Analytic formulas for marginal feature attributions of oblivious decision trees

Khashayar Filom

SIAM Conference on Mathematics of Data Science
October 2024

Explaining outcomes of a complex model

Setup

- the features are random variables $\mathbf{X} = (X_1, \ldots, X_n)$;
- the input-output function of the model $\mathbf{X} \mapsto f(\mathbf{X})$;
- f can be a linear model, a random forest, a neural net etc.

Explaining outcomes of a complex model

Setup

- the features are random variables $\mathbf{X} = (X_1, \ldots, X_n)$;
- the input-output function of the model $\mathbf{X} \mapsto f(\mathbf{X})$;
- f can be a linear model, a random forest, a neural net etc.

Goal

Interpreting the model via ranking the features based on feature attributions:

- ranking globally over the whole data;
- ranking features locally, i.e. for a given input $\mathbf{x}$.

Explaining outcomes of a complex model

Setup

- the features are random variables $\mathbf{X} = (X_1, \dots, X_n)$;
- the input-output function of the model $\mathbf{X} \mapsto f(\mathbf{X})$;
- f can be a linear model, a random forest, a neural net etc.

Goal

Interpreting the model via ranking the features based on feature attributions:

- ranking globally over the whole data; ✗
- ranking features locally, i.e. for a given input $\mathbf{x}$. ✓

ECOA/Regulation B require lenders to inform applicants of the primary reasons for decline or other adverse actions.

Game-theoretic local feature attributions

Štrumbelj-Kononenko 2010, Lundberg-Lee 2017

A very popular approach is to think of features as players of a game and then quantify their contributions based on a game value.

Game-theoretic local feature attributions

Štrumbelj-Kononenko 2010, Lundberg-Lee 2017

A very popular approach is to think of features as players of a game and then quantify their contributions based on a game value.

Three ingredients:

- features $\mathbf{X}$ and model f;
- games $S \mapsto v(S; \mathbf{X}, f)(\mathbf{x})$ $(S \subseteq \{1, \ldots, n\})$ assigned to every point $\mathbf{x}$;
- game value h quantifying contribution of i^{th} feature as $h_i[v]$ at given $\mathbf{x}$.

Game-theoretic local feature attributions

Štrumbelj-Kononenko 2010, Lundberg-Lee 2017

A very popular approach is to think of features as players of a game and then quantify their contributions based on a game value.

Three ingredients:

- features $\mathbf{X}$ and model f;

  Examples) f is a piecewise constant/linear function implemented by a tree ensemble/a ReLU network.

- games $S \mapsto v(S; \mathbf{X}, f)(\mathbf{x})$ ($S \subseteq \{1, \ldots, n\}$) assigned to every point $\mathbf{x}$;

  Examples) the conditional game $S \mapsto \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S]$ "true to the data" and the marginal game $S \mapsto \mathbb{E}[f(\mathbf{x}_S, \mathbf{X}_{-S})]$ "true to the model".

- game value h quantifying contribution of i^{th} feature as $h_i[v]$ at given $\mathbf{x}$.

  Example) the Shapley value:

$$\varphi_i[v] := \sum_{S \subseteq \{1,\ldots,n\} \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \left(v(S \cup \{i\} - v(S))\right).$$

Obstacles in computing game-theoretic local feature attributions

- **Features are almost never independent.**

- **Formulas for game values usually have exponentially many terms.**

Obstacles in computing game-theoretic local feature attributions

- **Features are almost never independent.**
  Conditional feature attributions (based on game $S \mapsto \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S])$ often differ from the marginal ones (based on game $S \mapsto \mathbb{E}[f(\mathbf{x}_S, \mathbf{X}_{-S})]$).

- **Formulas for game values usually have exponentially many terms.**

Obstacles in computing game-theoretic local feature attributions

- **Features are almost never independent.**

  Conditional feature attributions (based on game $S \mapsto \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S])$ often differ from the marginal ones (based on game $S \mapsto \mathbb{E}[f(\mathbf{x}_S, \mathbf{X}_{-S})])$.

  - Remedy: Grouping features based on dependencies and using coalitional variants of the Shapley value (e.g. the Owen value) unifies the two frameworks and yields more stable explanations [Miroshnikov-Kotsiopoulos-F.-Ravi Kannan 2022].

- **Formulas for game values usually have exponentially many terms.**

Obstacles in computing game-theoretic local feature attributions

- **Features are almost never independent.**

  Conditional feature attributions (based on game $S \mapsto \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S])$ often differ from the marginal ones (based on game $S \mapsto \mathbb{E}[f(\mathbf{x}_S, \mathbf{X}_{-S})])$.

  - Remedy: Grouping features based on dependencies and using coalitional variants of the Shapley value (e.g. the Owen value) unifies the two frameworks and yields more stable explanations [Miroshnikov-Kotsiopoulos-F.-Ravi Kannan 2022].

- **Formulas for game values usually have exponentially many terms.**

  E.g. the formula for the Shapley value has 2^{n-1} terms (n can be as large as 100 for a credit card acquisition model).

Obstacles in computing game-theoretic local feature attributions

- **Features are almost never independent.**

  Conditional feature attributions (based on game $S \mapsto \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S])$ often differ from the marginal ones (based on game $S \mapsto \mathbb{E}[f(\mathbf{x}_S, \mathbf{X}_{-S})])$.

  - Remedy: Grouping features based on dependencies and using coalitional variants of the Shapley value (e.g. the Owen value) unifies the two frameworks and yields more stable explanations [Miroshnikov-Kotsiopoulos-F.-Ravi Kannan 2022].

- **Formulas for game values usually have exponentially many terms.**

  E.g. the formula for the Shapley value has 2^{n-1} terms (n can be as large as 100 for a credit card acquisition model).

  - Remedy: Monte-Carlo approximation [Štrumbelj-Kononenko 2010 & 2014], [Kotsiopoulos-Miroshnikov-F.-Ravi Kannan 2023].

Obstacles in computing game-theoretic local feature attributions

- **Features are almost never independent.**

  Conditional feature attributions (based on game $S \mapsto \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S])$ often differ from the marginal ones (based on game $S \mapsto \mathbb{E}[f(\mathbf{x}_S, \mathbf{X}_{-S})])$.

  - Remedy: Grouping features based on dependencies and using coalitional variants of the Shapley value (e.g. the Owen value) unifies the two frameworks and yields more stable explanations [Miroshnikov-Kotsiopoulos-F.-Ravi Kannan 2022].

- **Formulas for game values usually have exponentially many terms.**

  E.g. the formula for the Shapley value has 2^{n-1} terms (n can be as large as 100 for a credit card acquisition model).

  - Remedy: Monte-Carlo approximation [Štrumbelj-Kononenko 2010 & 2014], [Kotsiopoulos-Miroshnikov-F.-Ravi Kannan 2023].
  - Remedy: Focusing on a specific type of models.

Obstacles in computing game-theoretic local feature attributions

- **Features are almost never independent.**

  Conditional feature attributions (based on game $S \mapsto \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S])$ often differ from the marginal ones (based on game $S \mapsto \mathbb{E}[f(\mathbf{x}_S, \mathbf{X}_{-S})]$).

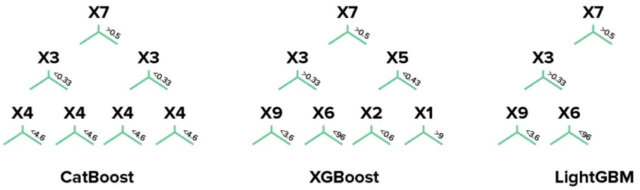
  - Remedy: Grouping features based on dependencies and using coalitional variants of the Shapley value (e.g. the Owen value) unifies the two frameworks and yields more stable explanations [Miroshnikov-Kotsiopoulos-F.-Ravi Kannan 2022].

- **Formulas for game values usually have exponentially many terms.**

  E.g. the formula for the Shapley value has 2^{n-1} terms (n can be as large as 100 for a credit card acquisition model).

  - Remedy: Monte-Carlo approximation [Štrumbelj-Kononenko 2010 & 2014], [Kotsiopoulos-Miroshnikov-F.-Ravi Kannan 2023].
  - Remedy: Focusing on a specific type of models.

Different boosting libraries construct trees differently.



Picture from Medium.

- The CatBoost library utilizes **oblivious (symmetric) decision trees** as base learners [Dorogush-Ershov-Gulin 2018].
- Despite this constraint, ensembles of symmetric trees demonstrate competitive predictive power [Ferov-Modrý 2016], [Hancock-Khoshgoftaar 2020].

Main result: a model-specific and inherently-interpretable approach

[F.-Miroshnikov-Kotsiopoulos-Ravi Kannan 2023] (10.3934/fods.2024021)

Let $\mathcal{T}$ be an ensemble of symmetric decision trees of depth $\leq d$ trained on a dataset D. (Typically, D is very large and $d \leq 10$.)

Main result: a model-specific and inherently-interpretable approach

[F.-Miroshnikov-Kotsiopoulos-Ravi Kannan 2023] (10.3934/fods.2024021)

Let $\mathcal{T}$ be an ensemble of symmetric decision trees of depth $\leq d$ trained on a dataset D. (Typically, D is very large and $d \leq 10$.)

- There is **an explicit formula** for marginal Shapley values of $\mathcal{T}$ solely in terms of the model's parameters. (In principle, it can be used to compute marginal Shapley values of any decision tree.)

Main result: a model-specific and inherently-interpretable approach

[F.-Miroshnikov-Kotsiopoulos-Ravi Kannan 2023] (10.3934/fods.2024021)

Let $\mathcal{T}$ be an ensemble of symmetric decision trees of depth $\leq d$ trained on a dataset D. (Typically, D is very large and $d \leq 10$.)

- There is **an explicit formula** for marginal Shapley values of $\mathcal{T}$ solely in terms of the model's parameters. (In principle, it can be used to compute marginal Shapley values of any decision tree.)
- The formula **can be generalized** for an axiomatically characterized family of game values (including variants of Shapley such as Banzhaf or Owen).

Main result: a model-specific and inherently-interpretable approach

[F.-Miroshnikov-Kotsiopoulos-Ravi Kannan 2023] (10.3934/fods.2024021)

Let $\mathcal{T}$ be an ensemble of symmetric decision trees of depth $\leq d$ trained on a dataset D. (Typically, D is very large and $d \leq 10$.)

- There is **an explicit formula** for marginal Shapley values of $\mathcal{T}$ solely in terms of the model's parameters. (In principle, it can be used to compute marginal Shapley values of any decision tree.)
- The formula **can be generalized** for an axiomatically characterized family of game values (including variants of Shapley such as Banzhaf or Owen).
- Based on this analytic solution, we designed **an algorithm for estimating marginal feature attributions** of $\mathcal{T}$ according to certain precomputed look-up tables.

Main result: a model-specific and inherently-interpretable approach

[F.-Miroshnikov-Kotsiopoulos-Ravi Kannan 2023] (10.3934/fods.2024021)

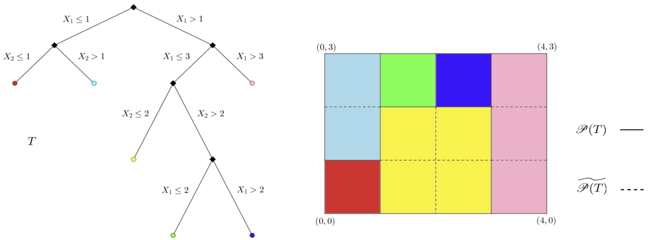
Let $\mathcal{T}$ be an ensemble of symmetric decision trees of depth $\leq d$ trained on a dataset D. (Typically, D is very large and $d \leq 10$.)

- There is **an explicit formula** for marginal Shapley values of $\mathcal{T}$ solely in terms of the model's parameters. (In principle, it can be used to compute marginal Shapley values of any decision tree.)
- The formula **can be generalized** for an axiomatically characterized family of game values (including variants of Shapley such as Banzhaf or Owen).
- Based on this analytic solution, we designed **an algorithm for estimating marginal feature attributions** of $\mathcal{T}$ according to certain precomputed look-up tables.
- The algorithm is **fast** (the computation complexity is $O(|\mathcal{T}| \cdot d)$) and **accurate** (variance of error $\propto \frac{1}{|D|}$).

What is special about oblivious (symmetric) trees?

What is special about oblivious (symmetric) trees?

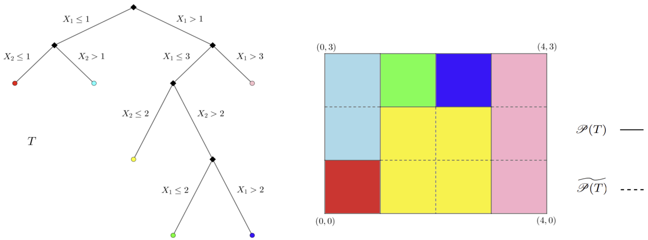
- For a tree T, marginal feature attributions based on a linear game value are piecewise constant, but only with respect to a grid partition $\widetilde{\mathscr{P}(T)}$, which is often finer than the tree's partition $\mathscr{P}(T)$. They coincide when T is symmetric.



Picture from 10.3934/fods.2024021.

What is special about oblivious (symmetric) trees?

- For a tree T, marginal feature attributions based on a linear game value are piecewise constant, but only with respect to a grid partition $\widetilde{\mathscr{P}(T)}$, which is often finer than the tree's partition $\mathscr{P}(T)$. They coincide when T is symmetric.



Picture from 10.3934/fods.2024021.

- Game value computations can be simplified by exploiting the symmetry.