Responsible Machine Learning*

Lecture 1: Interpretable Machine Learning Models

Patrick Hall

The George Washington University

May 19, 2020

Contents

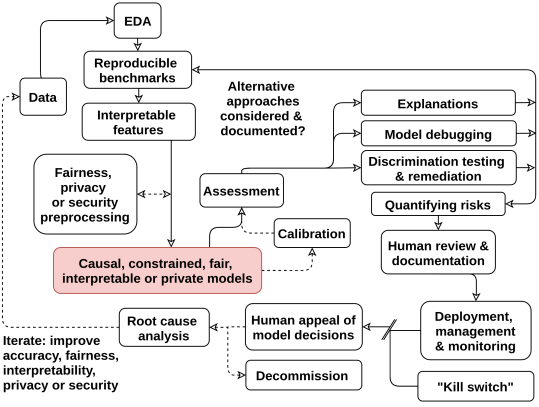
Grading and Policy

- 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

A Responsible Machine Learning Workflow†



† *A Responsible Machine Learning Workflow*

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 $\hat{\mathcal{Y}}$.

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 $\mathbf{x}^{(i)} = [x_0^{(i)}, x_1^{(i)}, \ldots, x_{P-1}^{(i)}]$, 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: $[(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}), (\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \ldots, (\mathbf{x}^{(N-1)}, \mathbf{y}^{(N-1)})]$.
- Each j-th input column vector of $\mathbf{X}$ is denoted as $X_j = [x_j^{(0)}, x_j^{(1)}, \ldots, x_j^{(N-1)}]^T$.

Notation

Models

- A type of machine learning 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

We will frequently refer to following terms and definitions today:

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

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 (**shapley**).

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

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

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

Partial Dependence and ICE

- Partial dependence (PD) plots are a widely-used method for describing the average predictions of a complex model across some partition of dataset for some interesting input feature (cite this - Friedman, Hastie, and Tibshirani, 2001)
- Individual conditional expectation (ICE) plots are a newer method that describes the local behavior of a complex model for a single input feature for each instances. PD and ICE can be combined in the same plot to compensate for known weaknesses of partial dependence, to identify the interactions of the prediction model, and to create a holistic portrait of the predictions of a complex model for some input feature (cite)
- In simplier terms, the "PD plot is the average of all the ICE plot lines"

GBM and Monotonic GBM

- Gradient Boosting Machine (GBM)is a type of ensemble decision tree model (sequential) that is fitted by minimizing a loss function averaged over a training data
- The MGBMs constrain typical GBM training to consider only tree splits that obey user-defined positive and negative monotonicity constraints, with respect to each input feature and a target feature, independently

Anatomy of Elastic Net Regression: L1 and L2 Penalty

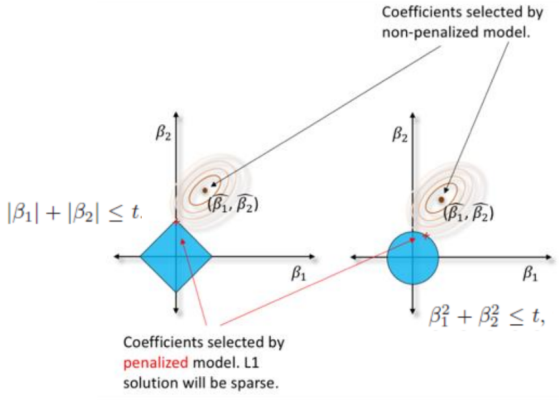
Iteratively Resweighted Least Square method with generalized Ridge (L2) and LASSOS (L1) penalty terms:

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

- 1: Least square minimization
- 2: Controls magnitude of penalties
- 3: Tunes balane between L1 and L2
- 4: L2/Ridge penalty term
- 5: L1/LASSO penalty term

Anatomy of Elastic Net Regression: L1 and L2 Penalty

Graphical Illustration of Shrinkage/Regularization Method:



Monotonic GBM: Shapley Value

MGBM and Shapley Value - insert content heres

Other Methods

- GA2M/EBM
- XNN
- Scalable Bayesian Rulest List(C. Rudin)
- CORELS(C. Rudin)

References

Link 1:
https://github.com/jphall663/

Link 2:
https://www.h2o.ai