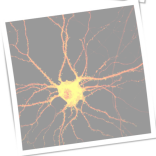
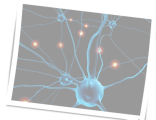
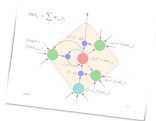
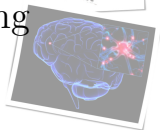


A quick introduction to machine learning

Spyros Samothrakis
Senior Lecturer, IADS
University of Essex
MiSoC

June 22, 2022



WELCOME/COURSE CONTENTS

- ▶ What will this course cover?
 - ▶ Day 1: An intro to machine learning (ML)
 - ▶ Day 1: ML labs
 - ▶ Day 2: An intro to causal inference
 - ▶ Day 2: ML and causal inference labs
- ▶ Textbooks?
 - ▶ Mitchell, T. M. (1997). Machine learning.¹
 - ▶ Bishop, C. M. (2006). Pattern recognition and machine learning. springer.²
 - ▶ Wasserman, L. (2013). All of statistics: a concise course in statistical inference. Springer Science & Business Media.³

¹<http://www.cs.cmu.edu/~tom/mlbook.html>

²<https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/>

³<http://www.stat.cmu.edu/~larry/all-of-statistics/index.html>

BETTER SCIENCE THROUGH DATA

Hey, Tony, Stewart Tansley, and Kristin M. Tolle. “Jim Gray on eScience: a transformed scientific method.” (2009).⁴

- ▶ Thousand years ago: empirical branch
 - ▶ You observed stuff and you wrote down about it
- ▶ Last few hundred years: theoretical branch
 - ▶ Equations of gravity, equations of electromagnetism
- ▶ Last few decades: computational branch
 - ▶ Modelling at the micro level, observing at the macro level
- ▶ Today: data exploration
 - ▶ Let machines create models using vast amounts of data

⁴<http://language.log.ldc.upenn.edu/myl/JimGrayOnE-Science.pdf>

BETTER BUSINESS THROUGH DATA

- ▶ There was a report by Mckinsey

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.⁵

- ▶ Urges everyone to monetise “Big Data”
- ▶ Use the data provided within your organisation to gain insights
- ▶ Has some numbers as to how much this is worth
- ▶ Proposes a number of methods, most of them associated with machine learning and databases

⁵<http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation>

WHY IS IT POPULAR NOW?

- ▶ **Algorithms + data + tools**
- ▶ Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science, 16(3), 199-231.⁶
- ▶ Anderson, P. W. (1972). More is different. Science, 177(4047), 393-396.⁷
- ▶ Pedregosa, et.al. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.⁸

⁶http://projecteuclid.org/download/pdf_1/euclid.ss/1009213726%20

⁷[https:](https://www.tkm.kit.edu/downloads/TKM1_2011_more_is_different_PWA.pdf)

[//www.tkm.kit.edu/downloads/TKM1_2011_more_is_different_PWA.pdf](https://www.tkm.kit.edu/downloads/TKM1_2011_more_is_different_PWA.pdf)

⁸[https:](https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf)

[//www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf](https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf)

SO THIS COURSE COVERS TOOLS

- ▶ ML theory
 - ▶ *Supervised learning, (non)-linear regression*
 - ▶ Understanding basic modelling
 - ▶ Confirming your model is sane
 - ▶ Tuning your model
 - ▶ **All within a very applied setting**
- ▶ Tools
 - ▶ Numpy
 - ▶ Scikit-learn

WHAT IS SUPERVISED LEARNING?

- ▶ Imagine someone gives you a group of smokers
 - ▶ And asks the question – what is there life expectancy?
- ▶ **Completely made up imaginary data**

SOME ABSTRACTION

- ▶ We are given inputs $x_0, x_1 \dots x_n$ and we are looking to predict y
- ▶ Let's plot!