Machine Learning Interpretability

The Good, the Bad, and the Ugly

Patrick Hall

H_2O.ai

Aug. 1 2018

Contents

Obligatory Front Matter

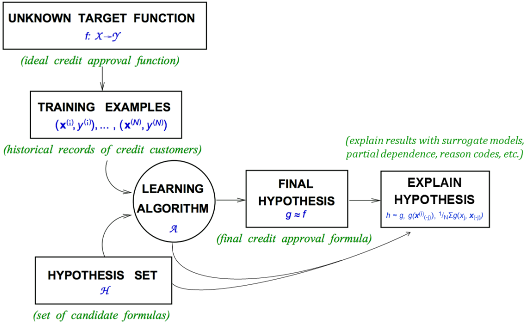
- **What is interpretation?** "The ability to explain or to present in understandable terms to a human." Doshi-Velez and Kim, 2017
- **What is a good interpretation?** "When you can no longer keep asking why." Gilpin et al., 2018
- **Why should you care?**
  - Addressing accidental or intentional discrimination.
  - 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. $\mathcal{Y} \subset \mathbb{R}^C$).

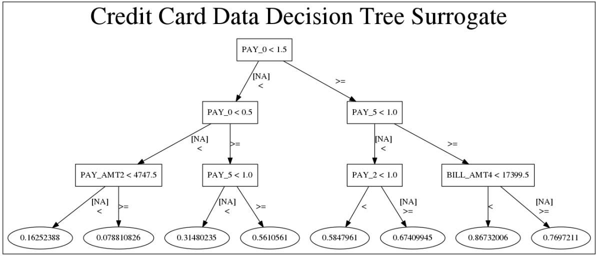
- **Dataset**. A dataset $\mathbf{D}$ consists of N tuples of observations: $[(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}), (\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \dots, (\mathbf{x}^{(N-1)}, \mathbf{y}^{(N-1)})], \mathbf{x}^{(i)} \in \mathcal{X}, \mathbf{y}^{(i)} \in \mathcal{Y}$.
  - The input data $\mathbf{X}$ 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 *Learning From Data*, Abu-Mostafa, Magdon-Ismail, and Lin, 2012.

Surrogate Decision Trees



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

Surrogate Decision Trees

- 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 error to standard 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 accuracy while retaining a high degree of interpretability.
- h_{tree} can provide low-fidelity explanations for model mechanisms in the original feature space if g is defined to include feature extraction.

Partial Dependence - *Description*

Partial Dependence - *Recommendations*

Individual Conditional Expectation (ICE) - *Description*

Individual Conditional Expectation (ICE) - *Recommendations*

Local Interpretable Model-agnostic Explanations (LIME) - *Description*

Local Interpretable Model-agnostic Explanations (LIME) - *Recommendations*

Tree Shap - *Description*

Tree Shap - *Recommendations*

Closing Recommendations

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.

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.

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.