The Importance of the Feature Importance

Mathematical Foundations of Machine Learning Course

Aviv Navon and Dana Kaner

September 2018

Table of contents

Random Forest and Wavelet Decomposition

The Classic Random Forest [Breiman,1996]

$T = \{x_i, f(x_i) = y_i\}_{i=1}^{N} \in (\Omega_0 \in \mathbb{R}^n, \mathbb{R})$.

- Draw B bootstrap samples $\{T_j\}_{j=1}^{B}$ of $0 < P < 100$ percents of T
- Fit a decision tree for each one, consider m covariates in each split
- Average all tree predictions

Denote $\Omega \subseteq \Omega_0$, $c_\Omega = \frac{1}{\#\{x_i \in \Omega\}} \sum_{x_i \in \Omega} f(x_i)$. The next split:

$$\min_{\Omega' \cup \Omega'' = \Omega} \sum_{x_i \in \Omega'} \left(f(x_i) - c_{\Omega'}\right)^2 + \sum_{x_i \in \Omega''} \left(f(x_i) - c_{\Omega''}\right)^2$$

Splitting Criteria Equivalence [Elisha and Dekel, 2016]

$$\max_{\Omega' \cup \Omega'' = \Omega} \|\psi_{\Omega'}\|_2^2 + \|\psi_{\Omega''}\|_2^2$$

where $\psi_{\Omega'} = I_{\Omega'}(c_{\Omega'} - c_\Omega)$ and $\|\psi_{\Omega'}\|_2^2 = \|c_{\Omega'} - c_\Omega\|_2^2 \# \left\{x_i \in \Omega'\right\}$.

The Wavelet Representation of Random Forest

$$\tilde{f}(x) = \frac{1}{B} \sum_{j=1}^{B} \sum_{\Omega \in t_j} \psi_{\Omega}(x)$$

A "pruned" representation of the ensemble [Elisha and Dekel, 2016]:

$$\|\psi_{\Omega_{k1}}\|_2 \geq \|\psi_{\Omega_{k2}}\|_2 \geq \|\psi_{\Omega_{k3}}\|_2 \dots$$

Final prediction: $\hat{f}_M(x) = \frac{1}{B} \sum_{m=1}^{M} \psi_{k_m}(x)$.

Representation Size Selection

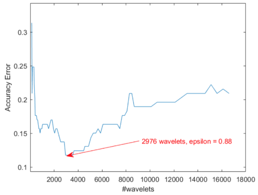


Figure 1: Illustration of the choice of M on the validation set [Elisha and Dekel]

Approaches of Feature Importance

Approaches of Feature Importance

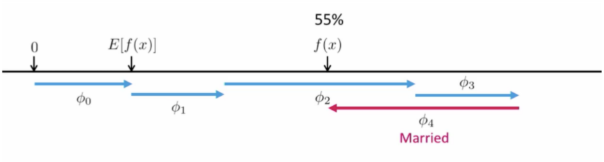
Global Feature Importance of feature i:

- **Gain:** Reduction of loss/impurity contributed by all splits by i
- **Split Count:** Number of times feature i was used to split the trees
- **Permutation:** Randomly permute the values of i in the test set
- **Wavelet decomposition based:** choose $\epsilon = \|\psi_{k_{M^*}}(x)\|_2$

$$s_i^\tau = \frac{1}{B}\sum_{j=1}^{B} \sum_{\Omega \in t_j \wedge v_i, \|\psi_\Omega\| \geq \epsilon} \|\psi_\Omega\|_2^\tau$$

Global&local Feature Importance of feature i:

- **SHAP:** Given a feature i and an observation x, $E[f(x) \mid x_i]$.
  The global FI is the mean of absolute values of the SHAP values.



Modeling Methodology

Modeling Methodology

Model parameters are chosen using 5-fold CV over a grid of parameters.

We consider the following Feature Importance methods:

- Gain on RF and GB
- SHAP on RF and GB
- Permutation FI on RF
- Wavelet-based FI

We compare the top k features selection using varied models:

- **Regression:** SVM, RF, linear regression
- **Classification:** SVM, RF, logistic regression

Regression Task

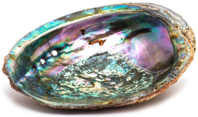
Regression Task

Abalone Dataset

Abalone's age (number of rings) $\approx$ physical measurements

Data description - number of rings with respect to:

- Sex
- Length
- Diameter
- Height
- Whole weight
- Shucked weight
- Viscera weight
- Shell weight



Classification Task

Classification Task

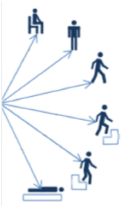
Human Activity Recognition Data

HAR $\approx$ accelerometer and gyroscope data

The data is originated in recordings of 30 subjects performing activities of daily living.

HAR classes:

- walking
- walking upstairs
- walking downstairs
- sitting
- standing
- laying



Results

Results

In order to further investigate the results, use the following **link** and explore our Shiny app.

For an HTML view of the results, use this **link**, or visit the github repo using this **link** for the full code.