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. A penalty term can pick a comprehensible hypothesis out of a hypothesis space that mostly holds others: least absolute shrinkage and selection operator (Lasso) drives most weights of a linear model to zero, leaving a hypothesis a user can read off, and explainable empirical risk minimization (EERM) charges the departure from the predictions one user supplies 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. Thus, interpretable ML methods require suitable choices for the hypothesis space and the training algorithm.

Examples of interpretable ML include empirical risk minimization (ERM) with 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). Fig. 1 depicts a decision tree for medical triage in a hospital. This system involves two tests of vital signs, allowing the staff to trace every prediction by reading the tree.

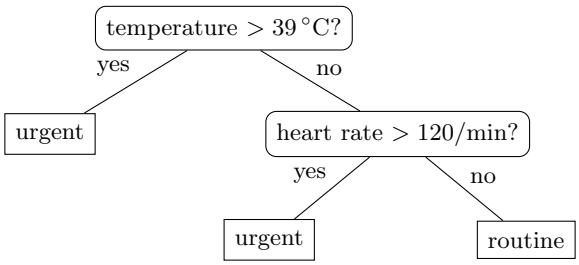

Figure 1: An interpretable hypothesis: a shallow decision tree for medical triage. Every prediction follows from at most two explicit tests of the features, so a human user can trace and anticipate the predictions

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. A patient at 38.5 °C with a heart rate of 110/min is routine; a rise to 38.9 °C leaves that prediction unchanged, while a rise to 39.5 °C turns it urgent.

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 is constructed for it. A hypothesis space containing only such hypotheses 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, such as the medical triage of Fig. 1, using an interpretable model can be preferable to explaining an opaque model, since post hoc explanations can be unfaithful to the hypothesis they explain (Rudin, 2019).

An interpretable learned hypothesis can also be obtained without restricting the hypothesis space beforehand. A penalty term added to the objective function of ERM charges the hypotheses a user cannot follow, so training picks a comprehensible one out of a hypothesis space that mostly holds the others. Least absolute shrinkage and selection operator (Lasso) does this for linear models, which are no more comprehensible than a deep net once they weigh hundreds of features. Lasso uses the ℓ_1-norm $\|\mathbf{w}\|_1$ of the weights $\mathbf{w}$ as the penalty term, which drives most weights to exactly zero (Hastie et al., 2009, Sect. 3.4). The learned hypothesis is then a weighted sum of a few features, so a user can name the features the prediction rests on and how much each one moves it. Explainable empirical risk minimization (EERM) extends this idea to the training of

an arbitrary model. Its penalty term is built from the predictions that one specific user supplies for the data points of the training set, charging the learned hypothesis for what those predictions do not already account for (see EERM). It therefore delivers explainability to that user rather than interpretability, and it leaves the hypothesis space as it is. EERM is a sibling of interpretable ML, not an instance of it: the two reach the same end by different forms of regularization, model pruning here and loss penalization there. The user predictions need not be elicited one data point at a time. They might be obtained as the predictions of a simpler proxy model that the user considers interpretable, so that the penalty term measures the departure from that proxy. In Fig. 2 the user predictions are those of a least-squares line.

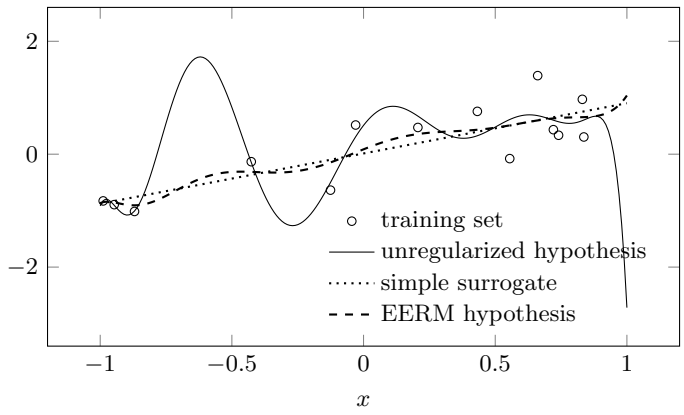

Figure 2: The idea of EERM (Zhang et al., 2024): training a high-dimensional model (polynomials of degree 10) with a penalty term proportional to the average squared discrepancy to the user summary (a least-squares surrogate; dotted line). The unregularized hypothesis (thin solid line) overfits the training set, while EERM delivers a hypothesis (dashed line) that stays close to the surrogate and generalizes better. 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. Molnar (2025). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 3rd ed., Ebook.

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

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

4. 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.