Lectures on Machine Learning

NBER, Saturday, July 18th, 2015

Susan Athey & Guido Imbens

Outline

In these lectures we will discuss some methods from the machine learning (ML) literature that we think will be useful for economists. There has been a fast growing literature in computer science and related fields developing new, and modifying existing, methods for analyzing large data sets. This literature builds heavily on traditional statistical methods, though often with new terminology and new questions. Many of these methods are focused on finding patterns in data sets that are useful for prediction and classification. The focus in this literature is typically on methods that "work," more than on deriving asymptotic (large sample) results of the type that are common in the econometrics and mathematical statistics literature. There is also less emphasis on confidence intervals and standard errors. On the other hand there is a heavy emphasis on out-of-sample comparisons, in particular cross-validation. There is generally also less emphasis on causal effects as opposed to prediction.

Here we discuss some of the most prominent methods. These include regularized regression, where least squares parameter estimates are shrunk towards zero, with the most popular method LASSO. We also look at regression trees and their extensions, including random forests. We will also discuss unsupervised learning methods such as clustering algorithms. We discuss some of the principles, and in particular how these methods relate to methods that are traditionally more familiar to economists, and how they can be implemented in practice.

We also discuss some of the more recent literature where the focus is on modifying these methods to estimate causal effects.

Breiman (2001a) is a very good easy to read paper on the differences in culture between the traditional statistics literature (which is very much like the econometrics literature) and the machine learning literature. If there is one paper to read before the lectures, it is that one. Wu et al. (2008) is a good example of the difference in the cultures: the goal is often to derive algorithms rather than statistical properties. Even if some of the algorithms (like the EM algorithm and nearest-neighbor methods) are familiar to economists, the philosophy

in the ML literature is very different. Tibshirani (1996) is the original paper on the Lasso, and gives a very good introduction to the basic ideas. The recent (2015) book by Hastie and Tibshirani is very good, and easier to read than Hastie, Tibshirani, and Friedman (2011) which is the classic reference. Breiman, Friedman, Olshen, and Stone (1984) is the classic book on trees. Breiman (2001b) is a good paper on random forests, which are also discussed in Hastie, Tibshirani, and Friedman (2011). Alpaydin (2014) is a highly recommended introduction to machine learning. Leskovec, Rajaraman, and Ullman (2014) is a very readable book that includes great discussions of unsupervised learning methods. Vapnik (1998) is a comprehensive book on support vector machines, which is another prominent technique. Schapiro and Freund (2012) contains an in-depth discussion of boosting methods by two of the original contributors to that literature. The forthcoming book by Efron and Hastie (see Efron's website) is likely to be a classic. Tan, Steinbach and Kumar (2006) and Newman (2006) discuss unsupervised learning methods. Imbens and Rubin (2015) is a general book on causality. Athey and Imbens (2015) discuss new machine learning methods for causal effects; Athey (2015) provides a short overview of three distinct directions for research in this area. Belloni, Chernozhukov and Hansen (2011, 2014) survey the recent econometric literature. Recent applications and discussions in the economics literature include Varian (2014), Einav and Levin (2014), Bajari, Nekipelov, Ryan, and Yang (2015), Chernozhukov, Hansen, and Spindler (2015), Gilchrist and Glassberg Sands (2015), and Kleinberg, Ludwig, Mullainathan, and Obermeyer (2015).

References

Aplpaydin, E., (2014), *Introduction to Machine Learning*, Third Edition MIT Press.

Athey, S. (2015), "Machine Learning and Causal Inference for Policy Evaluation," forthcoming in *Proceedings of the 21th ACM SIGKDD international conference on Knowledge Discovery and Data Mining.* http://faculty-gsb.stanford.edu/athey/documents/AtheyKDDfinal.pdf

Athey, S., and G. Imbens (2015), "Machine Learning Methods for Estimating Heterogeneous Causal Effects" NBER working paper.

Bajari, P., D. Nekipelov, S. Ryan, and M. Yang, (2015), "Machine Learning Methods for Demand Estimation," *American Economic Review*, Papers and Proceedings, Vol. 105(5): 481-85.

Belloni, A., V. Chernozhukov, and C. Hansen, (2011), "Inference Methods for High-Dimensional Sparse Econometric Models," *Advances in Economics & Econometrics*, Econometric Society World Congress.

Belloni, A., V. Chernozhukov, and C. Hansen, (2014), "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, 28(2): 29-50.

Breiman, L. (2001a), "Statistical Modeling: The Two Cultures," *Statistical Science*, Vol. 16(3): 199-215.

Breiman, L. (2001b), "Random forests," *Machine Learning*, 45, 5-32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone, (1984), *Classification and Regression Trees*, Wadsworth.

Chernozhukov, V., C. Hansen, and M. Spindler, (2015), "Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments," *American Economic Review*, Papers and Proceedings, Vol. 105(5): 486-90.

Efron, B., and T. Hastie, (2016), *Computer-Age Statistical Inference*.

Einav, L., and Levin, J., (2014), "Economics in the age of big data." *Science*, 346(6210), 1243089.

D. Gilchrist, and E. Glassberg Sands, (2015), "Something to Talk About: Social Spillovers in Movie Consumption," Unpublished Manuscript, http://scholar.harvard.edu/files/dgilchrist/files/socialspillo

Hastie, T., Tibhshirani, R. and M. Wainwright, (2015), *Statistical Learning with Sparsity: The Lasso and Generalizations*, Chapman and Hall.

Hastie, T., R. Tibshirani, and J. Friedman, (2011), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, Springer.

Imbens, G., and D. Rubin, (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer, (2015), "Prediction Policy Problems," *American Economic Review*, Papers & Proceedings, 105(5): 491495.

Leskovec, J., A. Rajaraman, and J. Ullman, (2014), *Mining of Massive Datasets*, second edition, Cambridge University Press.

Newman, M. (2006), "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, 103(23), 8577-8582.

Schapiro, R., and Y. Freund, (2012), *Boosting*, MIT Press.

Tan, P., M. Steinbach, and V. Kumar, (2006), *Introduction to Data Mining*, Addison Wesley.

Tibshirani, R., (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistal Society*, Series B., Vol. 58, No. 1, pages 267-288.

Vapnik, V., (1998), *Statistical Learning Theory*, Wiley.

Varian, H., (2014), "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, 3-27.

Wu, X., V. Kumar, R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand, and D. Steinberg, (2008), "Top 10 algorithms in data mining," *Knowledge in Information Systems*, 14:137.

Schedule

1:30-1.35 Welcome

1:35-2.30 Introduction to Supervised ML Concepts and Algorithms (Guido Imbens)

2:30-2.40 Break

2:40-3.35 Introduction to Unsupervised ML Concepts and Algorithms (Guido Imbens)

3:35-3.45 Break

3:45-4.40 Machine Learning and Causal Inference (Susan Athey)

4:40-4.50 Break

4:50-5.45 Unsupervised Learning: Applications to Networks and Text Mining (Susan Athey)