PHILOSOPHY OF DATA SCIENCE

Fall 2022 Syllabus

Instructor: Ignacio Ojea Quintana.
Meeting Time: Wednesdays 10am to 12pm
Location: Ludwigstr. 31 - 021 (floor plan)
Office Hours: Wednesdays 8am-10am
Contact: ignacio.ojea@lrz.uni-muenchen.de
Uni Link: Link
Google Drive Link: Link

Course Description

The purpose of this course is to provide a philosophically critical introduction to Data Science. We will cover the most basic and rudimentary techniques, while we also inquire about central philosophical questions:

- What is Data Science? Does it constitute a paradigm shift or is it just statistics on (computational) steroids?
- How does it relate to the old problem of induction?
- Does it have inherent problems (p-hacking) which also scale badly sociologically (publication bias)?
- Is *interpreting* data good science? Is data *given*? A study of exploratory data analysis with real data.
- What is data bias and what can we do about it? The COMPASS case.
- What is a good *scientific explanation* in Data Science? Are *good explanations* in tension with *good predictions*?
- How powerful are Neural Networks? The Universal Approximator Theorem.

For the technical content we will take a moderate approach. The idea is for students to leave the course with a basic understanding of the theoretical material, as well as being able to use programming to do some basic data analysis themselves. We will use a textbook but mostly custom notes in the form of jupyter notebooks (see GDrive) to cover the fundamentals: probabilities, basic statistics, hypothesis testing, regression models, decision trees, and neural networks.

Prerequisites

This class is planned so that students work on three different skills: philosophy, mathematics, and coding. For this reason it is expected of the students that are competent in at least two of them. Nevertheless, the coding as well as much of the mathematics will be presented from scratch.

Required Texts

- All texts will be made available digitally.
- The textbook we will be using is Joel Grus' *Data Science from Scratch: First Principles with Python* (2nd Ed), GitHub, [3].

Grading

There will be two components to the evaluation, namely writing a philosophical essay and performing a data analysis by yourself using the techniques presented.

In order to help students build up to the final project, optional exercises will be uploaded throughout the course.

TENTATIVE COURSE SCHEDULE

Please note that all readings and due dates are subject to change.

Please do the readings before attending to class.

Note: Dates will be corrected according to the academic calendar. This course is about twelve weeks long, factoring in public holidays.

Week 1: *Introduction: Is Data Science a paradigm change?* **Wed 19/10**

Required:

- Chapter 1 of the textbook, *Introduction*.
- Refresh yourself of Kuhn´s *Structure*, [9].

Optional:

- Scientific Revolutions (SEP) Link

Week 2: *Probability and Statistics Fundamentals* **Wed 26/10**

Required:

- Chapter 5 of the textbook, *Statistics*.
- Chapter 6 of the textbook, *Probability*.

Optional:

- See jupyter notebook on GDrive.

Week 3: *The Problem of Induction* **Wed 2/11**

Required:

- G. Harman & S. Kulkarni, *Reliable Reasoning*, Chapter 1, [4].
- P. Lipton, *Inference to the Best Explanation*, Chapter 1, [10].

Optional:

- See jupyter notebook on GDrive.

Week 4: *Hypothesis Testing and Statistical Inference* **Wed 9/11**

Required:

- Chapter 7 of the textbook, *Hypothesis and Inference*.

Optional:

- J. Cohen, "The earth is round (p<.05)", [1].
- P. Meehl, "Theory-testing in psychology and physics: A methodological paradox", [11].
- P. Meehl, "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology", [12].

Week 5: *P-Hacking and Publication Bias* **Wed 16/11**

Required:

- J. Ioannidis, "Why Most Published Research Findings Are False", [6].
- R. Nuzzo, "Statistical Errors: P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume", [13].

Optional:

- M. Kotzen, "Multiple Studies and Evidential Defeat", [8].

Week 6: *Working with Actual Data: Preprocessing and Exploratory Analysis* **Wed 23/11**

Required:

- Chapter 9 of the textbook, *Getting Data*.
- Chapter 10 of the textbook, *Working with Data*.
- J. Tukey, *Exploratory Data Analysis* (Preface and Section 1A), [14].

Optional:

- See jupyter notebook on GDrive.

Week 7: *Regression Models* **Wed 30/11**

Required:

- Chapter 14 of the textbook, *Simple Linear Regressions*.
- Chapter 15 of the textbook, *Multiple Regressions*.

Optional:

- Chapter 16 of the textbook, *Logistic Regression*.
- R. M. Dawes, "The robust beauty of improper linear models in decision making", [2].

Week 8: *Bias and Accuracy: The COMPASS Case* **Wed 7/12**

Required:

- Pro-Publica article on Machine Bias.

- J. Kleinberg et al., "Inherent Trade-Offs in the Fair Determination of Risk Scores", [7].

Optional:

- See jupyter notebook on GDrive.

Week 9: *Decision Trees and Forests* **Wed 14/12**

Required:

- Chapter 17 of the textbook, *Decision Trees*.

Optional:

- See jupyter notebook on GDrive.

Week 10: *Explanations in Data Science* **Wed 21/12**

Required:

- Chapter 11 of textbook, *Machine Learning* (on over-fitting).
- Explainable AI.

Optional:

- SEP Link.
- T. Yarkoni and J. Westfall, "Choosing prediction over explanation in psychology: Lessons from machine learning", [15].

Week 11: *Artificial Neural Networks* **Wed 11/1**

Required:

- Chapter 18 of the textbook, *Neural Networks*.

Optional:

- See jupyter notebook on GDrive.

Week 12: *Neural Networks as Universal Approximators (but open to change)* **Wed 18/1**

Required:

- K. Hornik, M. Stinchcombe, H. White, Halbert, "Multilayer Feedforward Networks are Universal Approximators", [5].

Optional:

- A visual proof of the theorem, by M. Nielsen.

Examination Week:

Final Project Due
Final Paper Due

References

[1] Jacob Cohen. The earth is round (p ¡ .05. *American Psychologist*, pages 997–1003, 1994.

[2] Robyn M. Dawes. The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7):571–582, 1979.

[3] J. Grus. *Data Science from Scratch: First Principles with Python*. O'Reilly Media, 2019.

[4] G. Harman and S. Kulkarni. *Reliable Reasoning: Induction and Statistical Learning Theory*. Jean Nicod Lectures. MIT Press, 2012.

[5] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[6] John Ioannidis. Why most published research findings are false. *PLoS medicine*, 2:e124, 09 2005.

[7] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807, 2016.

[8] Matthew Kotzen. Multiple studies and evidential defeat. *Noûs*, 47(1):154–180, 2013.

[9] Thomas S. Kuhn. *The structure of scientific revolutions*. University of Chicago Press, Chicago, 1970.

[10] Peter Lipton. *Inference to the Best Explanation*. London and New York: Routledge, 1991.

[11] Paul E. Meehl. Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34(2):103–115, 1967.

[12] Paul E. Meehl. Theoretical risks and tabular asterisks: Sir karl, sir ronald, and the slow progress of soft psychology. *ournal of Consulting and Clinical Psychology*, 46(4):806–834, 1992.

[13] Regina Nuzzo. Statistical errors: p values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. *Nature*, 130, 01 2014.

[14] J.W. Tukey. *Exploratory Data Analysis*. Number v. 2 in Addison-Wesley series in behavioral science. Addison-Wesley Publishing Company, 1977.

[15] Tal Yarkoni and Jacob Westfall. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6):1100–1122, 2017. PMID: 28841086.