CXPlain: Causal Explanations for Model Interpretation under Uncertainty

Patrick Schwab ( @schwabpa ), and Walter Karlen ( @mhsd_ethz )
Institute of Robotics and Intelligent Systems, ETH Zurich, Switzerland

1 Introduction

Feature importance estimates inform users about the degree to which given inputs influence the output of a predictive model, and they are crucial for understanding, validating, and interpreting machine-learning models. How can we provide accurate importance scores quickly for any model, and estimate their uncertainty?

2 Causal explanations (CXPlain)

The main idea behind CXPlain is to train an explanation model to explain a given model (Figure 1). This framework has the advantage that we do not need to retrain or adapt the original model to explain its decisions.

\[ \Delta \epsilon_{X,i} = \epsilon_{X \setminus \{i\}} - \epsilon_X \]

Figure 1. A conceptual overview of causal explanation models.

To train CXPlain, we transform the task of producing feature importance estimates for a given model into a supervised learning task by using a causal objective [1, 2].

3 Usage

1 TRAIN MODEL
2 TRAIN CXPLAIN
3 EXPLAIN MODEL
4 VISUALISE RESULT

x_train, y_train = ...
x_test = ...
model = ...
model.fit(x_train, y_train)

explain = CXPlain(model, builder, masking, loss)
explain.fit(x_train, y_train)

explain(explain(x_test))

TRY IT YOURSELF AT github.com/id908b/cxplain

5 Components

Model Structure. In this work, we focus on neural explanation models. However, in principle, any supervised model could be used.

Causal objective. We use a causal objective that quantifies the marginal contribution of a feature towards the model’s accuracy [1, 2].

Masking Operation. We use a masking operation, such as zero masking [2,6], to estimate each feature's marginal contribution.

6 Conclusion

We presented CXPlain, a new method for learning to estimate feature importance for any machine-learning model. We demonstrated that CXPlain is fast at explanation time, accurate, and that we are able to estimate its attribution uncertainty using bootstrap resampling.

7 References


NeurIPS | 2019