UNIVERSITY OF COPENHAGEN

Standard Models for Explainable ML

Tim Mondorf

Student, MSc in Computer Science, Department of Computer Science (DIKU)

D3A, Nyborg Strand, 1-2 February 2024

Contact Information:
Phone: +45 42765599
Email: `cjk681@alumni.ku.dk`

Abstract

Explainable Machine Learning (ML) is a property not just of models but of ML's scientific identity. I analyze this scientific identity through a case study of the work of the ACM task force on Data Science and of four seminal articles including Vaswani et. al. 'Attention is all you Need'. I conclude that ML's scientific identity is disparate and reflects a preference for probabilistic and heuristic rather than general statements. I challenge this preference and discuss a number of ways to potentially render ML less disparate and more explainable. I amend the Standard Model for Machine Learning, originally proposed by Hu et al. in 2022, in order to facilitate a structured discussion of accuracy, model desiderata, context and ethical and legal requirements across all ML model categories. I propose a stronger articulation of deductive versus inductive processes when models learn from data. I propose more consistent symbolic generalizations as well as more intuitive methods for teaching ML. These initiatives will promote Explainable ML.

About me

- MSc economics, 1994
- External lecturer at KU and CBS
- IT career with IBM, Dell, others
- BSc in Machine Learning and Data Science, DIKU, 2024
- Student, MSc Computer Science, DIKU 2024
- This presentation is my project on Explainable ML (Mondorf, 2023, p. 1, [4])
- It adopts the perspective of a newcomer to ML with a background in another science, i.e., economics

What is a standard model and why do we need it?

- One model that comprises all ML models as special cases
- Why?
  - Words, definitions, notation matter in innovation
  - Confirmed by literature such as Hadamard (Hadamard, 1945, chap. 7, p. 84, [1]) and Kuhn (Kuhn, 2012, 1969 postscript, p. 182 [2])
  - Less support for the idea that notation etc. must be standardized
  - The recent focus on Explainable ML makes standard models relevant
    - *"What if ... we would only have to learn and explain one model"*
    - *"What if ... all new models could be explained with reference to the standard model*

Hu and Xing Toward a Standard Model of ML

- Hu and Xing Toward a Standard Model of Machine Learning, Harvard Data Science Review (2022) (Hu, 2022, p. 1, [6])
  - Based on Lagrangian convex optimization
  - Minimize (-entropy + error + penalty term)
  - Subject to 'soft' secondary conditions on parameters
  - Secondary condition can be increased but this will increase the penalty term
- Will not here prove the two most important points:
  - all existing ML-models fit as special cases of the Hu and Xing-model
  - the model is non-trivial
- Focus on what the standard model could mean for Explainable ML

Adapting Hu and Xing's Standard Model

- Mondorf (Mondorf, 2023, p. 1, [4])
  - Optimize a function of (error + model desiderata + soft context)
  - Subject to hard secondary condition
  - Model desiderata: entropy, sparsity, smoothness etc.
  - Soft context: Relevant information outside of training data
  - Secondary condition can reflect ethical and legal restrictions
  - Such as examples from training data that must not be reproduced, blacklist words in language models et

Standard models = general statements = Explainable ML

- Make transparent the trade-off between accuracy and model desiderata
- Can we make the general statement that for two models with the same accuracy, we prefer
  - The model that is right when it comes to surprising predictions (entropy)?
  - The simple model (sparsity, Occam's Razor)?
  - The model that avoids drastic updates to weights (smoothness)?
- Economics derives strong intuition from optimization under secondary conditions
- What intuition can ML derive? What choices were made to maintain accuracy while respecting ethical and legal requirements?
- ML as a science is often heuristic or probabilistic
- Deductive-nomological model of scientific explanation (Woodward, 2021, sect. 2, [5])
- Standard models can facilitate general statements
- General statements can facilitate Explainable Machine Learning

The death of the hyper-parameter

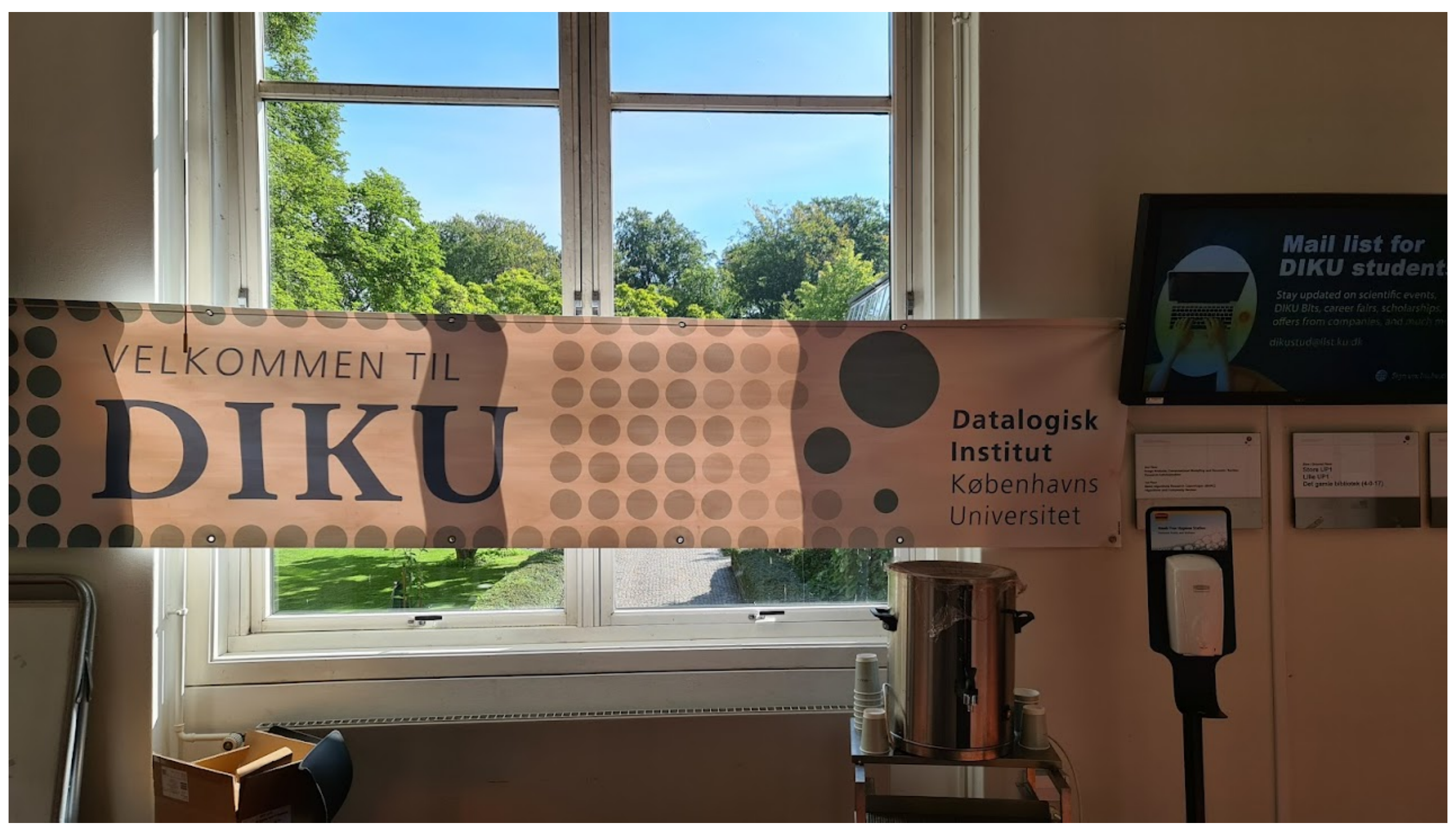
- Contemporary ML articles such as Vaswani 'Attention is all you need' (Vaswani, 2017, p.1, [7]) often have
  - A general research theme
  - A model that learns weights from data in a structured and well-documented process
  - A number of "hyper-parameters" such as activation functions, network architecture, supporting mathematical functions that are selected in a process that is often unstructured, heuristic and not replicable
- Selection of 'hyper-parameters' can also be automated (Gesmundo)
- A standard model would clearly separate parameters into
  - Exogeneous variables: everything that does NOT change during the experiment
  - Endogeneous variables: everything that is subject to change through ANY process during the experiment
  - ... similar to the intuition of economic models

Explainable ML is a property of ML's scientific identity

- A task force under the Association for Computing Machinery (ACM) worked from 2017 to 2021 on defining data science and ML (ACM, 2021, p.1, [3])
- Emphasized multidisciplinarity (computer science, statistics, mathematics, application domains)
- The word 'broad' appears 32 times over the 132 pages of the final report
- The official curricula for Danish ML programmes have the same 'broad' approach
- Justified from a ressource perspective, but commits the ML sin of "too rich a hypothesis space = overfitting"

As ML's scientific identity matures, will Explainable ML become easier?

- Proposal:
  - 'Computer science first': Define a clear hierachy of sciences for ML
  - ML is a proper subset of computer science.
  - The subset that analyzes uncertainty and ambiguity better than traditional computer science



Link to project

https://github.com/TimMondorf/ Does-Machine-Learning-need-a-Theory-of-Everything

References

[1] Jaques Hadamard. The mathematician's mind. Princeton University Press, 1945.

[2] Thomas S. Kuhn. The structure of scientific revolutions - 50th anniversary edition. The University of Chicago Press, 2012.

[3] Andrea Danyluk Paul Leidig. Computing competencies for undergraduate data science curricula acm data science task force. Association for Computing Machinery, 2021.

[4] Tim Mondorf. Does machine learning need a theory of everything? BSc-project, DIKU, 2023.

[5] James Woodward Lauren Ross. Scientific explanation. Stanford Encyclopedia of Philosophy, 2021.

[6] Zhiting Hu Eric Xing. Toward a standard model for machine learning. Harvard Data Science Review, 2022.

[7] Ashish Vaswani Noam Shazeer Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Łukasz Kaiser Illia Polosukhin. Attention is all you need. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017.

Acknowledgements

Thanks to professor Henrik Kragh Sørensen, Ph.d., Department of Science Education (IND) for supervising the project, and to all teachers at DIKU.