Statistical Learning in the Age of "Big Data" and Machine Learning

Miriam Frey and Mario Larch

Chair of Empirical Economics

June 26, 2017

1 General Issues

The course "Statistical Learning in the Age of "Big Data" and Machine Learning" addresses master students of Business Administration, Economics, Internationale Wirtschaft und Governance, and Philosophy & Economics. Advanced interested bachelor students may also participate.
Nowadays it is not uncommon that one wants to analyse "big data", i.e. really big datasets (such as credit card transaction data, scanner data of a firm, huge firm-level balance-sheet datasets, patent data, tweets, google search queries,...). Such big datasets lead to new challenges to estimation (see Varian, 2014):

1. due to the sheer size, one may need more powerful data manipulation tools.

2. with large datasets one may end up with more potential predictors than appropriate for estimation with classical estimation methods.

3. with large datasets one may have the possibility to allow for more flexible relationships than simple linear models.

Hence, traditional estimation methods may have to be complimented or even replaced by alternative estimators or estimation strategies. A new branch in computer science, “machine learning”, provides exactly such tools, such as decision trees, support vector machines, boosting, and random forests. These methods need a lot of data, but given that, allow for more effective ways to model complex relationships.

The seminar consists of two parts:

1. Student presentations will cover the recent, and most widely used methods of statistical learning. We provide the topics for the presentation (including the respective reading material). Our seminar will be based on the book James, Witten, Hastie, and Tibshirani (2013), which is available for download in our library and will be abbreviated from now on by “ISL”. Chapters one, two and three will be compulsory reading for everyone. Please register in the e-learning for the seminar to obtain all information and materials. Our first **compulsory introductory meeting** will be on **October 24th, 2017, 2pm-4pm (c.t.)** (the room will be announced in the e-learning), where we will discuss organizational issues and allocate topics. The date for the **presentation** of your topics is **December 8th and 9th, 2017** (depending on the number of participants). Please send us your **presentation slides** at least two days in advance.

2. To practice the learned methods, you will also implement the estimator of your topic with the provided *R* code from the book. Note that chapter 2 also contains an introduction to *R*. You will write a seminar thesis about your method, your reproduction and apply your method to an example of your choice. Datasets and suggestions are also contained in the book and the accompanying homepage (www.statlearning.com).

Interested students are asked to **sign up** by sending an email to Sandra Hörath (vwl6@uni-bayreuth.de) before **October 20th, 2017**, indicating your previous knowledge in statistics and empirical economics. Furthermore, please give three preferences for topics, including at least one from the list of the core topics.

For further questions concerning course details please contact Mario Larch (mario.larch@uni-bayreuth.de) or Miriam Frey (miriam.frey@uni-bayreuth.

de).

2 Requirements and Assessed Course Work

Requirements
In order to participate in the course, interest in and good knowledge of empirical economics at the level of Advanced Empirical Economics I is expected. Ideally, this includes some programming knowledge.

Assessed Course Work
The assessed course work consists of a term paper at the end about your topic including a reproduction of the example in *R* with possible extensions and further sensitivity analysis regarding implementation choices. In the term paper, you should on the one hand introduce thoroughly into the topic and on the other hand explain in detail the methods used and specifically show how the specific analysis is conducted and carefully interpret (as far as possible) your results. The term paper should consist of about 20,000 characters (including space characters).
After the presentations you write your term paper based on the knowledge you gained in the course. The date of submission of the term paper will be March 31st, 2018 (of course an earlier submission is possible at any time).

Language and Formal Requirements
The language of the course (and hence your presentation and presentation slides) is English. Hence, all the literature is in English. Your term paper can be written in German or in English, even though we suggest to write it in English. For more details concerning the formal requirements of the written assignments please see the style sheet available in German (Hinweis zur Formatierung von Seminar- und Abschlussarbeiten) and in English (Formal requirements for seminar papers and bachelor's/master's theses at the Chair of Economics VI: Empirical Economics).

3 Target Group

The course addresses students from the following degree courses:

- Betriebswirtschaftslehre (MA): as substitute for “Advanced Empirical Economics II” (which is part of the bloc “B 1 Forschungsmethoden” and of the bloc “V Empirische Wirtschaftsforschung”).
- Economics (MA): as substitute for “Advanced Empirical Economics II” (which is part of the specialization “Modelltheorie”) and “International Labor Markets” (which is part of the specialization “Internationale Wirtschaft”) or as “Individueller Schwerpunkt”.
- Internationale Wirtschaft und Governance (MA): as substitute for “Advanced Empirical Economics II” (which is part of the specialization “Ökonomische Modellbildung und empirische Analyse”) and “International Labor Markets” (which is part of the bloc “Internationale Wirtschaft”) or as “Individueller Schwerpunkt”.
- Philosophy and Economics (MA): as electives course.
- History and Economics (MA): as specialization.

Additionally, interested bachelor students may participate.

4 Reading List

In order to have a common base for discussion in class, you are all required to read the first three chapters of James, Witten, Hastie, and Tibshirani (2013). As further useful general background, we provide the following reading list:

- Varian (2014),
- Hastie, Tibshirani, and Friedman (2009) (also available for download in our library),

- six articles published from a symposium on “Recent Ideas in Econometrics” published in the *Journal of Economic Perspectives*, vol. 31 no. 2, Spring 2017:
  - Athey and Imbens (2017),
  - Low and Meghir (2017),
  - Stock and Watson (2017),
  - Mullainathan and Spiess (2017),
  - Powell (2017),
  - Angrist and Pischke (2017).

5 Presentation Topics

Notes:

- When you indicate your preferences for topics, you have to choose at least one from the list of core topics.
- Depending on the number of participants, topics may be allocated to teams of two. You are therefore also free to indicate potential partners for the presentation topic.
- All section references refer to ISL if not explicitly otherwise mentioned.
- Data and code for the book ISL can be downloaded at www.StatLearning.com.
- The programs are in *R*. *R* is a widely used statistical software package, that contains many machine learning algorithms. It is open source, and you can therefore download and us it for free. To do so, go to the official homepage https://www.r-project.org/. If you prefer a nice user interface, you may want to use *RStudio* (https://www.rstudio.com/). It provides you with a graphical user interface. Otherwise, it uses *R*. *R* and *RStudio* are both available in the computer labs.

- There are many books about R, which can be free download as pdf in our library, as for example Zuur, Ieno, and Meesters (2009).
- You may also want to consult the more advanced treatment of the topics dealt with in ISL in Hastie, Tibshirani, and Friedman (2009) for your chosen topic for the seminar presentation and thesis.

5.1 Core Topics

1. Logistic Regression (Sections 4.3 and 4.6.2).
2. Linear Discriminant Analysis (Sections 4.4 and 4.6.3).
3. Subset Selection (Sections 6.1 and 6.5).
4. Shrinkage Methods (Sections 6.2 and 6.6).
5. Dimension Reduction Methods (Sections 6.3 and 6.7).
6. Regression Trees (Sections 8.1.1 and 8.3.2).
7. Classification Trees (Sections 8.1.2 and 8.3.1).
8. Bagging and Random Forests (Sections 8.2.1, 8.2.2 and 8.3.3).
9. Boosting (Sections 8.2.3 and 8.3.4).
10. Support Vector Classifiers (Sections 9.2 and 9.6.1).

5.2 Further Topics

1. Cross-Validation (Sections 5.1 and 5.3.1-5.3.3).
2. The Bootstrap (Sections 5.2 and 5.3.4).
3. Polynomial Regression and Step Functions (Sections 7.1, 7.2 and 7.8.1).
4. Splines (Sections 7.4, 7.5 and 7.8.2).
5. Generalized Additive Models (Sections 7.7 and 7.8.3).

6. Support Vector Machines (Sections 9.3 and 9.6.2).
7. Support Vector Machines with More than Two Classes (Sections 9.4 and 9.6.4).
8. Principal Components Analysis (Sections 10.2 and 10.4).
9. Clustering Methods (Sections 10.3 and 10.5).

6 Overview of Important Dates

- October 20th, 2017: registration deadline.
- October 24th, 2017, 2pm-4pm (c.t.): compulsory introductory meeting, allocation of presentation topics.
- December 6th, 2017: submission deadline for presentation slides.
- December 8th and 9th, 2017: student presentations.
- March 31st, 2018: submission deadline for seminar papers.

References

Angrist, J., and J.-S. Pischke (2017): “Undergraduate Econometrics Instruction: Through Our Classes, Darkly,” *Journal of Economic Perspectives*, 31(2), 125–144.

Athey, S., and G. Imbens (2017): “The State of Applied Econometrics: Causality and Policy Evaluation,” *Journal of Economic Perspectives*, 31(2), 3–32.

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Springer, New York, 2 edn.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013): *An Introduction to Statistical Learning - with Applications in R*. Springer, New York, 1 edn.

Low, H., and C. Meghir (2017): “The Use of Structural Models in Econometrics,” *Journal of Economic Perspectives*, 31(2), 33–58.

Mullainathan, S., and J. Spiess (2017): “Machine Learning: An Applied Econometric Approach,” *Journal of Economic Perspectives*, 31(2), 87–106.

Powell, J. (2017): “Identification and Asymptotic Approximations: Three Examples of Progress in Econometric Theory,” *Journal of Economic Perspectives*, 31(2), 107–124.

Stock, J., and M. Watson (2017): “Twenty Years of Time Series Econometrics in Ten Pictures,” *Journal of Economic Perspectives*, 31(2), 59–86.

Varian, H. (2014): “Big Data: New Tricks for Econometrics,” *Journal of Economic Perspectives*, 28(2), 3–28.

Zuur, A., E. Ieno, and E. Meesters (2009): *A Beginner’s Guide to R*. Springer, New York, 1 edn.