Machine Learning Methods for Economists

Stephen Hansen, stephen.hansen@economics.ox.ac.uk

1 Textbooks / Overview Material

The following textbooks will cover all relevant material for the course and much more:

1. Hastie et al. 2009: good introduction to data science.
2. Hastie et al. 2015: recent textbook covering LASSO and extensions.
3. Manning et al. 2009: good introduction to basics of information retrieval (MRS in references below).
4. Murphy 2012: probabilistic, and in particular Bayesian, perspective on machine learning (KM in references below).

Purchasing these is not required, and I will provide self-contained lecture notes. Grimmer and Stewart (2013), Bholat et al. (2015), and Gentzkow et al. (2017) provide accessible introductions to text mining and machine learning.

2 Penalized Regression

Background:

- Meinshausen and Bühlmann (2006)
- Zou (2006)
- Meinshausen and Bühlmann (2010)

Economic Applications:

- Belloni et al. (2014b)
- Belloni et al. (2014a)
- Athey and Imbens (2016)
- Wager and Athey (2018)

3 Text as Data

Background:

- MRS 1, 2.2, 6.1-6.3
- KM 2.5.4, 3.3-3.4

4 Unsupervised Learning

4.1 Finite mixture models and EM algorithm

Background:

- KM 11

4.2 Singular value decomposition

Background:

- MRS 18
- Deerwester et al. (1990)

Applications:

- Boukus and Rosenberg (2006)
- Hendry and Madeley (2010)
- Acosta (2014)

4.3 Latent Dirichlet allocation

Background:

- KM 27.1-27.3.2, 27.3.1-27.3.6; 21
- Blei et al. (2003)
- Blei and Lafferty (2009)
- Wainwright and Jordan (2008)

Applications and extensions:

- Quinn et al. (2010)

- Hansen et al. (2014)
- Hansen and McMahon (2015)
- Mueller and Rauh (2016)
- Blei and Lafferty (2006)
- Roberts et al. (2016)

5 Generative Models for Text Regression

Background:

- MRS 13
- Mcauliffe and Blei (2008)
- Taddy (2013)
- Taddy (2015)

Applications:

- Gentzkow and Shapiro (2010)

References

Acosta, J. M. (2014). FOMC responses to calls for transparency: Evidence from the minutes and transcripts using latent semantic analysis. Mimeograph, University of Stanford.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

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

Belloni, A., Chernozhukov, V., and Hansen, C. (2014b). Inference on Treatment Effects after Selection among High-Dimensional Controls. *Review of Economic Studies*, 81(2):608–650.

Bholat, D., Hansen, S., Santos, P., and Schonhardt-Bailey, C. (2015). Text mining for central banks. Centre for Central Banking Studies, Handbook No. 33, Bank of England.

Blei, D. and Lafferty, J. (2006). Dynamic topic models. In *Proceedings of the 23rd Internaional Conference on Machine Learning*, pages 113–120.

Blei, D. and Lafferty, J. (2009). Topic models. In Srivastava, A. and Sahami, M., editors, *Text Mining: Classification, Clustering, and Applications*. Taylor & Francis, London, England.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Boukus, E. and Rosenberg, J. V. (2006). The information content of FOMC minutes. Mimeograph, Federal Reserve Bank of New York.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Gentzkow, M., Kelly, B. T., and Taddy, M. (2017). Text as Data. NBER Working Papers 23276, National Bureau of Economic Research, Inc.

Gentzkow, M. and Shapiro, J. M. (2010). What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*, 78(1):35–71.

Grimmer, J. and Stewart, B. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, pages 1–31.

Hansen, S. and McMahon, M. (2015). Shocking Language: Understanding the Macroeconomic Effects of Central Bank Communication. *Journal of International Economics*. forthcoming.

Hansen, S., McMahon, M., and Prat, A. (2014). Transparency and Deliberation within the FOMC: a Computational Linguistics Approach. CEPR Discussion Papers 9994, C.E.P.R. Discussion Papers.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2 edition.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Number 143 in Monographs on Statistics and Applied Probability. CRC Press.

Hendry, S. and Madeley, A. (2010). Text mining and the information content of bank of canada communications. Working Paper 2010-31, Bank of Canada.

Manning, C. D., Raghavan, P., and Shütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge University Press.

Mcauliffe, J. D. and Blei, D. M. (2008). Supervised Topic Models. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 121–128. Curran Associates, Inc.

Meinshausen, N. and Bühlmann, P. (2006). High-Dimensional Graphs and Variable Selection with the LASSO. *The Annals of Statistics*, 34(3):1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society Series B*, 72(4):417–473.

Mueller, H. and Rauh, C. (2016). Reading between the lines: Prediction of political violence using newspaper text.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning. MIT Press.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., and Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

Roberts, M. E., Stewart, B. M., and Airoldi, E. M. (2016). A Model of Text for Experiments in the Social Sciences. *Journal of the American Statistical Association*, 111(515):988–1003.

Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108.

Taddy, M. (2015). Distributed Multinomial Regression. *The Annals of Applied Statistics*, 9(3):1394–1414.

Wager, S. and Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*. forthcoming.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429.