CS6462

Probabilistic and Explainable AI

Lesson 26

Explainability in the Context of AI

by Emil Vassev

April 7, 2025

AI and Understandability

AI and ML:

- black box models:
    - non-transparent models that process high-dimensional input data in a non-linear and nested fashion to reach probabilistic decisions
    - deep neural networks (DNNs), support vector machines (SVMs), random forests (RFs), and ensemble models (EMs)
- pros:
    - automatically discover patterns and structures in large amount of data in an automated manner
    - great success in a number of different learning tasks, e.g., image recognition and natural language processing
- cons:
    - complex models with a lack of a straight-forward explanation
    - difficult (if not possible) to understand the decisions suggested by AI systems – Can we trust AI?

AI and Explainability

Explainability:

- facilitates the understanding of various aspects of an AI model
- provides notions of transparency - humans understand the inner side of the model
- provides model's insights that can be utilized by various stakeholders*:
  - *data scientists*: benefit when debugging models or when looking for ways to improve models' performance
  - *business owners*: care about the models' fit in the business strategy and purpose
  - *model-risk analysts*: challenge the models, in order to check for robustness and approve for deployment
  - *regulators*: inspect the models' reliability and the impact of models' decisions on customers
  - *consumers*: require transparency about how decisions are taken, and how they could potentially affect them

* V. Belle, I. Papantonis (2021). Principles and Practice of Explainable Machine Learning

AI and Explainability (cont.)

Definitions of Explainable AI:

Definition 1 (D. Gunning):

- Explainable AI *1)* produces more explainable models while maintaining a high level of learning performance (e.g., prediction accuracy), and 2) enables humans to understand, appropriately trust, and effectively manage the emerging generation of AI partners.

Definition 2 (A. Arrieta et al.):

- Explainable AI is a system that produces details or reasons to make its functioning clear or easy to understand.

Goals of Explainable AI:

- trustworthiness, causality, transferability, informativeness, confidence, fairness, accessibility, interactivity and privacy awareness

- D. Gunning (2017). Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA).
- A. Arrieta et al. (2019) Explainable Artificial Intelligence (xai): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. arXiv preprint arXiv:1910.10045

Perspectives of Explainability

Levels of transparency:

> "Transparency stands for a human-level understanding of the inner workings of the model." (Lipton, 2016)

- *simulatability*:
  - model is simple enough so it can be simulated by a human:
  - models that fall in this category:
    - simple and compact models
    - simple cases of otherwise complex models (e.g., neural networks with no hidden layers) could potentially fall into this category
  - insufficiency of simplicity: large number of simple rules would prohibit a human to calculate the model's decisions
- *decomposability*:
  - a model can be broken down into parts - input, parameters and computations
  - explain model's parts is easier

Z. C. Lipton (2016). The Mythos of Model Interpretability. arXiv preprint arXiv: 1606.03490

Perspectives of Explainability (cont.)

Levels of transparency (cont.):

- *algorithmic transparency*:
  - models' inner operations and the output are transparent to humans
  - requirement for a model to fall into this category: must provide insights, so a user can inspect it through a mathematical analysis
  - models that fall in this category:
    - models that classify instances based on some similarity measures
    - models with approximated solutions: complex models (e.g., neural networks) that construct an elusive loss function where the solution to the training objective is approximated
- *overall perception*:
  - decision trees, linear regressions – transparent models
  - random forests, deep learning – black box (opaque) models

Perspectives of Explainability (cont.)

Explainability evaluation criteria:*

- *comprehensibility*: the extent to which extracted representations are firmly comprehensible, and thus can be measured by levels of transparency
- *fidelity*: the extent to which extracted model representations accurately capture the opaque models from which they were extracted
- *accuracy*: the ability of extracted representations to accurately predict unseen model examples
- *scalability*: the ability of extracted representations to scale to opaque models with large input spaces and large numbers of weighted connections
- *generality*: the extent to which extracted representations require special training procedures or restrictions on opaque models

* M. Craven, J. Shavlik (1999). Rule Extraction: Where Do We Go from Here. University of Wisconsin.

Perspectives of Explainability (cont.)

Types of explanations:*

- *text explanations*: symbol-based explainable representations:
  - natural language text
  - propositional symbols that explain the model's behavior by defining abstract concepts that capture high-level processes
- *visual explanations*: generated visualizations of models:
  - inherit challenges: e.g., human inability to grasp more than 3 dimensions
  - provide insights on the decision boundaries and inter-interactions of models' features
  - used as complementary techniques: especially when appealing to a non-expert audience
- *local explanations:* explain how a model operates considering specific interests
  - do not necessarily generalize to a global scale that represents a model's overall behavior
  - approximate the model around the instance a user wants to explain

* A. Arrieta et al. (2019). Explainable Artificial Intelligence (Xai): Concepts, Taxonomies, Opportunities and Challenges toward Responsible Ai. arXiv preprint arXiv:1910.10045

Perspectives of Explainability (cont.)

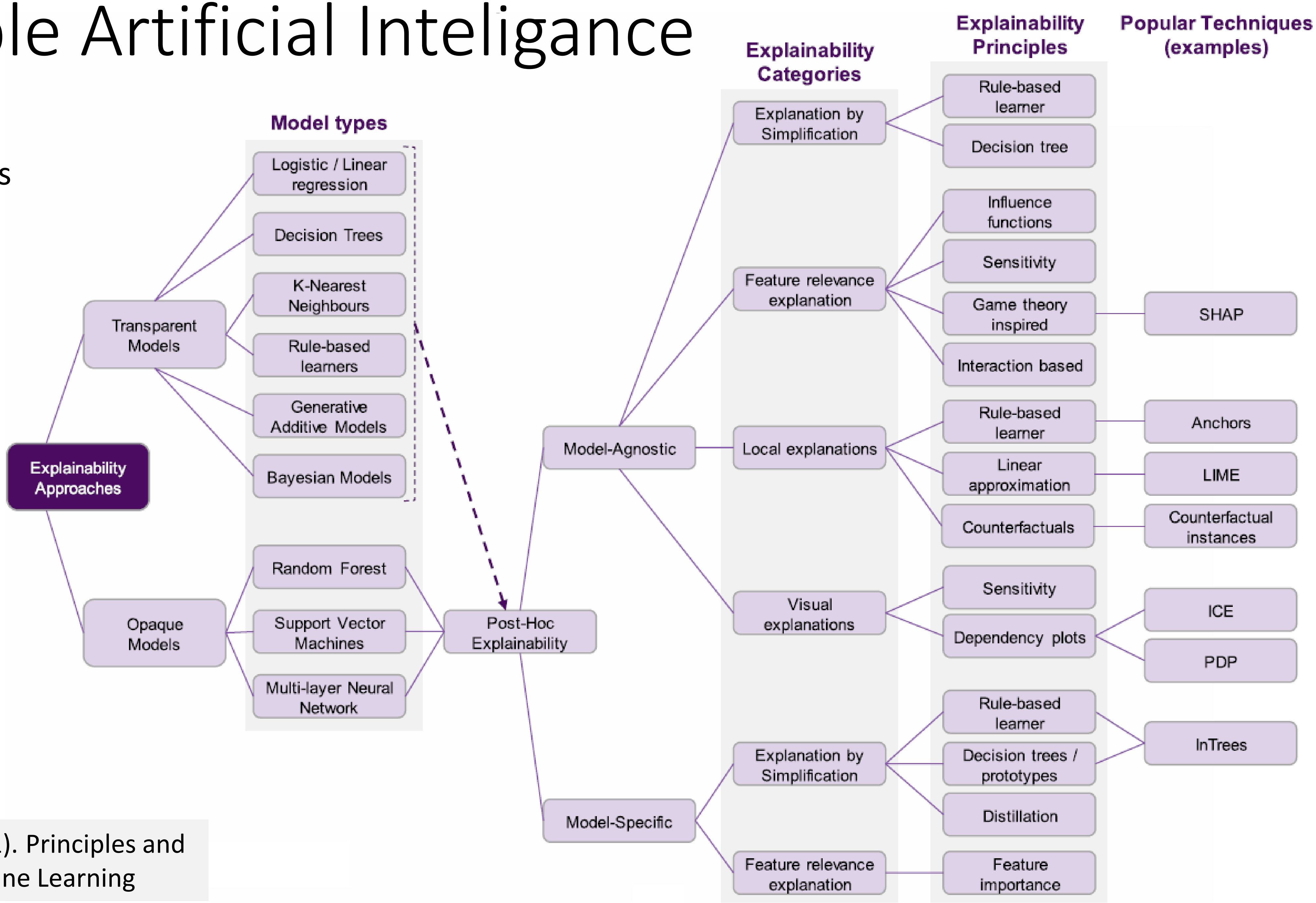
Types of explanations (cont.):

- *explanations by example*: extract representative instances from the training dataset to demonstrate how the model operates:
  - address the way how humans approach explanations in many cases, where they provide specific examples to describe a more general process
  - the training data must be in a form that is comprehensible by humans (e.g., images), because training datasets with hundreds of variables are difficult to follow and understand
- *explanations by simplification*: refer to techniques that approximate an opaque model using a simpler one (easier to interpret)
  - the simple model must be flexible enough so it can approximate the complex model accurately
  - measured by comparing the accuracy of these two models
- *feature-relevance explanations*: explain model's results by quantifying the influence of each input variable:
  - input variables are ranked by importance
  - provide some insights about the model's reasoning procedure

Explainable Artificial Inteligance

*Taxonomy framework**:*

- *tier 1*: arranges models in terms of explainability classes: transparent models and opaque models
- *tier 2*: subsequent frameworks are based on this taxonomy - elaboration on the distinction between transparent and opaque ML models
- *tier 3*: capabilities of the explainability approaches (XAI capability framework)



* V. Belle, I. Papantonis (2021). Principles and Practice of Explainable Machine Learning

Summary

Explainability in the Context of AI

AI and Understandability

AI and Explainability

Perspectives of Explainability

Explainable Machine Learning

Next Lesson:

- Transparent Machine Learning

Thank You!

Questions?