OrdShap: Feature Position Importance for Sequential Black-Box Models

Davin Hill, Brian L. Hill, Aria Masoomi, Vijay S. Nori, Robert E. Tillman & Jennifer Dy

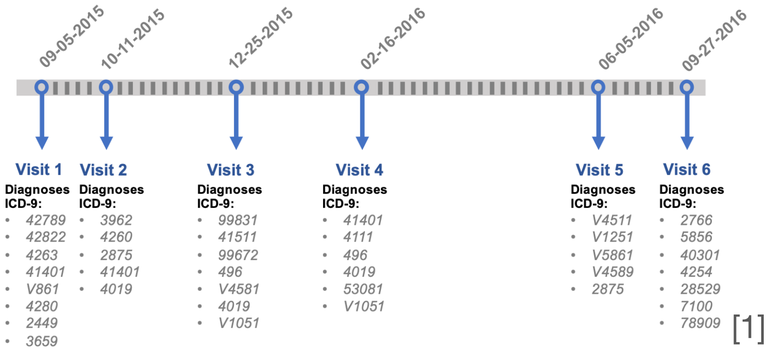
Northeastern University

Optum

Making Predictions with Sequential Models

Data

Longitudinal Data



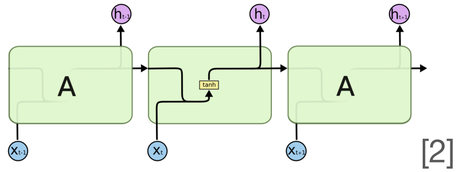
[1]

Natural Language

The	movie	was	surprisingly	bad

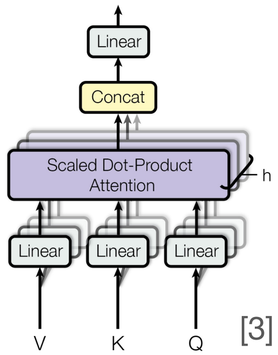
Models

RNN



[2]

Transformers



[3]

Tasks

Hospital Length-of-Stay

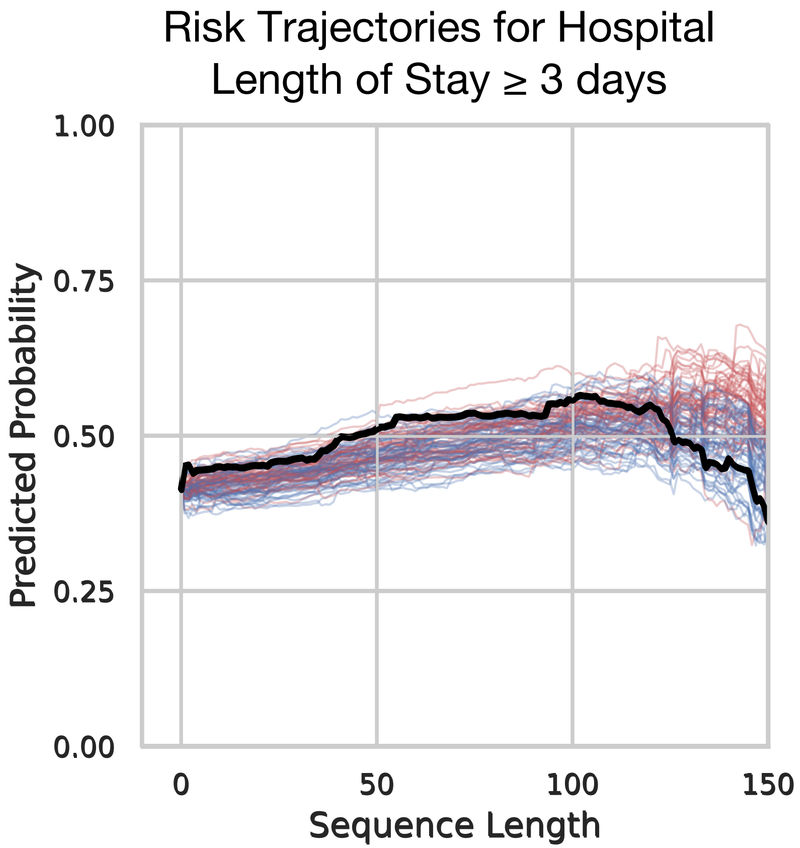
Mortality

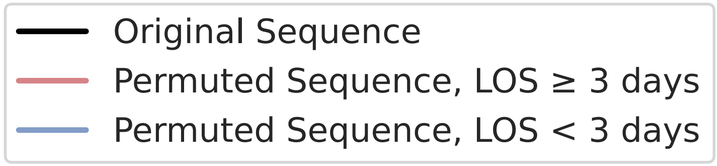
Sentiment Analysis

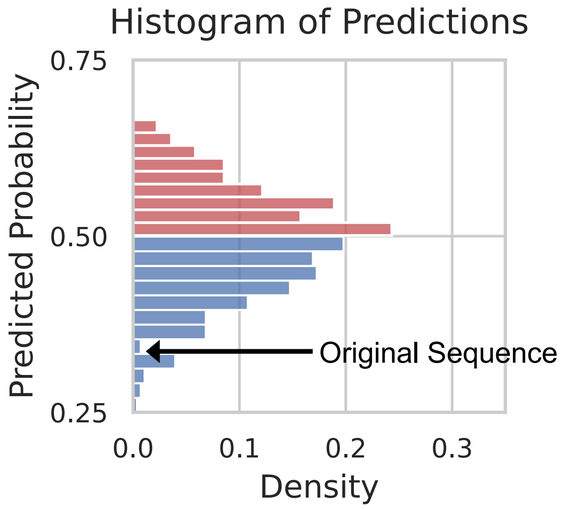
Purchase Intent

[1] Baytas et al.,Patient Subtyping via Time-Aware LSTM Networks. [2] https://colah.github.io/posts/2015-08-Understanding-LSTMs/ [3] Vaswani et al., Attention is all you need.

Permuting Ordering of Data Samples







Existing Attributions do not capture feature ordering, however permuting order significantly affects model output

Overview of OrdShap

Challenges

- How do we distentangle importance due to **1) Feature Value**, and **2) Feature Position**?
- How do we efficiently approximate this attribution?

Contributions

- We propose **OrdShap**, an attribution for Sequential models
- We establish a theoretical connection between OrdShap and Sanchez-Bergantinos Values
- We propose 2 algorithms to approximate OrdShap

Quantifying Positional Dependency

Average over Subsets

Average over Permutations

$$\gamma_{i,\ell}(N,\tilde{\omega}) = \sum_{\substack{S\subseteq N \\ i\in S}} \sum_{\substack{\sigma\in\mathfrak{S}_N \\ \sigma^{-1}(i)=\ell}} \frac{(|S|-1)!(|N|-|S|)!}{(|N|-1)!|N|!} \left[\tilde{\omega}(S,\sigma) - \tilde{\omega}(S\setminus\{i\},\sigma)\right]$$

Conditioned on Permuting $i \to \ell$

Weighting

Marginal decrease from removing i

$\mathfrak{S}_S$ Symmetric Group on S = {1,…|S|}

ω Characteristic function on Permutations

Toy Example: Position and Value Importance

Combined Value = 0

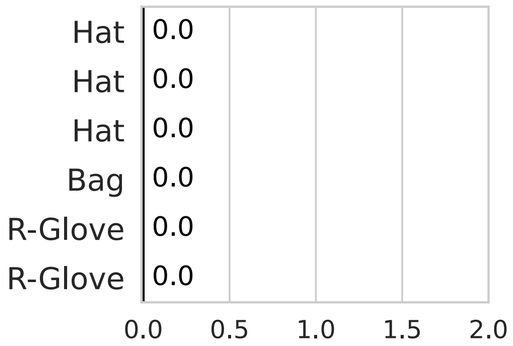
1. Value Importance

$$\bar{\gamma}_i(N,\tilde{\omega}) = \frac{1}{|N|}\sum_{\ell \in N} \gamma_{i,\ell}(N,\tilde{\omega})$$

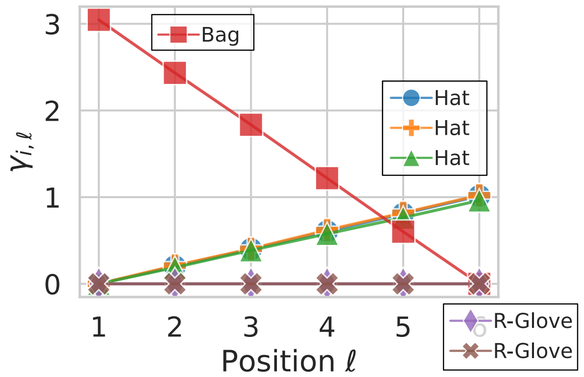
2. Position Importance

$$\arg\min_{\beta_i} \left[(\ell - \bar{\ell})\beta_i - (\gamma_{i,\ell}(N,v) - \bar{\gamma}_i)\right]^2$$

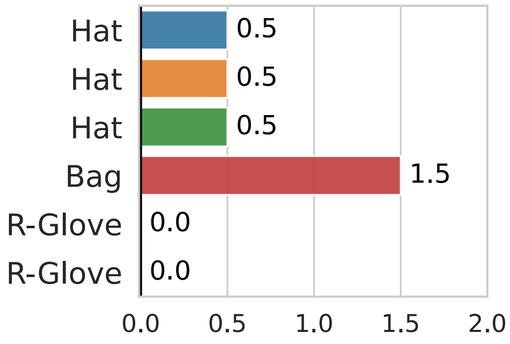
Shapley Values



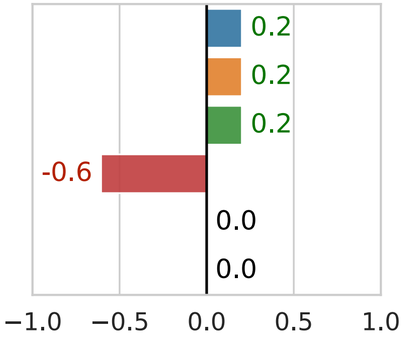
OrdShap Values



OrdShap-VI



OrdShap-PI



Positive OrdShap-PI: More important when later in the sequence

Theorem 1: OrdShap-VI fulfills Shapley Axioms of Efficiency*, Symmetry*, Null Player*, Additivity*

OrdShap Approximation

Naïve Calculation: $\mathcal{O}(d!\,2^d\,\delta_f)$

Monte Carlo Sampling Algorithm: $\mathcal{O}(dKL\delta_f + d^2KL)$

Least-Squares Algorithm with Monte Carlo Sampling: $\mathcal{O}(KL\delta_f + d^2KL + d^3)$

d : # features

δ_f : Model Evaluation

K, L : # MC samples

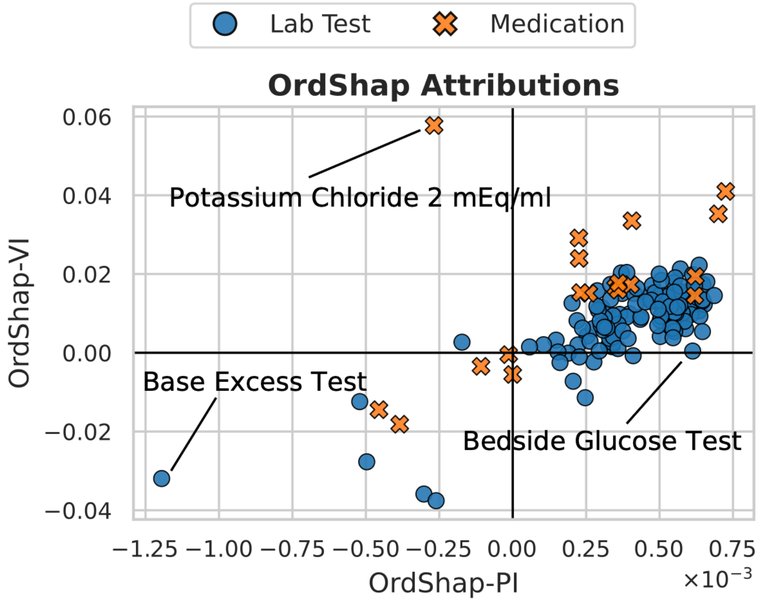
$$\min_{\alpha_1,\ldots,\alpha_n,\beta_1,\ldots,\beta_n} \sum_{\substack{S\subseteq N \\ S\neq\varnothing,N}} \sum_{\sigma\in\mathfrak{S}_N} \mu(|S|) \left[\underbrace{\sum_{i\in S}\alpha_i}_{\text{Value Importance}} + \underbrace{\sum_{i\in S}\left[\sigma^{-1}(i) - \bar{\ell}\right]\beta_i}_{\text{Position Importance}} - \underbrace{\left[\tilde{\omega}(S,\sigma) - \tilde{\omega}(\varnothing,\sigma_N^{id})\right]}_{\text{Model Output}} \right]^2$$

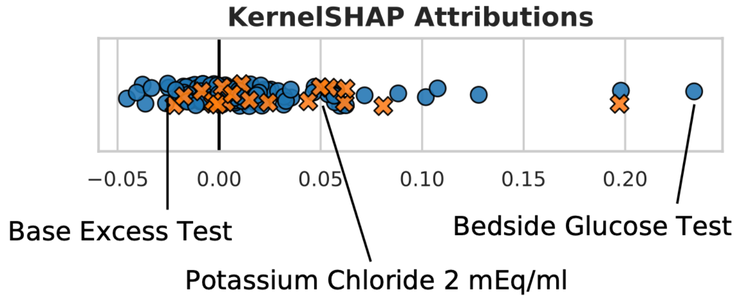
$$s.t. \quad \sum_{i\in N}\alpha_i = \frac{1}{|N|!}\sum_{\sigma\in\mathfrak{S}_N}\tilde{\omega}(N,\sigma) + \tilde{\omega}(\varnothing,\sigma_N^{id})$$

Theorem 2: Optimal α, β corresponds to OrdShap

Experiments

- Quantitative Comparisons using Inclusion / Exclusion AUC
- Evaluation on Synthetic Data
- Qualitative Examples on Medical Tokens
- Execution Time Results
- Sensitivity Analysis





Thanks for Listening!

OrdShap: Feature Position Importance for Sequential Black-Box Models

Davin Hill, Brian L. Hill, Aria Masoomi, Vijay S. Nori, Robert E. Tillman & Jennifer Dy

Northeastern University

Optum