interpretable machine learning (interpretable ML)

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Abstract

Interpretable machine learning (interpretable ML) refers to machine learning (ML) methods whose hypothesis space contains only hypotheses that a human user can comprehend directly, such as sparse linear models or shallow decision trees. Such methods maximally enable the user to understand how a prediction changes when the features of a data point change, and how the predictions change when a different training set is used. An interpretable hypothesis is its own explanation; post hoc explanations of the predictions of an opaque learned hypothesis can instead be unfaithful. Restricting the hypothesis space to interpretable hypotheses also acts as a form of regularization. Interpretable surrogates serve opaque hypotheses as well: local interpretable model-agnostic explanations (LIME) approximates them locally by sparse linear hypotheses, and explainable empirical risk minimization (EERM) penalizes the discrepancy to a simple surrogate during training.

Definition

Interpretable machine learning (ML) refers to ML methods whose hypothesis space contains only hypotheses that a human user can comprehend directly (see interpretability). Such methods maximally enable the user to understand both the input-output behavior of a learned hypothesis and the inner workings of training. Ultimately, an interpretable ML method allows the user to understand how a prediction changes when the features of a data point change, and how the predictions change when a different training set is used.

Examples include linear models that combine a small number of meaningful features, shallow decision trees whose predictions follow from a few explicit tests, generalized additive models (GAMs), and lists of decision rules (Molnar, 2025). For example, a medical triage system can use a decision tree with two tests of vital signs: the staff can trace every prediction by reading the tree (see Fig. 1).

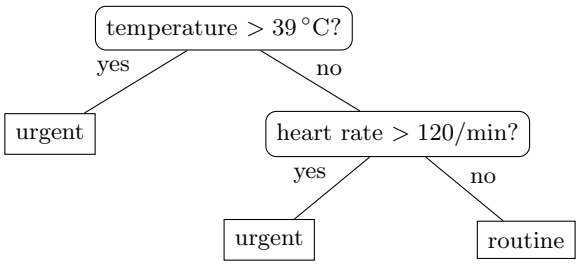

Figure 1: An interpretable hypothesis: a shallow decision tree for medical triage. Every prediction follows from at most two explicit tests of the features, so a human user can trace and anticipate the predictions without additional explanations

An interpretable hypothesis makes the effect of feature changes explicit. For a linear model $h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, changing the feature x_j by an amount Δ changes the prediction by exactly $w_j \Delta$. Each weight states how strongly one feature affects the prediction. For the decision tree of Fig. 1, a prediction changes only when a feature crosses one of the explicit test thresholds. An interpretable ML method also makes the effect of training set changes traceable. This dependence is studied under stability, which formalizes an ML method as a map from a training set to a learned hypothesis. For linear regression, this map is available in closed form through the normal equations, so the user can compute how a perturbed label in the training set shifts the predictions (see linear regression).

An interpretable hypothesis is its own explanation: no separate explanation machinery is required. Such a model is called intrinsically interpretable (Molnar, 2025). This contrasts with explainable artificial intelligence (XAI), which constructs post hoc explanations for the predictions of a possibly opaque learned hypothesis. For high-stakes applications, using an interpretable model is often preferable to explaining an opaque model, since post hoc explanations can be unfaithful to the hypothesis they explain (Rudin, 2019).

An explanation is faithful when it reflects the actual computation of the hypothesis it explains. For a surrogate hypothesis serving as explanation, faithfulness can be quantified by the agreement between its predictions and those of the explained hypothesis on a test set. A post hoc explanation that is faithful everywhere would coincide with the explained hypothesis itself. Any simpler explanation must therefore deviate somewhere (Rudin, 2019). Moreover, on structured (tabular) data, hypotheses learned within an interpretable model often achieve accuracy comparable to hypotheses learned within an opaque model (Rudin, 2019).

In an influential taxonomy, interpretable models are called transparent, with three degrees of transparency: simulatability (a human user can reproduce the computation of the hypothesis), decomposability (every part of the hypothesis, such as a single test in a decision tree, admits an intuitive description), and algorithmic transparency (the behavior of the training algorithm is understood)

(Lipton, 2018). The two questions above map onto this taxonomy: simulatability and decomposability concern the input-output behavior of a hypothesis, while algorithmic transparency concerns the dependence of the learned hypothesis on the training set. Definitions of the underlying property, and its treatment by terminology standards, are discussed under interpretability.

Regulation uses the neighboring notion of transparency rather than defining interpretability: the EU AI Act requires that a high-risk artificial intelligence system (high-risk AI system) be designed so that its operation is sufficiently transparent to enable deployers to interpret the output of the artificial intelligence system (AI system) and use it appropriately (European Parliament and Council of the European Union, 2024, Art. 13). An interpretable ML method supports this requirement, since the deployer can trace every prediction back all the way to the choice of training set (see right to explanation).

Restricting the hypothesis space to interpretable hypotheses is a form of regularization: it prunes the model and thereby reduces the risk of overfitting. Least absolute shrinkage and selection operator (Lasso) illustrates this connection: its penalty term favors sparse linear models, which use few features and are therefore easier to interpret (Hastie et al., 2009, Sect. 3.4). Explainable empirical risk minimization (EERM) also uses a penalty term, one that targets interpretability directly (Zhang et al., 2024). Its penalty term measures the discrepancy between a hypothesis and a simple surrogate, steering the training of a high-dimensional model toward hypotheses that stay close to the surrogate (see Fig. 2).

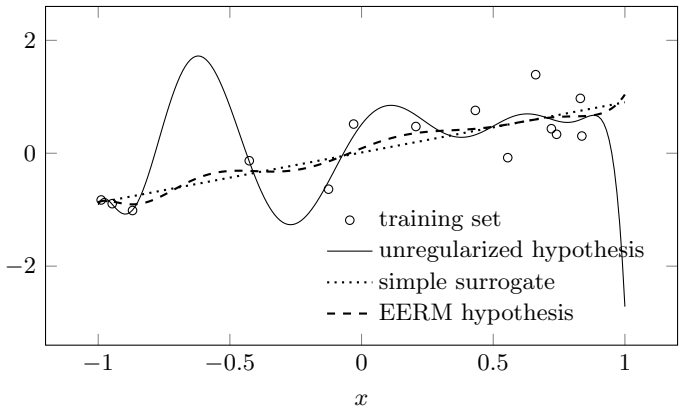

Figure 2: The idea of EERM (Zhang et al., 2024): training a high-dimensional model (polynomials of degree 10) with a penalty term proportional to the average squared discrepancy to a simple surrogate (dotted; a least-squares line). The unregularized hypothesis (thin solid) overfits the training set, while the EERM hypothesis (dashed) stays close to the surrogate and generalizes better. Data generated by `pythondemos/interpretableml.py`

Interpretable hypotheses also serve ML methods whose hypothesis space is not interpretable: a learned hypothesis $\widehat{h}$ can be approximated locally, near a given data point, by a linear hypothesis that depends on only a few features (see Fig. 3). This local sparse surrogate is the idea of local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016).

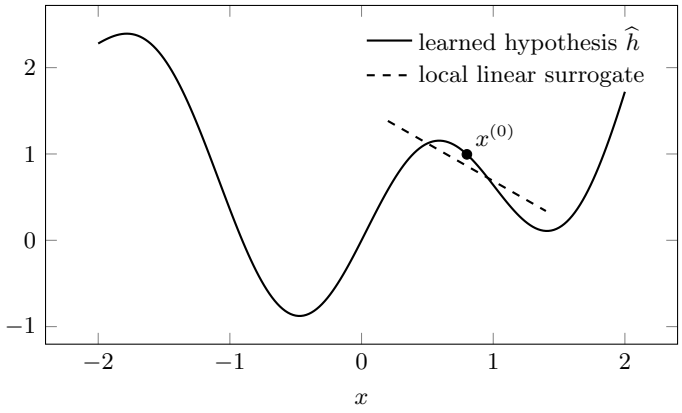

Figure 3: The idea of LIME: near the data point with feature $x^{(0)} = 0.8$, the learned hypothesis $\widehat{h}$ (solid) is approximated by a linear hypothesis (dashed) fitted with proximity weights. For data points with several features, this local surrogate depends on only a few of them. Data generated by `pythondemos/interpretableml.py`

Cross References

interpretability, explainability, XAI, decision tree, linear model, GAM, Lasso, regularization, stability, LIME, EERM, transparency, EU AI Act, trustworthy AI

References

1. European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). Official Journal of the European Union, L series, Jul. 12. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

2. Lipton (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3), pp. 31–57.

3. Molnar (2025). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 3rd ed., Ebook.

4. Ribeiro et al. (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. In: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 1135–1144.

5. Zhang et al. (2024). Explainable empirical risk minimization. Neural Comput. Appl., 36(8), pp. 3983–3996.

6. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.

7. Rudin (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. Nature Mach. Intell., 1(5), pp. 206–215.