Machine Learning Interpretability

The Good, the Bad, and the Ugly

Patrick Hall

H_2O.ai

Aug. 1 2018

Contents

Obligatory Front Matter

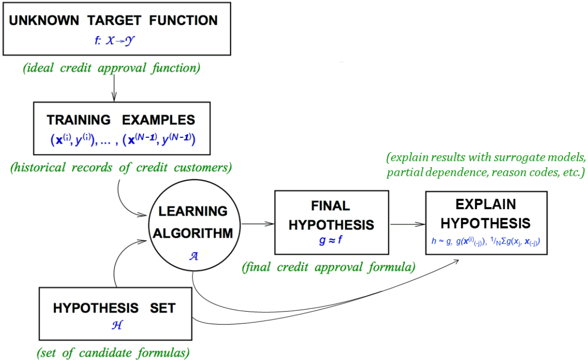
- **What is interpretation?** “The ability to explain or to present in understandable terms to a human.” Doshi-Velez and Kim, 2017
- **What is a good interpretation?** "When you can no longer keep asking why." Gilpin et al., 2018
- **Why should you care?**
  - Understanding of an impactful and quickly expanding set of technologies.
  - Addressing accidental or intentional discrimination.
  - Preventing malicious hacking and adversarial attacks.
  - Enabling regulatory compliance and increased financial margins.

Notation

- **Spaces**.
  - The input features come from a set $\mathcal{X}$ contained in a P-dimensional input space (i.e. $\mathcal{X} \subset \mathbb{R}^P$).
  - The output responses come from a set $\mathcal{Y}$ contained in a C-dimensional output space (i.e. $\mathcal{Y} \subset \mathbb{R}^C$).

- **Dataset**. A dataset $\mathbf{D}$ consists of N tuples of observations: $[(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}), (\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \dots, (\mathbf{x}^{(N-1)}, \mathbf{y}^{(N-1)})], \mathbf{x}^{(i)} \in \mathcal{X}, \mathbf{y}^{(i)} \in \mathcal{Y}$.
  - The input data $\mathbf{X}$ is composed of the set of row vectors $\mathbf{x}^{(i)}$.
    - let $\mathcal{P}$ be the set of features $\{X_0, X_1, \dots, X_{P-1}\}$, where $X_j = \left[x_j^{(0)}, x_j^{(1)}, \dots, x_j^{(N-1)}\right]^T$.
    - then each i-th observation denoted as $\mathbf{x}^{(i)} = \left[x_0^{(i)}, x_1^{(i)}, \dots, x_{P-1}^{(i)}\right]$ is an instance of $\mathcal{P}$.

Proposed Updates to the Learning Problem



The learning problem. Adapted from Abu-Mostafa, Magdon-Ismail, and Lin, 2012.

Surrogate Decision Trees (DT)

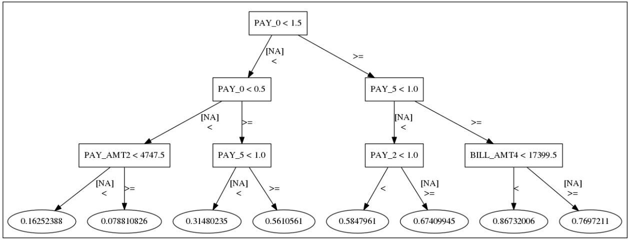


Figure: h_{tree} for Taiwanese credit card data Lichman, 2013, and for machine-learned GBM response function g.

- Given a learned function g and set of predictions $g(\mathbf{X})$, a surrogate DT can be trained: $\mathbf{X}, g(\mathbf{X}) \xrightarrow{\mathcal{A}_{\text{surrogate}}} h_{\text{tree}}$.
- h_{tree} displays a low-fidelity, high-interpretability flow chart of g's decision making process, and important features and interactions in g.

Surrogate Decision Trees (DT)

- Always use error measures to assess the trustworthiness of h_{tree}.
- Prescribed methods (Craven and Shavlik, 1996; Bastani, Kim, and Bastani, 2017) for training h_{tree} do exist. In practice, straightforward cross-validation approaches are typically sufficient.
- Comparing cross-validated training error to traditional training error can give an indication of the stability of the single tree model, h_{tree}.
- Hu et al., 2018 use local linear surrogate models, h_{GLM}, in h_{tree} leaf nodes to increase overall surrogate model fidelity while also retaining a high degree of interpretability.

Partial Dependence (PD) and Individual Conditional Expectation (ICE)

- Following Friedman, Hastie, and Tibshirani, 2001 a single feature $X_j \in \mathbf{X}$ and its complement set $X_{(-j)}$ (where $X_j \cup X_{(-j)} = \mathbf{X}$) is considered.
- $\text{PD}(X_j, g)$ for a given feature X_j is estimated as the average of the output of the learned function g, where all the components of X_j are set to a constant $x \in \mathcal{X}$, and $X_{(-j)}$ is left untouched.
- $\text{ICE}(\mathbf{x}_j^{(i)}, g)$ for a given observation $\mathbf{x}^{(i)}$ and feature X_j is estimated as the output of the learned function g where $x_j^{(i)}$ is set to a constant $x \in \mathcal{X}$ and $\mathbf{x}^{(i)} \in X_{(-j)}$ are left untouched.
- PD and ICE are usually plotted over some set of interesting constants $x \in \mathcal{X}$.

Partial Dependence (PD) and Individual Conditional Expectation (ICE)

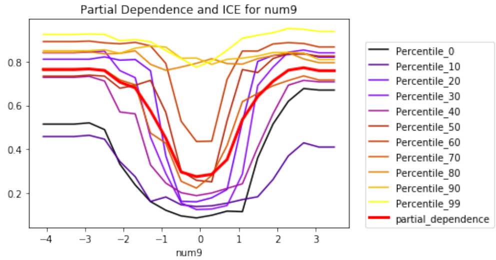
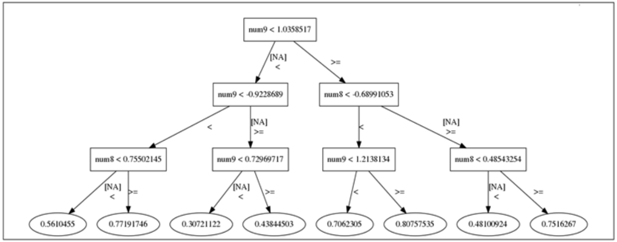


Figure: PD and ICE curves for $X_j = \text{num}_9$, for known signal generating function $f(\mathbf{X}) = \text{num}_1 * \text{num}_4 + |\text{num}_8| * \text{num}_9^2 + e$, and for machine-learned GBM response function g.

Overlaying PD and ICE curves is a succinct method for describing global and local prediction behavior and can be used to detect interactions. Goldstein et al., 2015

Partial Dependence (PD) and Individual Conditional Expectation (ICE)



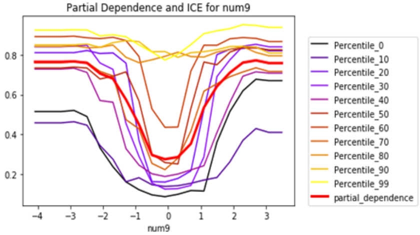


Figure: Surrogate DT, PD, and ICE curves for $X_j = \text{num}_9$, for known signal generating function $f(\mathbf{X}) = \text{num}_1 * \text{num}_4 + |\text{num}_8| * \text{num}_9^2 + e$, and for machine-learned GBM response function g.

Combining Surrogate DT models with PD and ICE curves is a convenient method for detecting, confirming, and understanding important interactions.

Local Interpretable Model-agnostic Explanations (LIME)

Ribeiro, Singh, and Guestrin, 2016 defines LIME for some observation $\mathbf{x} \in \mathcal{X}$:

$$\underset{h \in \mathcal{H}}{\arg\max}\, \mathcal{L}(g, h, \pi_{\mathbf{X}}) + \Omega(h)$$

Here g is the function to be explained, h is an interpretable surrogate model of g, often a linear model h_{GLM}, $\pi_{\mathbf{X}}$ is a weighting function over the domain of g, and $\Omega(h)$ limits the complexity of h.

Typically, h_{GLM} is constructed such that $\mathbf{X}^{(*)}, g(X^{(*)}) \xrightarrow{\mathcal{A}_{\text{surrogate}}} h_{\text{GLM}}$, where $\mathbf{X}^{(*)}$ is a generated sample, $\pi_{\mathbf{X}}$ weighs $\mathbf{X}^{(*)}$ samples by their Euclidean similarity to $\mathbf{x}$, local feature importance is estimated using $\beta_j x_j$, and L_1 regularization is used to induce a simplified, sparse h_{GLM}.

Local Interpretable Model-agnostic Explanations (LIME)

- LIME is ideal for creating low-fidelity, highly interpretable explanations for non-DT models and for neural network models trained on unstructured data, e.g. deep learning.
- Always use regression fit measures to assess the trustworthiness of LIME explanations.
- LIME can be difficult to deploy, but there are highly deployable variants. Hu et al., 2018; Hall et al., 2017
- Local feature importance values are offsets from a local intercept.
  - Note that the intercept in LIME can account for the most important local phenomena.
  - Generated LIME samples can contain large proportions of out-of-range data that can lead to unrealistic intercept values.

- To increase the fidelity of LIME explanations, try LIME on discretized input features and on manually constructed interactions.
- Use cross-validation to construct standard deviations or even confidence intervals for local feature importance values.
- LIME can fail, particularly in the presence of extreme nonlinearity or high-degree interactions.

Tree Shap

Shapley explanations are a class of additive, consistent local feature importance measures with long-standing theoretical support, Lundberg and Lee, 2017. For some observation $\mathbf{x} \in \mathcal{X}$, Shapley explanations take the form:

$$\phi_0 + \sum_{j=0}^{j=\mathcal{P}-1} \phi_j \mathbf{x}'_j$$

Here $\mathbf{x}' \in \{0,1\}^{\mathcal{P}}$ is a binary representation of $\mathbf{x}$ where 0 indicates missingness. Each ϕ_j is the local feature importance value associated with x_j.

- Calculating Shapley values directly is typically infeasible, but they can be estimated in different ways.
- Tree Shap is a specific implementation of Shapley explanations that leverages DT structures to disaggregate the contribution of each x_j to $g(\mathbf{x})$ in a DT or DT-based ensemble model. Lundberg, Erion, and Lee, 2018

Tree Shap

- Tree Shap is ideal for high-fidelity explanations of DT-based models, perhaps even in regulated applications.
- Local feature importance values are offsets from a global intercept.
- LIME can be constrained to become Shapley explanations.
- A similar, popular method known as *treeinterpreter* appears untrustworthy when applied to GBM models.

Closing Recommendations

- Monotonically constrained XGBoost, Surrogate DT, PD and ICE plots, and Tree Shap are a direct and open source way to create an interpretable nonlinear model.
- Global and local explanatory techniques are often necessary to explain a model.
- Use simpler low-fidelity or sparse explanations to understand more accurate and complex high-fidelity explanations.
- Seek consistent results across multiple explanatory techniques.
- Methods relying on generated data are sometimes unpalatable to users. They want to understand *their* data.
- Surrogate models can provide low-fidelity explanations for model mechanisms in original feature spaces if g is defined to include feature extraction or engineering.
- To increase adoption, production deployment of explanatory methods must be straightforward.

Software Examples and Resources

Comparison of Explanatory Techniques on Simulated Data:
https://github.com/h2oai/mli-resources/tree/master/lime_shap_treeint_compare

In-depth Explanatory Technique Examples:
https://github.com/jphall663/interpretable_machine_learning_with_python

"Awesome" Machine Learning Interpretability Resource List:
https://github.com/jphall663/awesome-machine-learning-interpretability

References

Abu-Mostafa, Yaser S., Malik Magdon-Ismail, and Hsuan-Tien Lin (2012). ***Learning from Data***. New York: AMLBook. URL: https://work.caltech.edu/textbook.html.

Bastani, Osbert, Carolyn Kim, and Hamsa Bastani (2017). "Interpreting blackbox models via model extraction." In: *arXiv preprint arXiv:1705.08504*. URL: https://arxiv.org/pdf/1705.08504.pdf.

Craven, Mark W. and Jude W. Shavlik (1996). "Extracting Tree-Structured Representations of Trained Networks." In: *Advances in Neural Information Processing Systems*. URL: http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.pdf.

Doshi-Velez, Finale and Been Kim (2017). "Towards a rigorous science of interpretable machine learning." In: *arXiv preprint arXiv:1702.08608*. URL: https://arxiv.org/pdf/1702.08608.pdf.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). ***The Elements of Statistical Learning***. New York: Springer. URL: https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf.

Gilpin, Leilani H et al. (2018). "Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning." In: *arXiv preprint arXiv:1806.00069*. URL: https://arxiv.org/pdf/1806.00069.pdf.

Goldstein, Alex et al. (2015). "Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation." In: *Journal of Computational and Graphical Statistics* 24.1.

References

Hall, Patrick et al. (2017). *Machine Learning Interpretability with H2O Driverless AI*. URL: `http://docs.h2o.ai/driverless-ai/latest-stable/docs/booklets/MLIBooklet.pdf`.

Hu, Linwei et al. (2018). "Locally Interpretable Models and Effects based on Supervised Partitioning (LIME-SUP)." In: *arXiv preprint arXiv:1806.00663*. URL: `https://arxiv.org/ftp/arxiv/papers/1806/1806.00663.pdf`.

Lichman, M. (2013). *UCI Machine Learning Repository*. URL: `http://archive.ics.uci.edu/ml`.

Lundberg, Scott M, Gabriel G Erion, and Su-In Lee (2018). "Consistent Individualized Feature Attribution for Tree Ensembles." In: *arXiv preprint arXiv:1802.03888*. URL: `https://arxiv.org/pdf/1706.06060.pdf`.

Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions." In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 4765–4774. URL: `http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf`.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). "Why should I trust you?: Explaining the predictions of any classifier." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 1135–1144. URL: `http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf`.