Interpretable Machine Learning

Instance-wise Feature Selection

The movie experience was awful Explain The movie experience was awful Predict

Learning goals

- Instance-wise feature selection
- Explain then predict models
- Optimizing using explanation data



INSTANCE-WISE FEATURE SELECTION

- Select a subset of features conditioned or based on the input instance
 - Two instances might not have the same feature mask
- Instance-wise feature selection similar to feature attribution important features are selected
- Unambiguous with respect to explanation (more on that later)



Standard ML





Predict

Standard ML





Interpretable Machine Learning - 3 / 13

Standard ML





The movie experience was awful

Explain ▼



















EXPLANATION DATA







The movie experience was awful

0 0 0 1 1

EXPLANATION DATA

The movie experience was awful





$$\begin{array}{ccc} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{1} \\ \mathcal{L}_{exp} = \frac{1}{|S|} \sum_{i=1}^{|S|} |S_{t^i}| \cdot \operatorname{BCE}\left(p^i, t^i\right) \end{array}$$

EXPLANATION DATA







0 0 0 1 1
$$\mathcal{L}_{exp} = \frac{1}{|S|} \sum_{i=1}^{|S|} |S_{t^i}| \cdot BCE(p^i, t^i)$$

Predictor Model

EXPLAIN THEN PREDICT USING EXPLANATION DATA

- Selecting features conditioned on individual instances results in better task performance in comparison to global feature selection
- Instance-wise feature selection has higher inherent sparsity
- The output of the feature-selection stage is the explanation
- The predictor depends solely on the masked input and therefore unambiguous with respect to explanation

