Shapley Values

A Game-Theoretic Approach to Model Interpretability

The "Game" Analogy

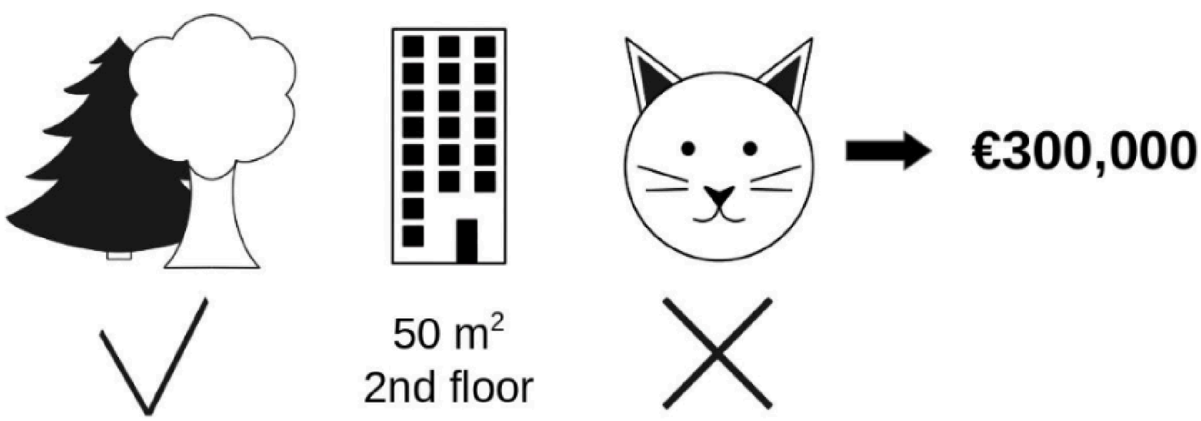
- Imagine a team of players collaborating to achieve a high score in a game. Some players contribute more than others.
- In machine learning
  - **Game:** The prediction task for a single data instance.
  - **Players:** The individual feature values of that instance.
  - **Payout:** The actual prediction minus the average.

Example

Avg prediction: € 310,000

Difference (gain): - € 10,000

How do we fairly distribute the -€10,000 "gain" among these four features?



Game Theory Definitions

- Shapley value: Assume there's a set of players $\mathcal{D} = \{1, \ldots, D\}$. And there's a subset of players $S \subseteq \mathcal{D}$. The value earned by S players is $v(S)$.
    - Now, we want to split the total payoff $v(\mathcal{D})$ among the players.
    - Shapley value $\psi_d(v)$: How much of this total payout is contributed by player d?
- Marginal contributions of player d to subset S: $C(d \mid S) = v(S \cup \{d\}) - v(S)$

Formal definition of Shapley value

- Weighted average of marginal contributions across all possible subsets that don't contain player d

$$\phi_d(v) = \sum_{S \subseteq \mathcal{D} \setminus \{d\}} \frac{1}{D\binom{D-1}{|S|}} C(d \mid S) = \sum_{S \subseteq \mathcal{D} \setminus \{d\}} \frac{1}{D\binom{D-1}{|S|}} \left[v(S \cup \{d\}) - v(S)\right]$$

- Intiutively, $\phi_d(v) = \mathbb{E}$ (contribution of d when joining a random group S)

Meaning of weight

- Denote the subset size $|S|$ as s. Without player d, there're $D-1$ players.
- Then, there're $\binom{D-1}{s}$ ways to choose S.
- The sum of all weights is

$$\sum_{s=0}^{D-1} \binom{D-1}{s} \frac{1}{D\binom{D-1}{|S|}} = \sum_{s=0}^{D-1} \frac{1}{D} = 1$$

Axioms of Shapley value

- **Efficiency:** The player contributions must add up to the total payout

$$\sum_{d=1}^{D} \phi_d(v) = v(\mathcal{D}) - v(\emptyset)$$

    - not just $v(\mathcal{D})$
- **Monotonicity:** If player d contributes at least as much in game v_1 as in v_2 for every coalition, i.e. $v_1(S \cup \{d\}) - v_1(S) \geq v_2(S \cup \{d\}) - v_2(S)\, \forall S$, then $\phi_d(v_1) \geq \phi_d(v_2)$
- **Symmetry:** The contributions of two players d_1 and d_2 should be the same if they contribute equally to all possible coalitions, i.e. $v(S \cup \{d_1\}) = v(S \cup \{d_2\})\ \forall S \Rightarrow$

$\phi_{d_1}(v) = \phi_{d_2}(v)$

- **Dummy:** A player d that does not change the predicted value – regardless of which coalition of feature values it is added to – should have a Shapley value of 0, i.e. $v(S \cup \{d\}) = v(S)\ \forall S \Rightarrow \phi_d(v) = 0$

Shapley in Machine Learning

- payout $v(\mathcal{D})$ → prediction $f(x)$
- player d → feature d
- team S → feature subsets S
- Shapley contribution $\phi_d(v)$ → $\phi_d(f, x)$

Problem

How to compute $v(S)$ s.t. $S \neq \mathcal{D}$?

Baseline imputer

- Let x' be a baseline value, eg. x' $= \mathbf{0}$ or $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_D) \in \mathbb{R}^D$.
- Then, define $v(S) = f(x_S, x'_{\bar{S}})$
  - keeping values for features in S
  - replace everything else with the baseline value
- Example
  - $x = (\text{age} = 60,\ \text{income} = \$100k)$
  - $x' = \bar{x} = (\text{age} = 40,\ \text{income} = \$50k)$
  - $S = \{\text{age}\}$
  - $x_{\text{new}} = (\text{age} = 60,\ \text{income} = \$50k)$
- Marginal contribution

$$C(d \mid S) = f(x_{S \cup \{d\}}, x'_{\overline{S \cup \{d\}}}) - f(x_S, x'_{\bar{S}})$$

  - when feature d is real V.S. when feature d is replaced by baseline
  - importance of d = prediction change when we swap x'_d → x_d
- Drawback

- no principled universal baseline
- explanations become reference-dependent
- imputed data can be unrealistic (off-manifold)

Marginal imputer

- $v(S) = \mathbb{E}_{X_{\bar{S}} \sim p(X_{\bar{S}})}[f(x_S, X_{\bar{S}})]$
  - keep the real values for features in S
  - fill the missing features with realistic values sampled from the dataset
- Assume training data is $x_1, \ldots, x_N$

$$v(S) = \frac{1}{N} \sum_{i=1}^{N} f(x_S, x_{i,\bar{S}})$$

- Drawback
  - Still, imputed data can be unrealistic (off-manifold)
  - eg. feature 1 = height, feature 2 = weight
  - If S = feature 1, $f(6'2'', 110), f(6'2'', 143), f(6'2'', 187)$

Height	Weight
4′11′′ (150 cm)	110 lb (50 kg)
5′7′′ (170 cm)	143 lb (65 kg)
6′2′′ (190 cm)	187 lb (85 kg)

Conditional imputer

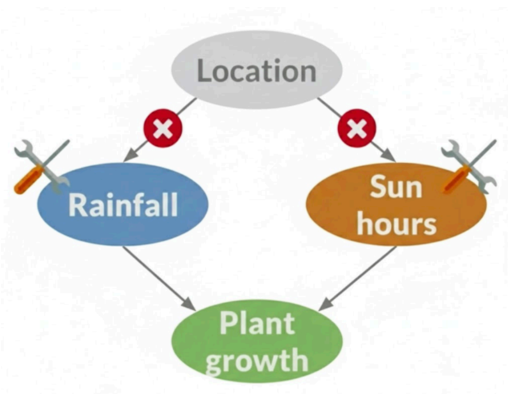
- $v(S) = \mathbb{E}_{X_{\bar{S}} \sim p(X_{\bar{S}} | X_S = x_S)}[f(x_S, X_{\bar{S}})]$
  - keep the real values for features in S
  - sample the rest in a way that is consistent with them,
    eg. weight ~ p(weight | height = 6′2′′)
- Drawback
  - the conditional distribution $p(X_{\bar{S}} \mid X_S = x_S)$ is usually unknown and hard to compute

Other imputers

- **Gaussian imputer**
  - assumption: $X \sim \mathcal{N}(\mu, \Sigma)$
  - $X_{\bar{S}} \mid X_S$ is also Gaussian with closed-form mean + covariance
  - works only for continuous features
  - fails for nonlinear relationship since $\mathbb{E}[X_{\bar{S}} \mid X_S]$ is a linear function of X_S
- **Gaussian copula imputer**
  - $(U_1, \ldots, U_d) = (F_1(X_1), \ldots, F_d(X_d))$ where F_j's are the empirical CDF
  - $(Z_1, \ldots, Z_d) = (\Phi^{-1}(U_1), \ldots, \Phi^{-1}(U_d)) \sim \mathcal{N}(\mathbf{0}, \Sigma)$ where Φ^{-1} is the standard inverse normal CDF

Causal Intervention

- **Intervention (causal)**
  - $v(S) = \mathbb{E}[f(x_S, X_{\bar{S}}) \mid do(X_S = x_S)]$
  - force $X_S = x_S$ and leave everything else unchanged
  - eg. originally, P(loc, rain, sun, growth) = P(loc) P(rain | loc) P(sun | loc) P(grow | rain, sun)
  - P′(loc, rain, sun, growth) = P(loc) P′(do(rain)) P′(do(sun)) P(grow | do(rain), do(sun))



- Marginal imputer achieves the causal intervention impact

Shapley Value Estimation

$$\phi_d(v) = \sum_{S \subseteq \mathcal{D} \setminus \{d\}} \frac{1}{D\binom{D-1}{|S|}} \, [v(S \cup \{d\}) - v(S)]$$

- **Problem**
  - For each feature d, there're 2^{D-1} such subsets to compute all subset $S \subseteq \mathcal{D} \setminus \{d\}$
  - Even worse, each term needs evaluating a value function
- **Solution:** turn SHAP into a regression problem!

KernelSHAP

- Through an additive approximation to $v_x(S)$

$$v_x(S) \approx \phi_\emptyset(f, x) + \sum_{d=1}^{D} \mathbf{1}\{d \in S\} \cdot \phi_d(f, x)$$

- KernelSHAP solves a regression problem

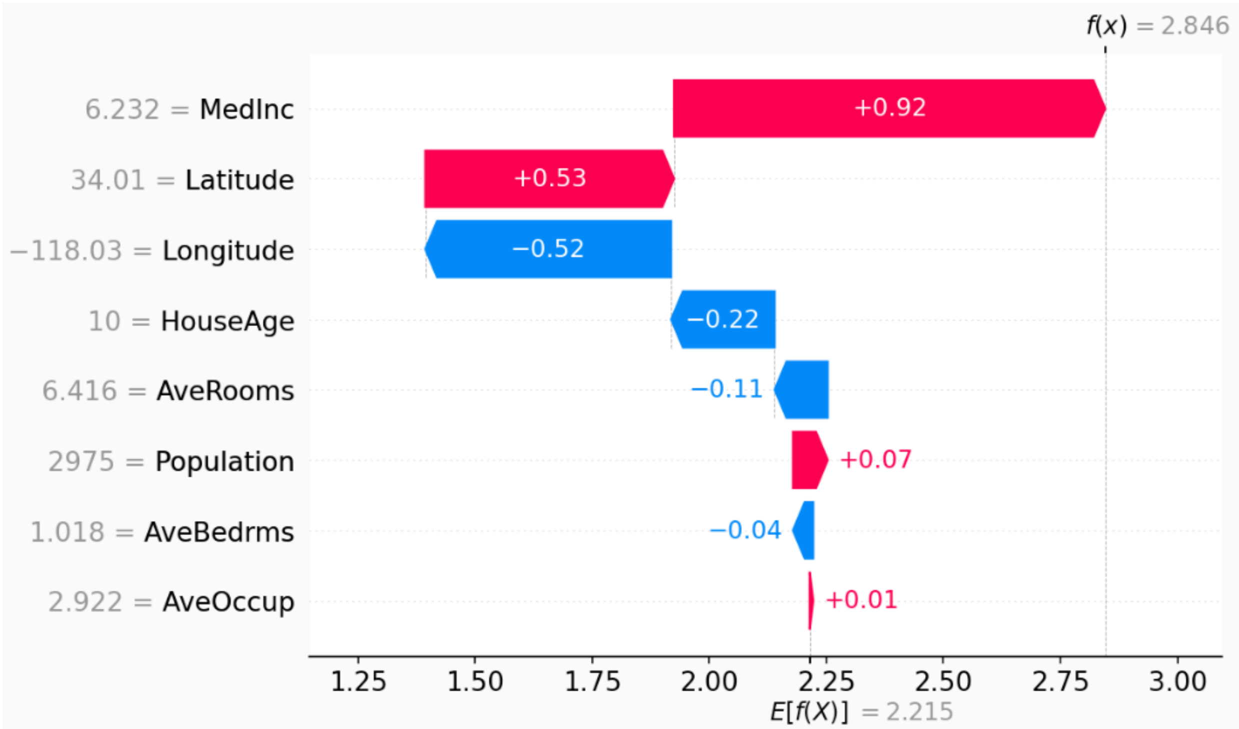
$$\phi = \arg\min_{\psi} \sum_{S \subseteq \mathcal{D}} \pi(S) \left[v_x(S) - \left(\phi_\emptyset(f, x) + \sum_{d=1}^{D} \mathbf{1}\{d \in S\} \cdot \phi_d(f, x) \right) \right]^2$$

where the unique weight is chosen as below to match the Shapley formula

$$\pi(S) = \frac{D-1}{\binom{D}{|S|} \cdot |S| \cdot (D - |S|)}$$

- Intuition of weights
  - singleton: can learn about this feature's isolated main effect on the prediction
  - all but one feature: can learn about this feature's total effect (main effect + interactions)
  - half of features: learn little about an individual feature's contribution

Visualization



Shapley Interaction

$$\phi_{ij} = \sum_{S \subseteq \mathcal{D} \setminus \{i,j\}} \frac{|S|!\,(D - |S| - 2)!}{2(D-1)!} \left[f_x(S \cup \{i,j\}) - f_x(S \cup \{i\}) - f_x(S \cup \{j\}) + f_x(S) \right]$$

→ the joint effect of features i and j co-occurring compared to the individual effects of i and j

Weight is chosen s.t. the axioms are satisfied, except for the efficiency axiom

eg. This film is not bad. → positive classification

- not: (-)
- bad: (-)
- not bad: (+)