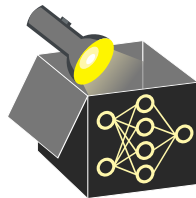


Interpretable Machine Learning

Instance-wise Feature Selection



The movie experience was awful



The movie experience **was awful**

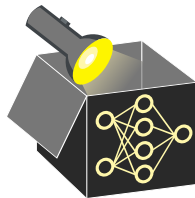


Learning goals

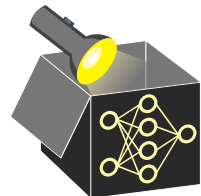
- Instance-wise feature selection
- Explain then predict models
- Optimizing using explanation data

INSTANCE-WISE FEATURE SELECTION

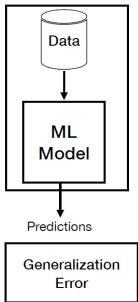
- Select a subset of features conditioned or based on the input instance
 - Two instances might not have the same feature mask
- Instance-wise feature selection similar to feature attribution — important features are selected
- Unambiguous with respect to explanation (more on that later)



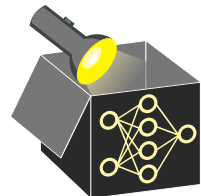
INSTANCE-WISE FEATURE SELECTION - EXAMPLE IN LANGUAGE



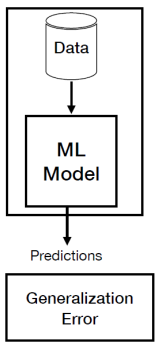
Standard ML



INSTANCE-WISE FEATURE SELECTION - EXAMPLE IN LANGUAGE



Standard ML

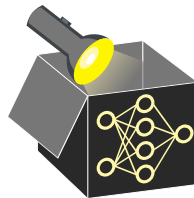


The movie experience was awful

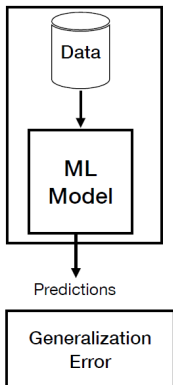


Predict

INSTANCE-WISE FEATURE SELECTION - EXAMPLE IN LANGUAGE



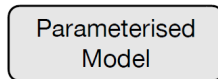
Standard ML



The movie experience was awful



Predict



Parameterised
Model

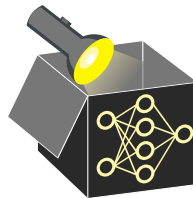
INSTANCE-WISE FEATURE SELECTION - EXAMPLE IN LANGUAGE

The movie experience was awful

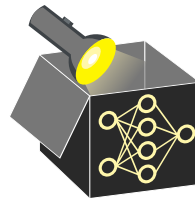


Explain

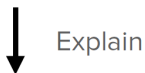
The movie experience **was awful**



INSTANCE-WISE FEATURE SELECTION - EXAMPLE IN LANGUAGE



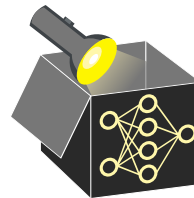
The movie experience was awful



The movie experience **was awful**



INSTANCE-WISE FEATURE SELECTION - EXAMPLE IN LANGUAGE



The movie experience was awful



Explain

Parameterised
Model

The movie experience **was awful**

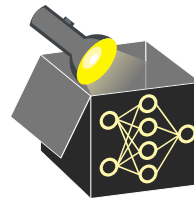


Predict

Parameterised
Model



INSTANCE-WISE FEATURE SELECTION - EXAMPLE IN LANGUAGE



The movie experience was awful



Explain

Parameterised
Model

The movie experience **was awful**



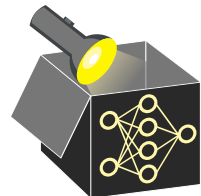
Ensure prediction is solely on the explanations

Predict

Parameterised
Model



INSTANCE-WISE FEATURE SELECTION - EXAMPLE IN LANGUAGE



The movie experience was awful



Explain

Parameterised Model

explanation

The movie experience **was awful**

Ensure prediction is solely on the explanations



Predict

Parameterised Model



EXPLANATION DATA

The movie experience was awful



Annotate



ML Model

Explain

The movie experience **was awful**

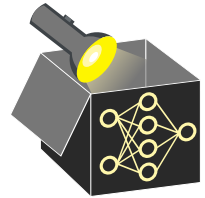
0

0

0

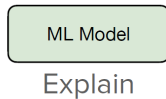
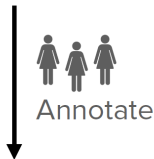
1

1



EXPLANATION DATA

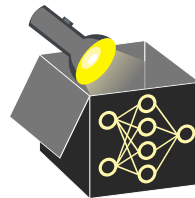
The movie experience was awful



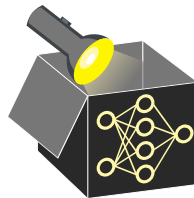
The movie experience **was awful**

0 0 0 1 1

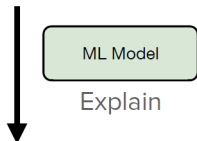
$$\mathcal{L}_{\text{exp}} = \frac{1}{|S|} \sum_{i=1}^{|S|} |S_{t^i}| \cdot \text{BCE}(p^i, t^i)$$



EXPLANATION DATA



The movie experience was awful



The movie experience **was awful**

0 0 0 1 1

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



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

