursework (30%) and final exam (70%).

Recommended Preparatory Readings

[1] Provost, F. and Fawcett, T. (2013). Data Science for Business. O'Reilly. [2] Runkler, T.A. (2012). Data Analytics. Springer.

[3] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer.

Students may choose to read any one of the above three books. They are listed in the ascending order in terms of the technical level, as [3] is technically the most advanced. [3] also illustrates how to implement data analytic methods in R.

References for R:

[4] Venables N. et. al. (2015). An Introduction to R. is available online at http://cran.r-project.org/doc/manuals/R-intro.pdf.

[5] Zuur, A., Ieno, E. and Meesters, E. (2009). A Beginners Guide to R. Springer. [1]