Introduction to Responsible Machine Learning*

Lecture 1: Interpretable Machine Learning Models

Patrick Hall

The George Washington University

May 21, 2020

Contents

Grading and Policy

- Grading:
  - $\frac{1}{3}$ Participation
  - $\frac{1}{3}$ Project GitHub or Kaggle kernel
  - $\frac{1}{3}$ Public Kaggle leaderboard score
- Project:
  - Kaggle competition using techniques from class
  - Individual or group (no more than 4 members)
  - Select team members ASAP
- Syllabus
- Webex office hours: Thurs. 5-6 pm or by appointment
- Class resources: `https://jphall663.github.io/GWU_rml/`

Overview

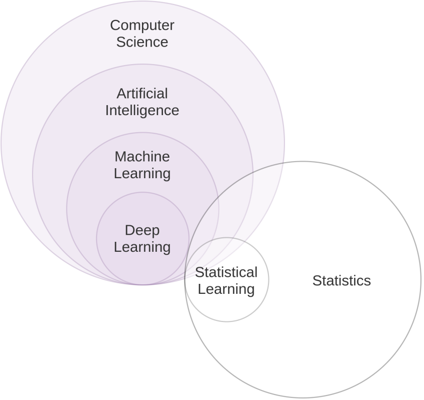
- **Class 1**: Interpretable Models
- **Class 2**: Post-hoc Explanations
- **Class 3**: Fairness
- **Class 4**: Security
- **Class 5**: Model Debugging
- **Class 6**: Best Practices

Responsible Artificial Intelligence

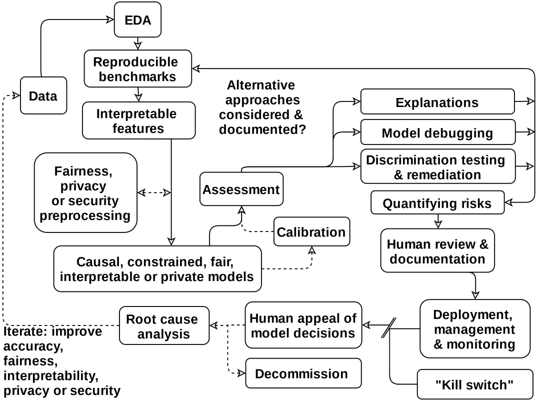
"Responsible Artificial Intelligence is about human responsibility for the development of intelligent systems along fundamental human principles and values, to ensure human-flourishing and well-being in a sustainable world."

— Virginia Dignum, ***Responsible Artificial Intelligence***

What About Machine Learning?

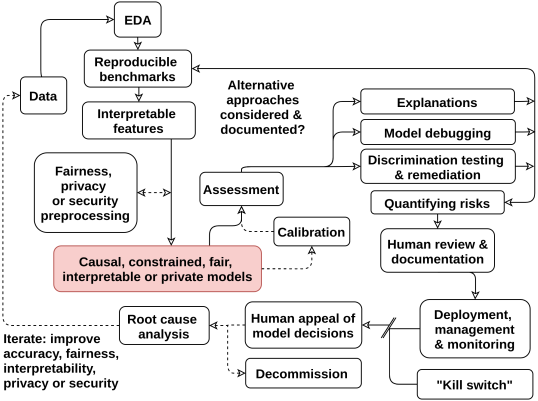


A Responsible Machine Learning Workflow


EDA
Reproducible benchmarks
Data
Interpretable features
Alternative approaches considered & documented?
Explanations
Model debugging
Discrimination testing & remediation
Quantifying risks
Fairness, privacy or security preprocessing
Assessment
Calibration
Causal, constrained, fair, interpretable or private models
Human review & documentation
Deployment, management & monitoring
Root cause analysis
Human appeal of model decisions
Decommission
"Kill switch"
Iterate: improve accuracy, fairness, interpretability, privacy or security

Source: *A Responsible Machine Learning Workflow*.

A Responsible ML Workflow: Interpretable Models


EDA
Reproducible benchmarks
Data
Interpretable features
Alternative approaches considered & documented?
Explanations
Model debugging
Discrimination testing & remediation
Fairness, privacy or security preprocessing
Assessment
Quantifying risks
Calibration
Human review & documentation
Causal, constrained, fair, interpretable or private models
Deployment, management & monitoring
Root cause analysis
Human appeal of model decisions
Iterate: improve accuracy, fairness, interpretability, privacy or security
Decommission
"Kill switch"

Source: *A Responsible Machine Learning Workflow*.

Interpretable ML Models

Doshi-Velez and Kim, 2017, define interpretable as, "the ability to explain or to present in understandable terms to a human."

There are many types of interpretable ML models. Some might be directly interpretable to non-technical consumers. Some are only interpretable to highly-skilled data scientists. Interpretability is not an on-and-off switch.

Interpretable models are crucial for documentation, explanation of predictions to consumers, finding and fixing discrimination, and debugging other problems in ML modeling pipelines. Simply put, **it is very difficult to mitigate risks you don't understand**.

There is not necessarily a trade-off between accuracy and interpretability, especially for structured data.

Background

We will frequently refer to the following terms and definitions today:

- Notation
- Pearson correlation
  - Measurement of the linear relationship between two input X_j features; takes on values between -1 and $+1$, including 0.
- Shapley value
- Partial dependence and individual conditional expectation (ICE)
- Gradient boosting machine (GBM)

Background: Notation

Spaces

- Input features come from the set $\mathcal{X}$ contained in a P-dimensional input space, $\mathcal{X} \subset \mathbb{R}^P$. An arbitrary, potentially unobserved, or future instance of $\mathcal{X}$ is denoted $\mathbf{x}$, $\mathbf{x} \in \mathcal{X}$.
- Labels corresponding to instances of $\mathcal{X}$ come from the set $\mathcal{Y}$.
- Learned output responses come from the set $\hat{\mathcal{Y}}$.

Background: Notation

Datasets

- The input dataset $\mathbf{X}$ is composed of observed instances of the set $\mathcal{X}$ with a corresponding dataset of labels $\mathbf{Y}$, observed instances of the set $\mathcal{Y}$.
- Each i-th observation of $\mathbf{X}$ is denoted as $\mathbf{x}^{(i)} = [x_0^{(i)}, x_1^{(i)}, \dots, x_{P-1}^{(i)}]$, with corresponding i-th labels in $\mathbf{Y}, \mathbf{y}^{(i)}$, and corresponding predictions in $\hat{\mathbf{Y}}, \hat{\mathbf{y}}^{(i)}$.
- $\mathbf{X}$ and $\mathbf{Y}$ consist of N tuples of observations: $[(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}), (\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \dots, (\mathbf{x}^{(N-1)}, \mathbf{y}^{(N-1)})]$.
- Each j-th input column vector of $\mathbf{X}$ is denoted as $X_j = [x_j^{(0)}, x_j^{(1)}, \dots, x_j^{(N-1)}]^T$.

Background: Notation

Models

- A type of machine learning (ML) model g, selected from a hypothesis set $\mathcal{H}$, is trained to represent an unknown signal-generating function f observed as $\mathbf{X}$ with labels $\mathbf{Y}$ using a training algorithm $\mathcal{A}$: $\mathbf{X}, \mathbf{Y} \xrightarrow{\mathcal{A}} g$, such that $g \approx f$.
- g generates learned output responses on the input dataset $g(\mathbf{X}) = \hat{\mathbf{Y}}$, and on the general input space $g(\mathcal{X}) = \hat{\mathcal{Y}}$.
- The model to be explained, tested for discrimination, or debugged is denoted as g.

Background: Shapley Value

Shapley explanations, including TreeSHAP and even certain implementations of LIME, are a class of additive, locally accurate feature contribution measures with long-standing theoretical support (Lundberg and Lee, 2017).

For some observation $\mathbf{x} \in \mathcal{X}$, Shapley explanations take the form:

$$\phi_j = \underbrace{\sum_{S \subseteq \mathcal{P} \setminus \{j\}} \frac{|S|!(\mathcal{P} - |S| - 1)!}{\mathcal{P}!}}_{\text{weighted average over all subsets in } \mathbf{X}} \underbrace{(S \cup \{j\}) - g_x(S)]}_{g \text{ "without" } x_j} \tag{1}$$

$$g(\mathbf{x}) = \phi_0 + \sum_{j=0}^{j=\mathcal{P}-1} \phi_j \mathbf{z}_j \tag{2}$$

Background: Partial Dependence and ICE

- Following Friedman, Hastie, and Tibshirani (2001) a single input feature, $X_j \in \mathbf{X}$, and its complement set, $\mathbf{X}_{\mathcal{P}\setminus\{j\}} \in \mathbf{X}$, where $X_j \cup \mathbf{X}_{\mathcal{P}\setminus\{j\}} = \mathbf{X}$ is considered. $\text{PD}(X_j, g)$ for a given feature X_j is estimated as the average output of the learned function $g(\mathbf{X})$ when all the components of X_j are set to a constant $x \in \mathcal{X}$ and $\mathbf{X}_{(-j)}$ is left unchanged.
- $\text{ICE}(x_j, \mathbf{x}, g)$ for a given instance $\mathbf{x}$ and feature x_j is estimated as the output of $g(\mathbf{x})$ when x_j is set to a constant $x \in \mathcal{X}$ and all other features $\mathbf{x} \in \mathbf{X}_{(-j)}$ are left untouched. Partial dependence and ICE curves are usually plotted over some set of constants $x \in \mathcal{X}$ (Goldstein et al., 2015).

Background: Gradient Boosting Machine

$$g^{\text{GBM}}(\mathbf{x}) = \sum_{b=0}^{B-1} T_b(\mathbf{x}; \Theta) \tag{3}$$

A GBM is a sequential combination of decision trees, T_b, where T_0 is trained to predict $\mathbf{y}$, but all subsequent T are trained to reduce the errors of T_{b-1}.

Anatomy of Elastic Net Regression

Generalized linear models (GLM) have the same basic functional form as more traditional linear models, e.g. ...

$$g^{\text{GLM}}(\mathbf{x}) = \beta_0 + \beta_1 x_0 + \beta_2 x_1 + \cdots + \beta_P x_{P-1} \tag{4}$$

... but are more robust to correlation, wide data, and outliers.

Anatomy of Elastic Net Regression: L1 and L2 Penalty

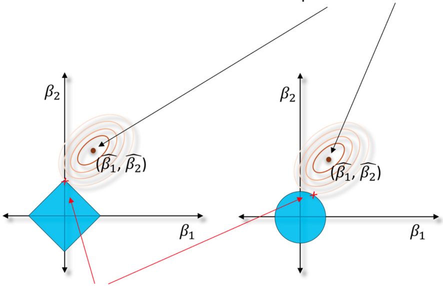
Iteratively reweighted least squares (IRLS) method with ridge (L_2) and LASSO (L_1) penalty terms:

$$\tilde{\beta} = \underset{\beta}{min}\Big\{ \underbrace{\sum_{i=0}^{N-1}(y_i - \beta_0 - \sum_{j=1}^{P-1} x_{ij}\beta_j)^2}_{1} + \underbrace{\lambda}_{2} \sum_{j=1}^{P-1}(\underbrace{\alpha}_{3}\underbrace{\beta_j^2}_{4} + (1 - \underbrace{\alpha}_{3})\underbrace{|\beta_j|}_{5})\Big\} \tag{5}$$

- 1: Least squares minimization
- 2: Controls magnitude of penalties
- 3: Tunes balance between L1 and L2
- 4: L_2/Ridge penalty term
- 5: L_1/LASSO penalty term

Graphical Illustration of Shrinkage/Regularization Method:

Coefficients selected by non-penalized model.


β_2
$(\widehat{\beta_1}, \widehat{\beta_2})$
β_1
β_2
$(\widehat{\beta_1}, \widehat{\beta_2})$
β_1

Coefficients selected by penalized model. L1 solution will be sparse.

Monotonic GBM (Gill et al., 2020)

Monotonic GBM (MGBM) constrain typical GBM training to consider only tree splits that obey user-defined positive and negative monotone constraints, with respect to each input feature, X_j, and a target feature, $\mathbf{y}$, independently. An MGBM remains an additive combination of B trees trained by gradient boosting, T_b, and each tree learns a set of splitting rules that respect monotone constraints, Θ_b^{mono}. A trained MGBM model, g^{MGBM}, takes the form:

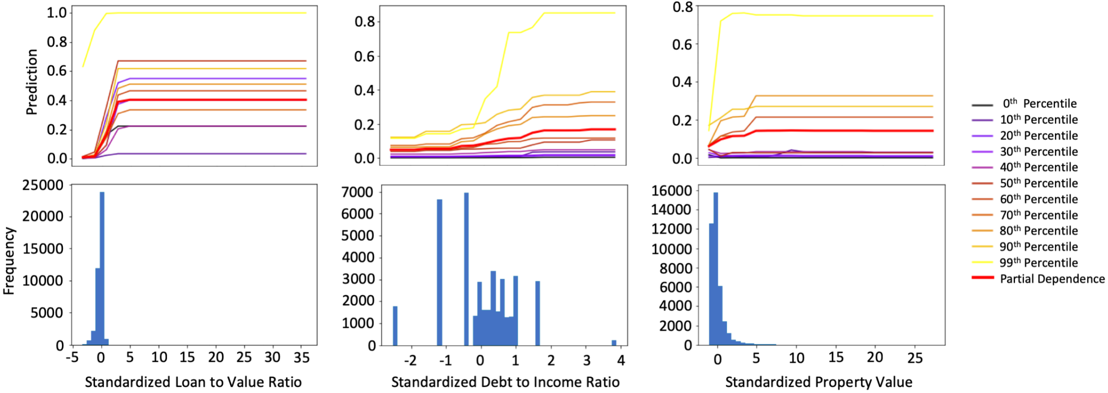
$$g^{\text{MGBM}}(\mathbf{x}) = \sum_{b=0}^{B-1} T_b\left(\mathbf{x}; \Theta_b^{\text{mono}}\right) \tag{6}$$

Monotone Constraints for GBM (Gill et al., 2020)

1. For the first and highest split in T_b involving X_j, any $\theta_{b,j,0}$ resulting in $T(x_j;\theta_{b,j,0}) = \{w_{b,j,0,L}, w_{b,j,0,R}\}$ where $w_{b,j,0,L} > w_{b,j,0,R}$, is not considered.
2. For any subsequent left child node involving X_j, any $\theta_{b,j,k\geq 1}$ resulting in $T(x_j;\theta_{b,j,k\geq 1}) = \{w_{b,j,k\geq 1,L}, w_{b,j,k\geq 1,R}\}$ where $w_{b,j,k\geq 1,L} > w_{b,j,k\geq 1,R}$, is not considered.
3. Moreover, for any subsequent left child node involving X_j, $T(x_j;\theta_{b,j,k\geq 1}) = \{w_{b,j,k\geq 1,L}, w_{b,j,k\geq 1,R}\}$, $\{w_{b,j,k\geq 1,L}, w_{b,j,k\geq 1,R}\}$ are bound by the associated $\theta_{b,j,k-1}$ set of node weights, $\{w_{b,j,k-1,L}, w_{b,j,k-1,R}\}$, such that $\{w_{b,j,k\geq 1,L}, w_{b,j,k\geq 1,R}\} < \frac{w_{b,j,k-\mathbf{1},L}+w_{b,j,k-\mathbf{1},R}}{2}$.
4. (1) and (2) are also applied to all right child nodes, except that for right child nodes $w_{b,j,k,L} \leq w_{b,j,k,R}$ and $\{w_{b,j,k\geq 1,L}, w_{b,j,k\geq 1,R}\} \geq \frac{w_{b,j,k-\mathbf{1},L}+w_{b,j,k-\mathbf{1},R}}{2}$.

Note that $g^{\mathrm{MGBM}}(\mathbf{x})$ is an addition of each full T_b prediction, with the application of a monotonic logit or softmax link function for classification problems. Moreover, each tree's root node corresponds to some constant node weight that by definition obeys monotonicity constraints, $T(x_j^{\alpha};\theta_{b,0}) = T(x_j^{\beta};\theta_{b,0}) = w_{b,0}$.

Partial Dependence and ICE:



A Burgeoning Ecosystem of Interpretable Machine Learning Models

- Generalized additive model (GAM) (Friedman, Hastie, and Tibshirani, 2001)
- GA2M / Explainable boosting machine (EBM) (Lou et al., 2013)
- Explainable Neural Network (XNN) (Vaughan et al., 2018)
- Rudin group:
  - *This looks like that* deep learning (Chen et al., 2019)
  - Scalable Bayesian rule list (Yang, Rudin, and Seltzer, 2017)
  - Optimal sparse decision tree (Hu, Rudin, and Seltzer, 2019)
  - Supersparse linear integer models (Ustun and Rudin, 2016)
  - and more ...
- RuleFit (Friedman and Popescu, 2008)

Acknowledgments

Thanks to Lisa Song for her continued assistance in developing these course materials.

References

Chen, Chaofan et al. (2019). "This Looks Like That: Deep Learning for Interpretable Image Recognition." In: *Proceedings of Neural Information Processing Systems (NeurIPS)*. URL: https://arxiv.org/pdf/1806.10574.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***. URL: https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf. New York: Springer.

Friedman, Jerome H., Bogdan E. Popescu, et al. (2008). "Predictive Learning Via Rule Ensembles." In: *The Annals of Applied Statistics* 2.3. URL: https://projecteuclid.org/download/pdfview_1/euclid.aoas/1223908046, pp. 916–954.

Gill, Navdeep et al. (2020). "A Responsible Machine Learning Workflow with Focus on Interpretable Models, Post-hoc Explanation, and Discrimination Testing." In: *Information* 11.3. URL: https://www.mdpi.com/2078-2489/11/3/137, p. 137.

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. URL: https://arxiv.org/pdf/1309.6392.pdf.

References

Hu, Xiyang, Cynthia Rudin, and Margo Seltzer (2019). "Optimal Sparse Decision Trees." In: *arXiv preprint arXiv:1904.12847*. URL: https://arxiv.org/pdf/1904.12847.pdf.

Lou, Yin et al. (2013). "Accurate Intelligible Models with Pairwise Interactions." In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.352.7682&rep=rep1&type=pdf. ACM, pp. 623–631.

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. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf. Curran Associates, Inc., pp. 4765–4774.

Ustun, Berk and Cynthia Rudin (2016). "Supersparse Linear Integer Models for Optimized Medical Scoring Systems." In: *Machine Learning* 102.3. URL: https://users.cs.duke.edu/~cynthia/docs/UstunTrRuAAAI13.pdf, pp. 349–391.

Vaughan, Joel et al. (2018). "Explainable Neural Networks Based on Additive Index Models." In: *arXiv preprint arXiv:1806.01933*. URL: https://arxiv.org/pdf/1806.01933.pdf.

Yang, Hongyu, Cynthia Rudin, and Margo Seltzer (2017). "Scalable Bayesian Rule Lists." In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. URL: https://arxiv.org/pdf/1602.08610.pdf.