Probabilistic Sufficient Explanations

Eric Wang **Pasha Khosravi** **Guy Van den Broeck**
ericzxwang@ucla.edu, {pashak, guyvdb}@cs.ucla.edu

Motivation

We want to generate local explanations for a given classifier.

- Logical Reasoning: Aims for 100% guarantee. Too strict, can result in complex explanations. Not always tractable to find.
- Model Agnostic: Hard to capture the dependencies between features. Generally, ignore feature distribution (can be fooled). For example, LIME and SHAP are in this category.

Proposed Solution: Choose a subset of given features and treat the rest as missing. Also want to provide some probabilistic guarantees about the outcome of the classifier while prioritizing "simpler" subsets.

Probabilistic Sufficiency

Two intuitive metrics which can be used to evaluate explanation quality.

- **Same Decision Probability** (SDP): probability the classifier will make the same decision when observing the rest of the features.

$$\text{SDP}_{\mathcal{C},\mathbf{x}}(\mathbf{z}) = \mathop{\mathbb{E}}_{\mathbf{m}\sim\Pr(\mathbf{M}|\mathbf{z})} [\![\mathcal{C}(\mathbf{zm}) = \mathcal{C}(\mathbf{x})]\!]. \qquad \Rightarrow \textit{intractable to compute [2, 1]}$$

- **Expected Prediction** (EP): how "confident" the classifier is on its decision.

$$\text{EP}(\mathbf{z}) = \mathop{\mathbb{E}}_{\mathbf{m}\sim\Pr(\mathbf{M}|\mathbf{z})} f(\mathbf{zm}). \qquad \Rightarrow \textit{for some distribution, classifier pairs [3, 4, 5]}$$

Connection between SDP and EP

$$\text{SDP}_{\mathcal{C},\mathbf{x}}(\mathbf{z}) > \frac{\text{EP}(\mathbf{z}) - T}{U(\mathbf{z}) - T}.$$

Probabilistic Sufficient Explanations

Idea: We want to provide good probabilistic guarantees while choosing a small subset of features.
We define the sufficient explanations to be

$$\text{SE}_k(\mathbf{x}) = \mathop{\text{argmax}}_{\mathbf{z}\subseteq\mathbf{x}} \text{EP}(\mathbf{z}) \quad \text{s.t. } |\mathbf{z}| \leq k$$

Out of these, we want the most likely ones:

$$\text{MLSE}_k(\mathbf{x}) = \mathop{\text{argmax}}_{\mathbf{z}\in\text{SE}_k(\mathbf{x})} \Pr(\mathbf{z})$$

Finding Sufficient Explanations

We use beam search algorithm guided by expected prediction to greedily find the subset of features that give us best guarantee. The iterative nature of beam search allows us to save explanations of different sizes.

Experiments

MNIST 3 vs 5 binary classification using decision forest classifier and probabilistic circuit for feature distribution

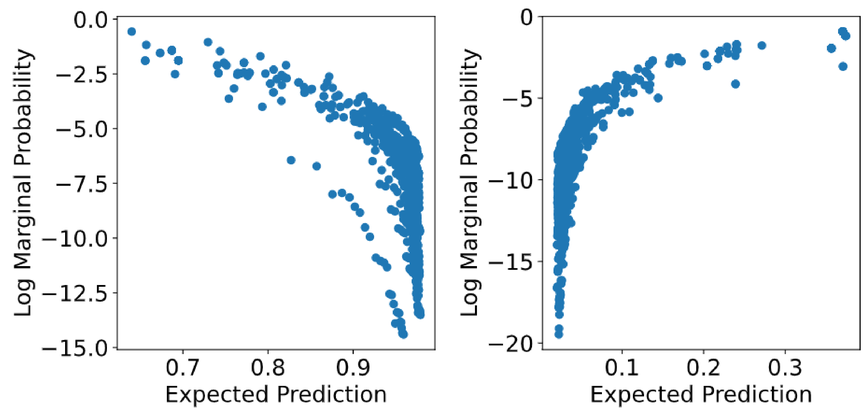
Correctly classified examples | Misclassified examples

From left to right: original image, anchors, ours (same size), ours (size 30)

- Chosen pixels mostly in upper part of image - where 3's and 5's differ
- White pixels show outline of predicted number; black pixels where the other number may be present

Method	$\lvert\text{EP}_{\mathcal{O}}(\mathbf{z})\rvert$	$\text{SDP}_{\mathcal{C},\mathbf{x}}(\mathbf{z})$
Anchors	0.75 ± 0.37	0.66 ± 0.08
MLSE_s	1.57 ± 0.29	0.86 ± 0.05
MLSE_{30}	3.75 ± 0.13	1.00 ± 0.00

$\Rightarrow$ *Our explanations have high expected predictions and high SDP*



$\Rightarrow$ *squeezing out small gains in expected prediction results in much less likely (more complex) explanations*

References

[1] Suming Jeremiah Chen, Arthur Choi, and Adnan Darwiche. "An exact algorithm for computing the same-decision probability". In: *IJCAI*. 2013.

[2] Arthur Choi, Yexiang Xue, and Adnan Darwiche. "Same-decision probability: A confidence measure for threshold-based decisions". In: *International Journal of Approximate Reasoning* (2012).

[3] Pasha Khosravi et al. "Handling Missing Data in Decision Trees: A Probabilistic Approach". In: *Artemiss Workshop at ICML*. 2020.

[4] Pasha Khosravi et al. "On Tractable Computation of Expected Predictions". In: *NeurIPS*. 2019.

[5] Pasha Khosravi et al. "What to Expect of Classifiers? Reasoning about Logistic Regression with Missing Features". In: *IJCAI*. 2019.