Top 25 Articles from NIPS 2017

Critique: A Unified Approach to Interpreting Model Predictions (2017)

Scott M. Lundberg, Su-In Lee

Presented by: Ronghao Zhang, Wenting Song

Procedure Clarity on DeepSHAP (DeepLIFT + SHAP)

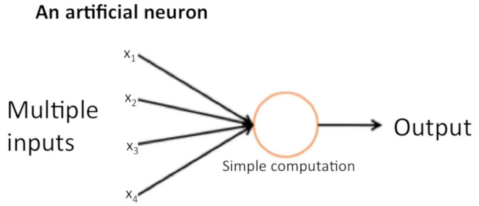
Background and procedures been presented not in enough detail to enable a reader to duplicate the connection between DeepLIFT and SHAP Value.

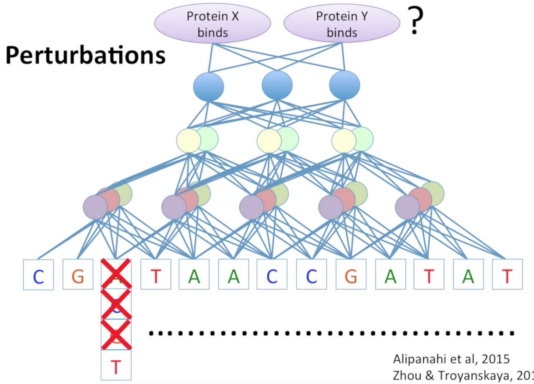
Missing Background Information:

- Neural network basics (layers, neurons and activation)
- DeepLIFT (Deep Learning Important FeaTures)

Computational performance improvements

- Previous approaches to identify important inputs in Deep Models (Perturbations vs DeepSHAP)





Linear SHAP

Let f be the original prediction model to be explained and g the explanation model.

$$\textit{Given a linear model } f(x) = \sum_{j=1}^{M} w_j x_j + b\textit{: } \phi_0(f, x) = b \textit{ and}$$

$$\phi_i(f, x) = w_j(x_j - E[x_j])$$

when the original model is already interpretable, i.e. the features **xj** are 0/1, the method should return the same model. However it seems from Corollary 1 this is not the case since **E[xj]** is nonzero.

Definition 1 Additive feature attribution methods *have an explanation model that is a linear function of binary variables:*

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \qquad (1)$$

where $z' \in \{0, 1\}^M$, M *is the number of simplified input features, and* $\phi_i \in \mathbb{R}$.

SHAP values Shortcomings

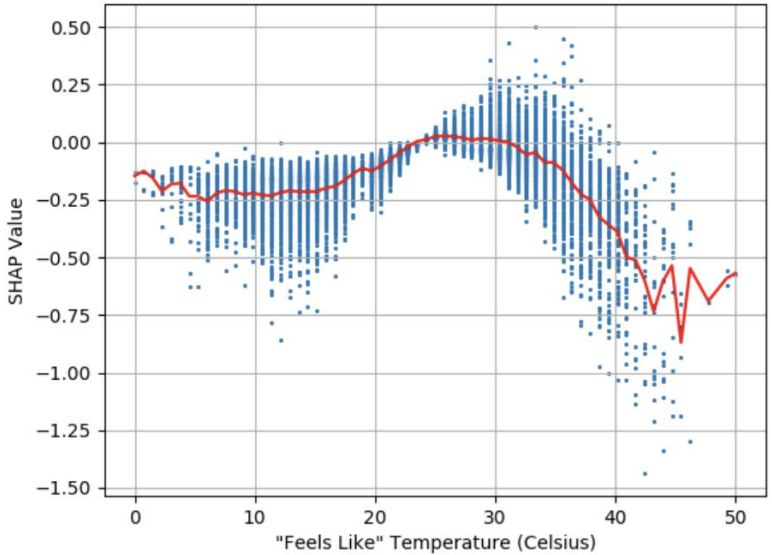
- Sensitive to high correlations among different features
  - Impact can be split in infinite number of ways
  - dividing impacts this way makes them look less important than if their impacts remained undivided.
- represent a descriptive approximation of the predictive model
  - cannot determine based on the SHAP values alone what the impact of this intervention will be.

Local -> Global?

- SHAP is local approximation method. It explains individual predictions by learning simple local approximations of a model around particular data points.

- global explanations describe the overall behavior of a model.
  - Fidelity(how well explanation matches predictions)
  - Accuracy(how well explanation predicts the origina

    label)



Tan, Sarah, et al. "Learning Global Additive Explanations for Neural Nets Using Model Distillation." arXiv preprint arXiv:1801.08640 (2018).

Useful links:

Scott M Lundberg's Website:

http://scottlundberg.com/

YouTube Introduction:

https://www.youtube.com/watch?v=wjd1G5bu_TY

Previous Comments and Critiques

https://media.nips.cc/nipsbooks/nipspapers/paper_files/nips30/reviews/2493.html

Top 25 Research and Papers in NIPS 2017:

https://www.twosigma.com/insights/article/25-of-our-favorite-papers-talks-presentations-and-workshops-from-nips-2017/