STAT3612 Lecture 7
Interpretable Machine Learning

Dr. Aijun Zhang

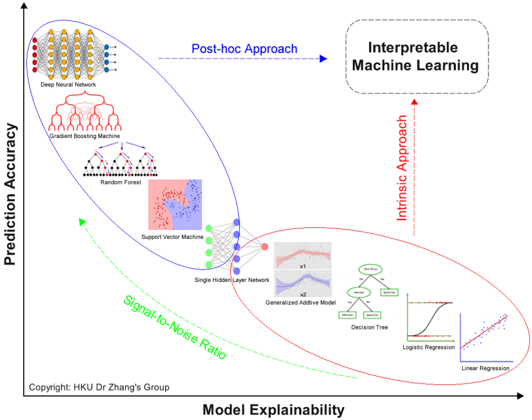
20 October 2020

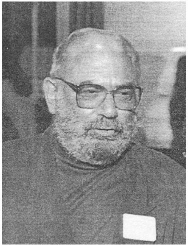
Department of 統計及精算學系
Statistics & Actuarial Science

Table of Contents

Interpretable Machine Learning

"Statistical Modeling: The Two Cultures" (Breiman 2001): Occam dilemma





Leo Breiman
(1928–2005)

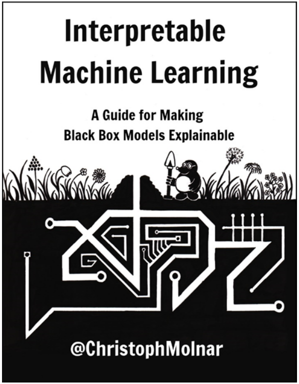
We discuss two IML approaches: Post-hoc vs. Intrinsic Interpretability ···

Global and Local Interpretability

- **Global interpretability:** to understand the modeled relationship between inputs and prediction target across entire data; to quantify global variable importance of each input variable.
- **Local interpretability:** to understand the model prediction for a single data point or a small region; to derive local variable importance for reason codes (i.e. plaintext explanations of individual prediction).
- **For white box models:** model-diagnostic and visualization methods for intrinsically interpretable models (GLM, GAM, Tree, GAIM/xNN ...)
- **For black box models:** model-agnostic post-hoc methods (VI, PDP, ICE, ALE, LIME, SHAP ...)
- Integrate global and local interpretability into data science workflow ...

A Black-cover Online Reference

- Free online book https://christophm.github.io/interpretable-ml-book/



Last updated: 2020-10-19

- Github: jphall663/awesome-machine-learning-interpretability

Table of Contents

Example: Boston House Prices

```
from sklearn.datasets import load_boston
data = load_boston()
print(data.DESCR)
```

```
.. _boston_dataset:

Boston house prices dataset
---------------------------

**Data Set Characteristics:**

    :Number of Instances: 506

    :Number of Attributes: 13 numeric/categorical predictive. Median Value (attribute 14) is usually the target.

    :Attribute Information (in order):
        - CRIM     per capita crime rate by town
        - ZN       proportion of residential land zoned for lots over 25,000 sq.ft.
        - INDUS    proportion of non-retail business acres per town
        - CHAS     Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
        - NOX      nitric oxides concentration (parts per 10 million)
        - RM       average number of rooms per dwelling
        - AGE      proportion of owner-occupied units built prior to 1940
        - DIS      weighted distances to five Boston employment centres
        - RAD      index of accessibility to radial highways
        - TAX      full-value property-tax rate per $10,000
        - PTRATIO  pupil-teacher ratio by town
        - B        1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
        - LSTAT    % lower status of the population
        - MEDV     Median value of owner-occupied homes in $1000's

    :Missing Attribute Values: None

    :Creator: Harrison, D. and Rubinfeld, D.L.
```

Intrinsically Interpretable Models

Refer to the supplementary Python notebook for the modeling results by:

- GLM (generalized linear models)
- Regularized GLM: Lasso and ElasticNet
- GAM (generalized additive models)

This also serves as a review of the materials we have discussed so far ...

Table of Contents

Black-box Models

- Run black-box models like DNN (deep neural networks) or XGBoost (extreme gradient boosting) for the Boston House data.
- Refer to the Python notebook.
- No worry about the details of these black-box models for now. We will discuss later.
- Observe the training and testing performances.
- Compare the results with intrinsically interpretable models (GLM, GAM)
- How can we explain the results by the black-box algorithms?

Post-hoc Model Explanation

Given a trained black box model (e.g. XGBoost, random forest, DNN), we can perform the following post-hoc explainability analysis:

Global explainability for the entire model:

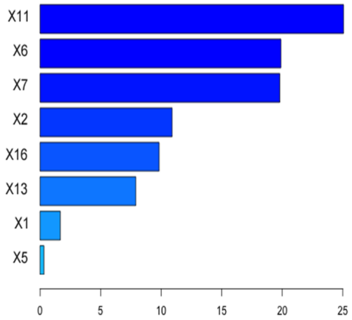
- **VI** (Variable Importance): rank-order the feature variables
- **PDP** (Partial Dependence Plot): check functional relationship
- **ICE** (Individual Conditional Expectation): per training instance
- **ALE** (Accumulated Local Effect): per training instance

Local explainability for the individual prediction:

- **LIME**: local surrogate model through permuted samples
- **SHAP**: based on the Shapley values from game theory

Variable Importance (VI)

- Breiman, L. (2001). Random forests. *Machine learning*, **45**(1), 5–32.



- VI per input variable is measured by model performance difference upon variable permutation or LOCO (Leave-One-Covariate-Out).

Partial Dependence Plot (PDP)

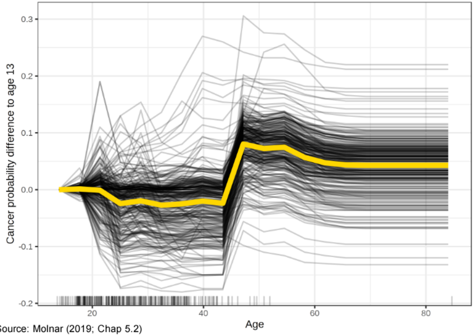
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, **29**, 1189–1232.

$$\mathrm{PDP}(x_j) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}\left(x_{i1}, \ldots, x_{i(j-1)}, x_j, x_{i(j+1)}, \ldots, x_{id}\right)$$

- It checks the marginal functional relationship between the individual feature and the predicted target.

Individual Conditional Expectation (ICE)

- Goldstein, A., et al. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, **24**(1), 44–65.

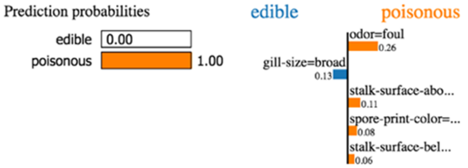


Source: Molnar (2019; Chap 5.2)

- See Molnar, C. (2019): Ch5.2 Individual Conditional Expectation (ICE)

Local Interpretability by LIME and SHAP

a) LIME



Feature	Value
odor=foul	True
gill-size=broad	True
stalk-surface-above-ring=silky	True
spore-print-color=chocolate	True
stalk-surface-below-ring=silky	True

b) SHAP

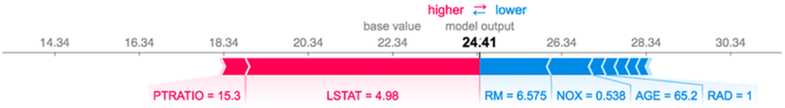


Table of Contents

How to Enhance Interpretability for Black-box Models?

"Intrinsic interpretability can only be induced from model constraints. "

Simple Statistical Models:

- Linear Regression
- Logistic Regression
- Decision Tree
- Generalized Additive Model

Interpretability Constraints:

- Additivity (or generalized additivity)
- Linearity (or piecewise linearity)
- Sparsity (principle of parsimony)
- Orthogonality (or near-orthogonality)
- Smoothness (or piecewise smoothness)
- Monotonicity (or partial monotonicity)
- Identifiability (subject to constraints)
- Prior experience and domain knowledge

How about complex models?

- In particular, Neural Networks

Thank You!

Q&A or Email ajzhang@umich.edu