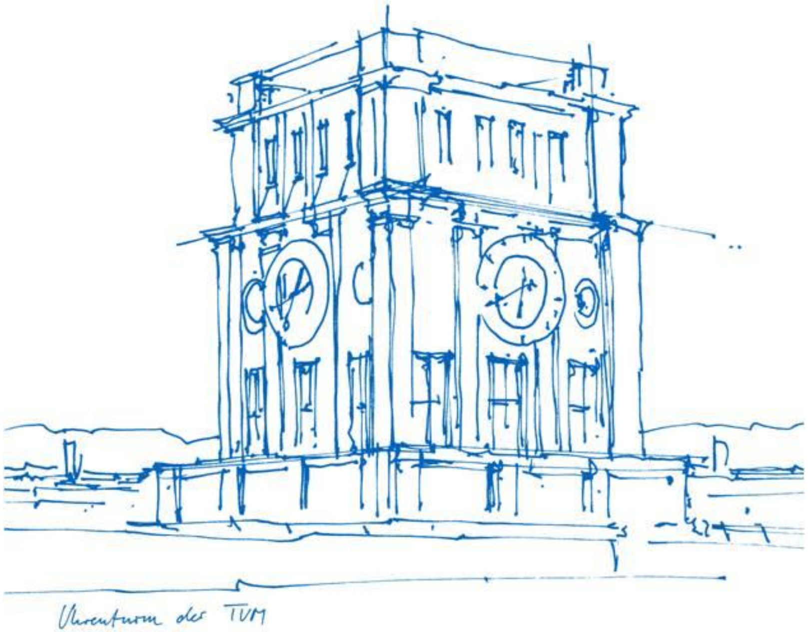
Opening the Black Box: Trends in explainable AI

General Information

Amjad Ibrahim

Prof. Dr. Alexander Pretschner



Technische Universität München

Fakultät für Informatik

Informatics 4 - Chair of Software and Systems Engineering

Assigned Topics

1. Overview of xAI: Requirements, stakeholders, concepts, taxonomies, Human-friendly Explanations
2. Global explanation methods
3. Model-Agnostic Methods
4. Counterfactual-based explanation
5. PROTOTYPES AND CRITICISMS
6. Tools and case-studies in different domains (medicine, law, finance, security)
7. Deep learning explanation (Multi-Layer Neural Networks)
8. Convolutional Neural Networks and Recurrent Neural Networks explanation
9. Explanations for AI Security: XAI and Adversarial Machine Learning
10. Explanations for Ensembles and Multiple Classifier Systems
11. Evaluation of xAI
12. Explanation by Feature relevance

Basic Literature (students should expand)

General (for all students): especially the general tree page 19

- Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI, Information Fusion, Volume 58, 2020, Pages 82-115, https://doi.org/10.1016/j.inffus.2019.12.012.

Basic Literature (students should expand)

Topic#1 : (Overview of xAI: Requirements, stakeholders, concepts, taxonomies, Human-friendly Explanations)

- Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.
- Miller, Tim. “Explanation in Artificial Intelligence: Insights from the Social Sciences.” Artificial Intelligence 267 (2019): 1–38. Crossref. Web.
- Brent Mittelstadt, Chris Russell, and Sandra Wachter, ‘Explaining explanations in ai,’ in Proceedings of the conference on fairness, accountability, and transparency, pp. 279–288. ACM (2019)

- Dhurandhar, Amit, et al. "Explanations based on the missing: Towards contrastive explanations with pertinent negatives." *Advances in Neural Information Processing Systems*. 2018.

Basic Literature (students should expand)

Topic#2 : (Global explanation methods)

- R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, ‘‘Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission,’’ in Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2015, pp. 1721–1730.
- B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, ‘‘Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model,’’ Ann. Appl. Statist., vol. 9, no. 3, pp. 1350–1371, 2015
- A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, ‘‘Synthesizing the preferred inputs for neurons in neural networks via deep generator networks,’’ in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2016, pp. 3387–3395.
- D. Erhan, A. Courville, and Y. Bengio, ‘‘Understanding representations learned in deep architectures,’’ Dept. d’Informatique Recherche Operationnelle, Univ. Montreal, Montreal, QC, Canada, Tech. Rep. 1355, 2010.

Basic Literature (students should expand)

Topic#3 : (Model-Agnostic Methods)

- Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. Chapter 5 https://christophm.github.io/interpretable-ml-book/agnostic.html
- M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier,'' in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 1135–1144

- O. Bastani, C. Kim, and H. Bastani. (2017). ''Interpretability via model extraction.'' [Online]. Available: https://arxiv.org/abs/1706.09773
- J. J. Thiagarajan, B. Kailkhura, P. Sattigeri, and K. N. Ramamurthy. (2016). ''TreeView: Peeking into deep neural networks via feature-space partitioning.'' [Online]. Available: https://arxiv.org/abs/1611.07429
- D. P. Green and H. L. Kern, ''Modeling heterogeneous treatment effects in large-scale experiments using Bayesian additive regression trees,'' in Proc. Annu. Summer Meeting Soc. Political Methodol., 2010, pp. 1–40.

Basic Literature (students should expand)

Topic#4: (Counterfactual-based explanation)

- S. Wachter, B. Mittelstadt, and C. Russell. (2017). ‘‘Counterfactual explanations without opening the black box: Automated decisions and the GDPR.’’ [Online]. Available: https://arxiv.org/abs/1711.00399
- X. Yuan, P. He, Q. Zhu, and X. Li. (2017). ‘‘Adversarial examples: Attacks and defenses for deep learning.’’ [Online]. Available: https://arxiv.org/abs/1712.07107
- R. M. J. Byrne, Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, 2019, pp. 6276–6282
- S. Sharma, J. Henderson, J. Ghosh, Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models, arXiv preprint arXiv:1905.07857 (2019).
- Bertossi, Leopoldo. "An ASP-Based Approach to Counterfactual Explanations for Classification." *arXiv preprint arXiv:2004.13237* (2020).

Basic Literature (students should expand)

Topic#5: (Prototypes and Criticisms)

- J. Bien and R. Tibshirani, ‘‘Prototype selection for interpretable classification,’’ Ann. Appl. Statist., vol. 5, no. 4, pp. 2403–2424, 2011.
- B. Kim, C. Rudin, and J. A. Shah, ‘‘The Bayesian case model: A generative approach for case-based reasoning and prototype classification,’’ in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 1952–1960.
- K. S. Gurumoorthy, A. Dhurandhar, and G. Cecchi. (2017). ‘‘ProtoDash: Fast interpretable prototype selection.’’ [Online]. Available: https://arxiv.org/abs/1707.01212
- B. Kim, R. Khanna, and O. O. Koyejo, ‘‘Examples are not enough, learn to criticize! criticism for interpretability,’’ in Proc. 29th Conf. Neural Inf. Process. Syst. (NIPS), 2016, pp. 2280–2288.
- Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. Chapter 6.3 https://christophm.github.io/interpretable-ml-book/proto.html

Basic Literature (students should expand)

Topic#6 (Tools and case-studies in different domains (medicine, law, finance, security)

- Arya, Vijay, et al. "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques." *arXiv preprint arXiv:1909.03012* (2019).
- M. Bojarski et al. (2016). ‘‘End to end learning for self-driving cars.’’ [Online]. Available: https://arxiv.org/abs/1604.07316
- J. Haspiel et al. (2018). Explanations and Expectations: Trust Building in Automated Vehicles, deepblue.lib.umich.edu. [Online]. Available: https://doi.org/10.1145/3173386.3177057
- A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell. (2017). ‘‘What do we need to build explainable AI systems for the medical domain?’’ [Online]. Available: https://arxiv.org/abs/1712.09923
- Z. Che , S. Purushotham, R. Khemani, and Y. Liu, ‘‘Interpretable deep models for ICU outcome prediction,’’ in Proc. AMIA Annu. Symp., 2017, pp. 371–380.
- S. Tan, R. Caruana, G. Hooker, and Y. Lou. (2018). ‘‘Detecting bias in black-box models using transparent model distillation.’’ [Online]. Available: https://arxiv.org/abs/1710.06169
- Lucic, Ana, Hinda Haned, and Maarten de Rijke. "Why does my model fail? contrastive local explanations for retail forecasting." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020.

Basic Literature (students should expand)

Topic#7 (Deep learning explanation (Multi-Layer Neural Networks)

- J. J. Thiagarajan, B. Kailkhura, P. Sattigeri, K. N. Ramamurthy, Treeview: Peeking into deep neural networks via feature-space partitioning (2016). arXiv:1611.07429.
- G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network (2015). arXiv: 1503.02531
- Z. Che, S. Purushotham, R. Khemani, Y. Liu, Interpretable deep models for ICU outcome prediction, in: AMIA Annual Symposium Proceedings, Vol. 2016, American Medical Informatics Association, 2016, p. 371.
- J. R. Zilke, E. L. Menc´ıa, F. Janssen, Deepred–rule extraction from deep neural networks, in: International Conference on Discovery Science, Springer, 2016, pp. 457–473.
- M. Sato, H. Tsukimoto, Rule extraction from neural networks via decision tree induction, in: IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222), Vol. 3, IEEE, 2001, pp. 1870–1875.
- Quan-shi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey. Frontiers of Information Technology & Electronic Engineering, 19(1):27–39, 2018.

Basic Literature (students should expand)

Topic#8 (Convolutional Neural Networks and Recurrent Neural Networks explanation)

- Harradon, Michael, Jeff Druce, and Brian Ruttenberg. "Causal learning and explanation of deep neural networks via autoencoded activations." arXiv preprint arXiv:1802.00541 (2018).
- Karpathy, Andrej, Justin Johnson, and Li Fei-Fei. "Visualizing and understanding recurrent networks." arXiv preprint arXiv:1506.02078 (2015)
- S. Bach, A. Binder, K.-R. Muller, W. Samek, Controlling explanatory heatmap resolution and ¨ semantics via decomposition depth, in: IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 2271–2275
- Q. Zhang, Y. Yang, H. Ma, Y. N. Wu, Interpreting CNNs via decision trees, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6261–6270.
- J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net (2014). arXiv:1412.6806.
- J. Clos, N. Wiratunga, S. Massie, Towards explainable text classification by jointly learning lexicon and modifier terms, in: IJCAI-17 Workshop on Explainable AI (XAI), 2017, p. 19.

Basic Literature (students should expand)

Topic#9 (Explanations for AI Security: XAI and Adversarial Machine Learning)

- S. J. Oh, B. Schiele, M. Fritz, Towards reverse-engineering black-box neural networks, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, 2019, pp. 121–144.
- I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples (2014). arXiv:1412.6572.
- I. J. Goodfellow, N. Papernot, P. D. McDaniel, cleverhans v0.1: an adversarial machine learning library (2016). arXiv:1610.00768
- B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndi ˇ c, P. Laskov, G. Giacinto, F. Roli, Evasion ´ attacks against machine learning at test time, in: Proceedings of the 2013th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part III, ECMLPKDD'13, 2013, pp. 387–402.
- Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, Y. Zheng, Recent progress on generative adversarial networks (gans): A survey, IEEE Access 7 (2019) 36322–36333.

Basic Literature (students should expand)

Topic#10 (Explanations for Ensembles and Multiple Classifier Systems)

- N. F. Rajani, R. J. Mooney, Ensembling visual explanations, in: Explainable and Interpretable Models in Computer Vision and Machine Learning, Springer, 2018, pp. 155–172.
- G. Tolomei, F. Silvestri, A. Haines, M. Lalmas, Interpretable predictions of tree-based ensembles via actionable feature tweaking, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017, pp. 465–474
- H. Deng, Interpreting tree ensembles with intrees (2014). arXiv:1408.5456.
- P. Domingos, Knowledge discovery via multiple models, Intelligent Data Analysis 2 (1-4) (1998) 187–202.
- S. Hara, K. Hayashi, Making tree ensembles interpretable (2016). arXiv:1606.05390.

Basic Literature (students should expand)

Topic#11 (Evaluation of xAI)

- F. Doshi-Velez and B. Kim. (2018). ‘‘Towards a rigorous science of interpretable machine learning.’’ [Online]. Available: https://arxiv.org/abs/1702.08608
- S. Mohseni and E. D. Ragan. (2018). ‘‘A human-grounded evaluation benchmark for local explanations of machine learning.’’ [Online]. Available: https://arxiv.org/abs/1801.05075
- A. Backhaus and U. Seiffert, ‘‘Quantitative measurements of model interpretability for the analysis of spectral data,’’ in Proc. IEEE Symp. Comput. Intell. Data Mining (CIDM), 2013, pp. 18–25.
- F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. Wallach. (2018). ‘‘Manipulating and measuring model interpretability.’’ [Online]. Available: https://arxiv.org/abs/1802.07810
- J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, B. Baesens, An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models, Decision Support Systems 51 (1) (2011) 141–154.
- S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable ai systems (2018). arXiv:arXiv:1811.11839.

Basic Literature (students should expand)

Topic#12 (Explanation by Feature relevance)

- A. Palczewska, J. Palczewski, R. M. Robinson, D. Neagu, Interpreting random forest classification models using a feature contribution method, in: Integration of Reusable Systems, Springer, 2014, pp. 193–218.
- A. Datta, S. Sen, Y. Zick, Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems, in: 2016 IEEE symposium on security and privacy (SP), IEEE, 2016, pp. 598–617.
- S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, 2017, pp. 4765–4774.
- W. Landecker, M. D. Thomure, L. M. Bettencourt, M. Mitchell, G. T. Kenyon, S. P. Brumby, Interpreting individual classifications of hierarchical networks, in: 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2013, pp. 32–38.

Finding Literature

- TUM Library
  - Informatik
  - Others…

- Online portals
  - Springer (www.springerlink.com/)
  - ACM (dl.acm.org/)
  - IEEE (ieeexplore.ieee.org/Xplore/guesthome.jsp)
  - Google Scholar (scholar.google.com)
  - Scopus (scopus.com)

Seminar Goals

- Critical reading and understanding
- Comparing
- Classification
- Writing an exposé
- Presentation skills

Task Overview

- Independent work
  - Read and understand concepts
  - Look for papers/material beyond the initial suggestions
    - E.g. Academic publication portals, TUM library etc.
    - No Wikipedia! (Except if a source is picked – discuss with the supervisor)
    - No blogs!
- Discuss with your colleagues
- Talk with your supervisor whenever required

Roadmap

- ~~Matching~~
- ~~Topic selection~~
- Literature review (talk to the supervisor by 04.11)
- First submission (27.11)
- Peer review (04.12)
- Final submission (18.01) 50 %
- Talks/Presentation (TBA) 50 %

Literature review (talk to the supervisor by 04.11)

- Prepare about 2 pages
  - Extended abstract
    - Introduction
    - Problem statement, research questions and goals
    - Short description of content of each subsection
    - Description of your own contribution/critique
  - Bibliography

Exposé (first and final submission)

- Max. 15 pages including appendix, LNCS format (advice: use Latex)
- No plagiarism!
  - blatant copy-paste, summarizing others' ideas/results without reference etc. <u>will result in immediate expulsion from the course.</u>

- Discussion of own contribution
- Complete bibliography
- Appendix, if needed

Content

- Don’t deviate from allotted topic
- Logical and contradiction-free reasoning
- Argue with proper sources
- If any contradictions in the source paper, don’t hide them.

Content

- Clear distinction between scientific facts and own logical conclusion
  - E.g. if something is “good” according to you, why?
  - Proper references

- Language
  - Easy to understand, simple (and short) sentences
  - Precise
  - Sensible titles
  - Sensible paragraphing

Content

- Tables and pictures
  - Cite sources
  - Must not be blurry
  - Large enough to be read in print
  - Must be referenced in text
  - Consistent numbering
- Bibliography
  - Must be referenced in text
  - Consistent numbering
  - Citation must include - Authors’ names, title, year of publication, venue (or publisher)

Presentation

- Ca. 20 minutes of talking
  - Clear, linear storyline.
  - Must match the exposé, but should not be a text dump
  - Possibility of discussing slides with supervisor

- Ca. 10 minutes of discussion
  - Be prepared for questions on the topic
  - Ask questions on the presented topic

Thanks!