On Tractable Computation of Expected Predictions

Pasha Khosravi **YooJung Choi** **Yitao Liang** **Antonio Vergari** **Guy Van den Broeck**

{pashak, yjchoi, yliang, aver, guyvdb}@cs.ucla.edu

Motivation

In presence of uncertainty, one needs to probabilistically reason about the ***expected predictions*** of regressors and classifiers.

Expected Prediction appears in many interesting applications such as *handling missing values, fairness, and data analysis*.

More generally, we want to ***compute the k-th moment*** of a predictive model f w.r.t. the feature distribution p:

$$M_k(f, p) \triangleq \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[(f(\mathbf{x}))^k\right]$$

Complexity results

In general, this computation is not tractable. It is hard even for simple pairs of models naive Bayes and logistic regression [2].

We consider ***expressive models*** represented as ***probabilistic circuits*** [1]:

- f is a *regression circuit* and p is a *generative circuit* with *different vtree* $\Rightarrow$ ***proved #P-Hard!*** ☹
- f is a *classification circuit* and p is a *generative circuit even with same vtree* $\Rightarrow$ ***proved NP-Hard!*** ☹
- f is a *regression circuit* and p is a *generative circuit* with the *same vtree* $\Rightarrow$ ***polytime algorithm!*** ☺

Generative and discriminative circuits

For p we consider a generative circuit like a ***probabilistic sentential decision diagram*** (**PSDD**) [3]

$$p_n(\mathbf{x}) = \begin{cases} \mathbb{1}_n(\mathbf{x}) & \text{if } n \text{ is a leaf,} \\ p_{\mathsf{L}}(\mathbf{x}^{\mathsf{L}}) \cdot p_{\mathsf{R}}(\mathbf{x}^{\mathsf{R}}) & \text{if } n \text{ is an AND gate,} \\ \sum_{i\in \mathsf{ch}(n)} \theta_i p_i(\mathbf{x}) & \text{if } n \text{ is an OR gate.} \end{cases}$$

$\Rightarrow$ *structured decomposable, smooth*

For regression, we employ a ***regression circuit*** (**RC**), defining:

$$g_m(\mathbf{x}) = \begin{cases} 0 & \text{if } m \text{ is a leaf,} \\ g_{\mathsf{L}}(\mathbf{x}^{\mathsf{L}}) + g_{\mathsf{R}}(\mathbf{x}^{\mathsf{R}}) & \text{if } m \text{ is an AND gate,} \\ \sum_{j\in \mathsf{ch}(m)} \mathbb{1}_j(\mathbf{x})(\phi_j + g_j(\mathbf{x})) & \text{if } m \text{ is an OR gate.} \end{cases}$$

For classification, we use a ***logistic circuit*** (**LC**) [4], modeling $f(\mathbf{x}) = \gamma \circ g_r(\mathbf{x}) = 1/(1+e^{-g_r(\mathbf{x})})$.

$\Rightarrow$ *structured decomposable, smooth, deterministic*

References

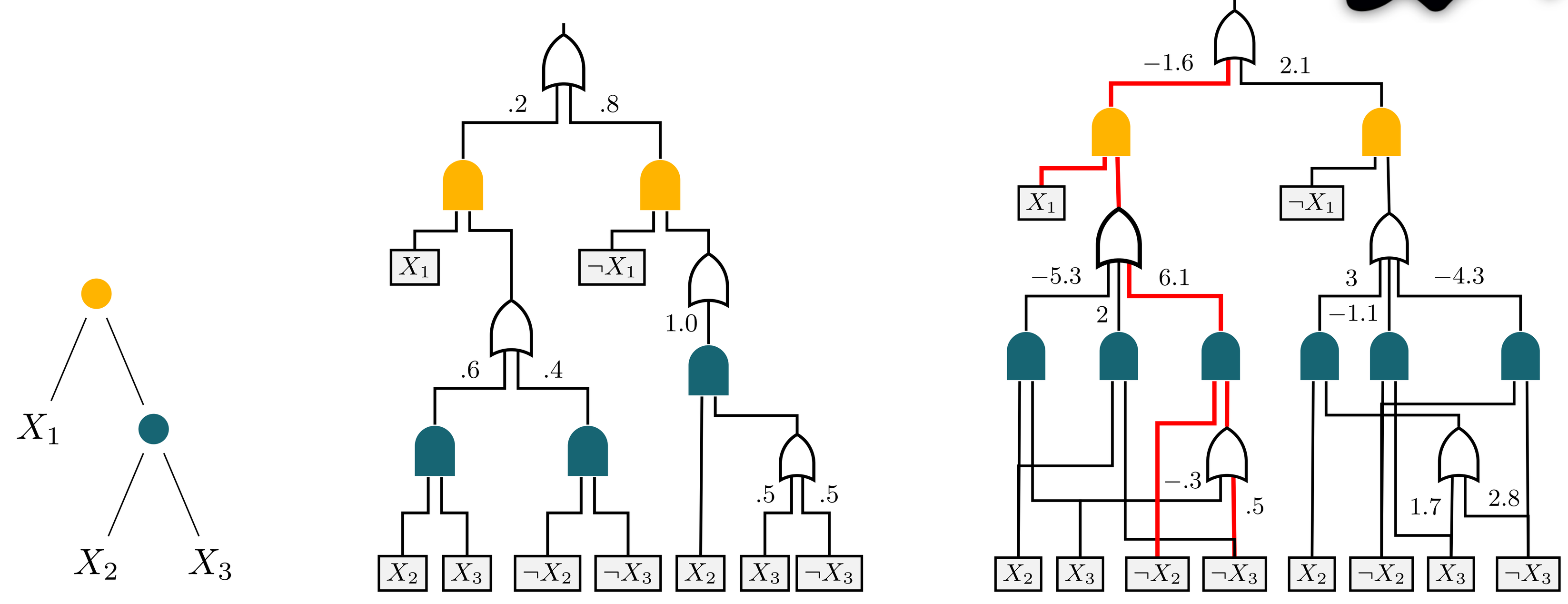
[1] Adnan Darwiche. "A Differential Approach to Inference in Bayesian Networks". In: *J.ACM* (2003).

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

[3] Doga Kisa et al. "Probabilistic sentential decision diagrams". In: *KR*. Vienna, Austria, July 2014.

[4] Yitao Liang and Guy Van den Broeck. "Learning Logistic Circuits". In: *AAAI*. 2019.

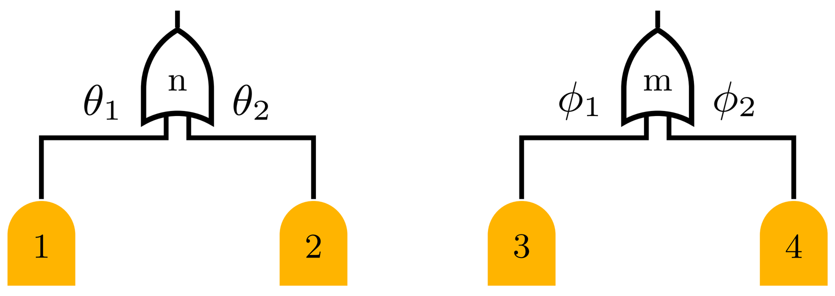
NeurIPS 2019 - Dec 2019, Vancouver, Canada



A vtree (left), and a generative (center) and discriminative circuit (right) conforming to it. Determinism is shown by a red "hot" wire.

Recursive moment decomposition

Recursively "pushes down" the computation to their children.



For example, for the pair of OR nodes (n, m) the computation involves solving subproblems $(1,3), (1,4), (2,3), (3,4)$.

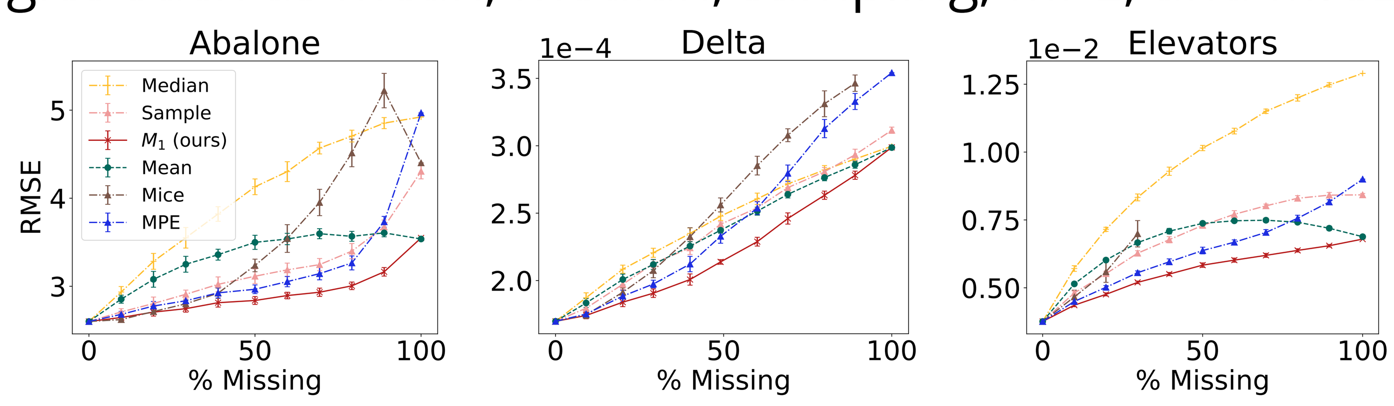
The k^{th} moments are computed ***exactly*** in $O(k^2 \cdot |p_n| \cdot |g_m|)$.

Reasoning with missing values

Given partial evidence $\mathbf{x}^o$ we want to compute

$$\mathbb{E}_{\mathbf{x}^m\sim p(\mathbf{x}^m|\mathbf{x}^o)}\left[f(\mathbf{x}^m\mathbf{x}^o)\right] = \frac{1}{p(\mathbf{x}^o)}\mathbb{E}_{\mathbf{x}^m\sim p(\mathbf{x}^m,\mathbf{x}^o)}\left[f(\mathbf{x}^m\mathbf{x}^o)\right]$$

Regression: Expected prediction outperforms many imputation strategies such as mean, median, sampling, MPE, and MICE.



Classification: We provide an approximation involving Taylor series and moments of g, which also outperforms several baselines (see paper).

Analyze behaviour of predictive models

Insurance dataset: yearly health insurance costs of people living in the USA.

Q1: "Difference of insurance costs between smokers and non smokers?"

$$M_1(f,\ p(.\,|\,\textit{Smoker})) - M_1(f,\ p(.\,|\,\textit{Non Smoker})) = 22{,}614$$

Q2: "Is the predictive model biased by gender?"

$$M_1(f,\ p(.\,|\,\textit{Female})) - M_1(f,\ p(.\,|\,\textit{Male})) = 974$$

Q3: "Expected cost of a female (F) smokers (S) with one child (C)" living in southeast (SE)? How about Std of the cost?"

$$\mu = M_1(f, p(.\,|\,\mathsf{F, S, C, SE})) = 30{,}974$$

$$\sigma = \sqrt{M_2(.) - (M_1(.))^2} = 11{,}229$$