On Tractable Computation of Expected Predictions

Pasha Khorasvi  Yoosung Choi  Yitao Liang  Antonio Vergari  Guy Van den Broeck
{pashak, yjchoi, yliang, aver, guyvdub}@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:

\[ M_k(f, p) = \mathbb{E}_{x \sim p(x)} [(f(x))^k] \]

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 the 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]

\[
p_n(x) = \begin{cases} 
1_n(x) & \text{if } n \text{ is a leaf,} \\
p_n(\chi^{1}_n) \cdot p_R(\chi^{R}_n) & \text{if } n \text{ is an AND gate,} \\
\sum_{j \in \text{ch}(n)} \theta_j p_j(x) & \text{if } n \text{ is an OR gate.} 
\end{cases}
\]

\[ g_m(x) = \begin{cases} 
0 & \text{if } m \text{ is a leaf,} \\
g_l(\chi^{1}_m) + g_R(\chi^{R}_m) & \text{if } m \text{ is an AND gate,} \\
\sum_{j \in \text{ch}(m)} I_j(x) (\phi_j + g_j(x)) & \text{if } m \text{ is an OR gate.} 
\end{cases} \]

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

\[ f(x) = \gamma \circ g(x) = 1/(1 + e^{-\gamma g(x)}) \]

\[ \Rightarrow \text{structured decomposable, smooth} \]

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 \(O(k^2 \cdot |p_u| \cdot |g_m|)\).

Reasoning with missing values

Given partial evidence \(x^o\) we want to compute

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

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?”

\[ M_1(f, p(. | \text{Smoker})) - M_1(f, p(. | \text{Non Smoker})) = 22,614 \]

Q2: “Is the predictive model biased by gender?”

\[ M_1(f, p(. | \text{Female})) - M_1(f, p(. | \text{Male})) = 974 \]

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

\[ \mu = M_1(f, p(. | F, S, C, SE)) = 30,974 \]

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

References


NeurIPS 2019 - Dec 2019, Vancouver, Canada