Introduction to Data Science

Reading List, Winter 2023/24

Oliver Ernst

January 29, 2024

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).
- Strang (2019)

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.
- Spiegelhalter (2019)
- Diaconis and Skyrms (2018): A wonderful and very accessible tour d'horizon of the foundational concepts of probability theory.

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)
- Shalev-Shwartz and Ben-David (2014), available online here.

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 former 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) Inspired 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 3

- Allen (1997) gives an easygoing and intuitive overview of linear regression methods.
- Mood, Graybill, and Boes (1974) in Chapter X gives a detailed exposition of hypothesis tests associated with linear regression models.
- Golub and Van Loan (2013) gives an encyclopaedic account of numerical linear algebra in theory and practice, including Cholesky and QR factorization, the SVD, least squares and generalizations.
- Lewis, Lakshmivarahan, and Dhall (2006) is a book on data assimilation and contains a very thorough and intuitive exposition of least squares, both from a purely deterministic and a statistical perspective.

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 5

- Cross validation is also discussed also in the ESL book (Hastie, Tibshirani, and Friedman, 2001, Section 7.2).
- Another popular method is known as *generalized cross validation (GCV)* Golub, Heath, and Wahba (1979).
- The Bootstrap was invented by Bradley Efron in the late 1970s (Efron, 1979)
- A nice introduction to the Bootstrap can be found in Efron (2013).

Chapter 6

- Model comparison:
  - Mallows' C_p statistic: introduced in 1964 by the English statistician Colin Lingwood Mallows. The original references as well as a modern statistical treatment can be found in Gilmour, 1996.
  - Akaike Information Criterion: first published by the Japanese statistician Hirotogu Akaike in 1969 Akaike, 1969 (cf. also
- Partial Least Squares:
  - Eldén (2004); Björck (2014) and the references therein give an account of the theoretical and algorithmic state of the art in PLS.
  - Mehmood and Ahmed (2016) gives an impression of current, in particular high-dimensional applications of PLS.

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).
- Boosting
  - A seminal reference to boosting methods is the paper Freund and Schapire (1997), where the *AdaBoost.M1* algorithm is introduced. See also the survey paper Friedman, Hastie, and Tibshirani (2000).
  - A comprehensive monograph on boosting methods is Schapire and Freund (2012).
  - A brief introduction to *Gradient Boosting* can be found in Hastie, Tibshirani, and Friedman, 2001, Section 10.1.
  - A more recent but all the more successful variation of gradient boosting is known as XGBoost Chen and Guestrin, 2016.

Chapter 9

- A comprehensive presentation of principal components analysis can be found in the book Jolliffe, 2002.

References

Akaike, Hirotugo (1969). "Fitting autoregressive models for prediction". In: *Annals of the Institute of Statistical Mathematics* 21, pp. 243–247. DOI: 10.1007/BF02532251 (cit. on p. 3).

Allen, Michael Patrick (1997). *Understanding Regression Analysis*. New Yor and London: Plenum Press. DOI: 10.1007/b102242 (cit. on p. 2).

Björck, Åke (2014). "Stability of two Direct Methods for Bidiagonalization and Partial Least Squares". In: *SIAM J. Matrix Anal. Appl.* 35.1, pp. 279–291. DOI: 10.1137/120895639 (cit. on p. 3).

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. 3).

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).
Chen, Tianqi and Carlos Guestrin (2016). *XGBoost: A Scalable Tree Boosting System*. arXiv: 1603.02754v3 [cs.LG] (cit. on p. 3).
Chollet, François (2018). *Deep Learning with Python*. Shelter Island: Manning Publications Co. URL: https://archive.org/details/ManningDeepLearningWithPython (cit. on p. 1).
Diaconis, Persi and Brian Skyrms (2018). *Ten Great Ideas About Chance*. Princeton, NJ: Princeton University Press (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. 2).
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 (1979). “Bootstrap Methods: Another Look at the Jackknife”. In: *Annals of Statistics* 7.1, pp. 1–26. DOI: 10.1214/aos/1176344552 (cit. on p. 3).
— (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 pp. 2, 3).
— (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).
Eldén, Lars (2004). “Partial least-squares vs. Lanczos bidiagonalization—I: Analysis of a projection method for multiple regression”. In: *Computational Statistics nad Data Analysis* 46, pp. 11–31. DOI: 10.1016/S0167-9473(03)00138-5 (cit. on p. 3).
Freedman, David, Robert Pisani, and Roger Purves (2007). *Statistics*. 4th. New York, London: W. W. Norton & Co. (cit. on p. 1).
Freund, Yoav and Robert E. Schapire (1997). “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting”. In: *Journal of Computer and System Sciences* 55, pp. 119–139. DOI: https://doi.org/10.1006/jcss.1997.1504 (cit. on p. 3).
Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2000). “Additive Logistic Regression: A Statistical View of Boosting”. In: *Annals of Statistics* 28.2, pp. 337–407. DOI: 10.1214/aos/1016218223 (cit. on p. 3).
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).
Gilmour, Steven G. (1996). “The Interpretation of Mallows’s C_p Statistic”. In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 45.1, pp. 49–56. DOI: 10.2307/2348411 (cit. on p. 3).
Golub, Gene H., Michael Heath, and Grace Wahba (1979). “Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter”. In: *Technometrics* 21.2, pp. 215–223. DOI: 10.1080/00401706.1979.10489751 (cit. on p. 3).
Golub, Gene H. and Charles F. Van Loan (2013). *Matrix Computations*. 4th. Johns Hopkins University Press (cit. on p. 2).
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 pp. 1, 3).
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).
Jolliffe, I. T. (2002). *Principal Component Analysis*. 2nd. Springer Series in Statistics. Springer. DOI: 10.1007/b98835.
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).

Lewis, John M., S. Lakshmivarahan, and Sudarshan Dhall (2006). *Dynamic Data Assimilation: A Least Squares Approach*. Cambridge University Press (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).

Mehmood, Tahir and Bilal Ahmed (2016). “The diversity in the applications of partial least squares: an overview”. In: *Chemometrics* 30, pp. 4–17. DOI: 10.1002/cem.2762 (cit. on p. 3).

Mood, Alexander McFarlane, Franklin A. Graybill, and Duane C. Boes (1974). *Introduction to the Theory of Statistics*. 3rd. McGraw-Hill (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 Mathematical 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. 3).

Schapire, Robert E. and Yoav Freund (2012). *Boosting: Foundations and Algorithms*. MIT Press (cit. on p. 3).

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

Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press (cit. on p. 1).

Spiegelhalter, David (2019). *The Art of Statistics: Learning from Data*. Penguin Books (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).

Strang, Gilbert (2019). *Linear Algebra and Learning from Data*. Wellesley-Cambridge Press (cit. on p. 1).

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

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