Introduction to Data Science

Reading List, Winter 2018

Oliver Ernst

January 16, 2019

Books

Textbooks

- James et al. (2013), available online here.
  This will be the primary source for the course.
- Hastie, Tibshirani, and Friedman (2001), available online here.
  A more technical and comprehensive precursor to (James et al., 2013).

Statistics

- Pichler (2018): Lecture notes for the TU Chemnitz undergraduate statistics class, which is recommended for all MSc Data Science students without an undergraduate math degree.
- Freedman, Pisani, and Purves (2007): A very elementary and non-technical introduction into statistical terminology and thinking.
- Williams (2010): A very lively and mathematicaly satisfying account of statistics and probability theory at the beginning graduate level.
- Efron and Hastie (2016): A very readable account of classical and modern statistical ideas, available online here.

Programming

- Grus (2015)
- Géron (2017), available online here.

Data Science

- MacKay (2003), a wonderful book on the connection between statistical inference and information theory. Avilable online here.
- Sutton and Barto (2018), available online here.
- Goodfellow, Bengio, and Courville (2016), available online here.
- Chollet (2018)
- Kelleher, Namee, and D'Arcy (2015)
- Schölkopf and Smola (2002)

Popular Science Books

- Bostrom (2014) Nick Bostrom, a Swedish philosopher at Oxford University, argues that if machine brains surpass human brains in general intelligence, then this new superintelligence could replace humans as the dominant lifeform on Earth.
- Domingos (2015) Outlines five tribes of machine learning: inductive reasoning, connectionism, evolutionary computation, Bayes' theorem and analogical modelling. The author explains these tribes to the reader by relating these to more familiar concepts of logic, connections made in the

brain, natural selection, probability and similarity judgements. Throughout the book, it is suggested that each different tribe has the potential to contribute to a unifying ”master algorithm”.
- O'Neil (2016) O'Neil, a mathematician and formeer Wall Street quant, analyses how the use of big data and algorithms in a variety of fields, including insurance, advertising, education, and policing, can lead to decisions that harm the poor, reinforce racism, and amplify inequality.
- Stephens-Davidowitz (2017) Ispired by Google Trends, former Google data scientist Seth Stephens-Davidowitz reveals what can be inferred about human desires, beliefs and prejudices from analyzing the vast logs of anonymous Google searches. A fascinating, if sobering, account.
- Fry (2018). An excellent exposition of the opportunities and dangers of data mining and machine learning in modern life, displayed across the chapters Power, Data, Justice, Medicine, Cars, Crime and Art. Somewhat more optimistic (balanced?) than (O'Neil, 2016).
- Harari (2018) A dismal look into the technological future by Silicon Valley's favorite philosopher.

What is Data Science?

- Bühlmann and Stuart (2016). A concise take on the role of math and stats within the emerging discipline of data science centering on models, high dimensionality and heterogeneity.
- Donoho (2017). Based on a presentation at the John Tukey 100th Birthday Celebration held in Princeton 2015, this overview traces the origins of the discipline, highlighting the role of statistics in the genesis of data science.
- Carmichael and Marron (2018)
- Mazzocchi (2015) A thoughtful discussion of Anderson's 'end of theory' proposition for data science, providing some epistemological background.

Chapter 4

- Bayes' theorem:
  - Efron (2013) : On the occasion of the 250th anniversary of Bayes' rule, eminent statistician Bradley Efron gives a very readable account of the dispute between Bayesians and frequentists delivered as the 85th Gibbs lecture at the 2012 Joint Mathematics Meeting.
  - Efron (2013) An executive summary of (Efron, 2013a).
  - McGrayne (2012), a popular science book on the history and real-world impact of Bayes' theorem
- Breast cancer screening:
  - Hoffrage and Gigerenzer (1998): How medical professionals can be taught to perform the calculations required to apply Bayes' rule.
  - Kerlikowske et al. (1996), Kerlikowske et al. (1996), A study determining the statistical parameters of mammography screening tests.

Chapter 8

- Tree-based methods
  - More details on optimal pruning of decision trees can be found in Breiman et al. (1984) (Chapter 10) and Ripley (1996) (Chapter 7).

Chapter 9

- Linear classifiers
  - A precursor of neural networks and machine learning in general was the pioneering work of the American statistician Frank Rosenblatt, whose *perceptron* was an analog device which could construct separating hyperplanes (Rosenblatt, 1958).
  - The solution of the optimization problem for finding the optimal separating hyperplane is discussed in Vapnik (2000) (Section 5.5).
- Kernel methods

– An excellent introduction into kernel-based methods and support vector machines can be found in Schölkopf and Smola (2002).

References

Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press (cit. on p. 1).

Breiman, Leo et al. (1984). *Classification and Regression Trees*. New York: Wadsworth (cit. on p. 2).

Bühlmann, Peter and Andrew M Stuart (2016). "Mathematics, Statistics and Data Science". In: *EMS Newsletter* 100, pp. 28–30 (cit. on p. 2).

Carmichael, Iain and J. S. Marron (2018). "Data science vs. statistics: two cultures?" In: *Japanese Journal of Statistics and Data Science* 1.1, pp. 117–138. DOI: 10.1007/s42081-018-0009-3 (cit. on p. 2).

Chollet, François (2018). *Deep Learning with Python*. Shelter Island: Manning Publications Co. URL: https://archive.org/details/ManningDeepLearningWithPython (cit. on p. 1).

Domingos, Pedro (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books (cit. on p. 1).

Donoho, David (2017). "50 Years of Data Science". In: *Journal of Computational and Graphical Statistics* 26.4, pp. 745–766. DOI: 10.1080/10618600.2017.1384734 (cit. on p. 2).

Efron, Bradley (2013a). "A 250-year argument: Belief, behavior, and the bootstrap". In: *Bulletin of the American Mathematical Society* 50.1, pp. 129–146. DOI: 10.1090/S0273-0979-2012-01374-5 (cit. on p. 2).

— (June 2013b). "Bayes' Theorem in the 21st Century". In: *Science* 340.6137, pp. 1177–1178. DOI: 10.1126/science.1236536 (cit. on p. 2).

Efron, Bradley and Trevor Hastie (2016). *Computer Age Statistical Inference: Algorithms, Evidence and Data Science*. Cambridge University Press. DOI: 10.1017/CBO9781316576533 (cit. on p. 1).

Freedman, David, Robert Pisani, and Roger Purves (2007). *Statistics*. 4th. New York, London: W. W. Norton & Co. (cit. on p. 1).

Fry, Hannah (2018). *Hello World: How to be Human in the Age of the Machine*. Doubleday (cit. on p. 2).

Géron, Aurélien (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts*. O'Reilly (cit. on p. 1).

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. MIT Press (cit. on p. 1).

Grus, Joel (2015). *Data Science from Scratch: First Principles with Python*. O'Reilly (cit. on p. 1).

Harari, Yuvalk Noah (2018). *21 Lessons for the 21st Century*. London: Jonathan Cape (cit. on p. 2).

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2001). *The Elements of Statistical Learning*. 2nd. Springer Series in Statistics. Springer. DOI: 10.1007/978-0-387-84858-7 (cit. on p. 1).

Hoffrage, Ulrich and Gerd Gigerenzer (May 1998). "Using Natural Frequencies to Improve Diagnostic Inferences". In: *Academic Medicine* 73.5, pp. 538–540 (cit. on p. 2).

James, Gareth et al. (2013). *An Introduction to Statistical Learning – with Applications in R*. corrected 7th printing. Springer. DOI: 10.1007/978-1-4614-7138-7 (cit. on p. 1).

Kelleher, John D., Brian Mac Namee, and Aoife D'Arcy (2015). *Fundamentals of Machine Learning for Predictive Data Analytics*. MIT Press (cit. on p. 1).

Kerlikowske, Karla et al. (1996a). "Effect of Age, Breast Density, and Family History on the Sensitivity of First Screening Mammography". In: *Journal of the American Medical Association* 276.1, pp. 33–38. DOI: 10.1001/jama.1996.03540010035027 (cit. on p. 2).

— (1996b). "Likelihood Ratios for Modern Screening Mammography: Risk of Breast Cancer Based on Age and Mammographic Interpretation". In: *Journal of the American Medical Association* 276.1, pp. 39–43. DOI: 10.1001/jama.1996.03540010041028 (cit. on p. 2).

MacKay, David (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press (cit. on p. 1).

Mazzocchi, Fulvio (2015). "Could Big Data be the end of theory in science?" In: *EMBO reports* 16.10, pp. 1250–1255. DOI: 10.15252/embr.201541001 (cit. on p. 2).

McGrayne, Sharon Bertsch (2012). *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*. Yale University Press (cit. on p. 2).

O'Neil, Cathy (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books (cit. on p. 2).

Pichler, Alois (Nov. 2018). “Selected Topics from Mathemartical Statistics”. Lecture notes for course given Winter Semester 2018/19 at TU Chemnitz, Germany (cit. on p. 1).

Ripley, Brian D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press (cit. on p. 2).

Rosenblatt, Frank (1958). “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain”. In: *Psychological Review* 65.6, pp. 386–408. DOI: 10.1037/h0042519 (cit. on p. 2).

Schölkopf, Bernhard and Alexander J. Smola (2002). *Learning with Kernels*. Cambridge, London: MIT Press (cit. on p. 1).

Stephens-Davidowitz, Seth (2017). *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are*. Dey Street Books (cit. on p. 2).

Sutton, Richard S. and Andrew G. Barto (2018). *Reinforcement Learning: An Introduction*. 2nd. Complete online draft. MIT Press (cit. on p. 1).

Vapnik, Vladimir N. (2000). *The Nature of Statistical Learning*. 2nd. Statistics for Engineering and Information Science. Springer. DOI: 10.1007/978-1-4757-3264-1 (cit. on p. 2).

Williams, David (2010). *Weighing the Odds: A Course in Probability and Statistics*. Cambridge University Press. DOI: 10.1017/CBO9781139164795 (cit. on p. 1).