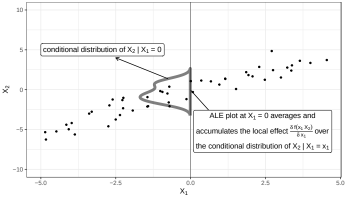
Interpretable Machine Learning

Feature Effects
Accumulated Local Effect (ALE) plot

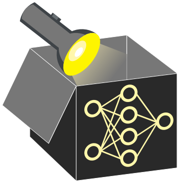


Learning goals

- Understand ALE plots
- Difference between ALE and PD plots

ACCUMULATED LOCAL EFFECTS (ALE)

▸ "Apley, Zhu" 2020

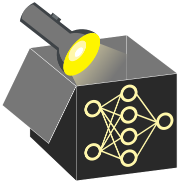


ALE plots estimate marginal effect of a feature by accumulating its local effects (integrating partial derivatives), evaluated in regions supported by the data.

Computation Steps:

1. **Estimate local effects** $\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}$ (via finite differences)
   $\Rightarrow$ Removes unwanted main effects of other features $\mathbf{x}_{-S}$ (unlike M plots)

ACCUMULATED LOCAL EFFECTS (ALE) ▸ "Apley, Zhu" 2020



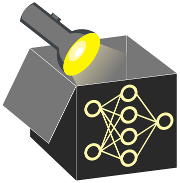
ALE plots estimate marginal effect of a feature by accumulating its local effects (integrating partial derivatives), evaluated in regions supported by the data.

Computation Steps:

1. **Estimate local effects** $\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}$ (via finite differences)
   $\Rightarrow$ Removes unwanted main effects of other features $\mathbf{x}_{-S}$ (unlike M plots)
2. **Average local effects** over conditional distr. $\mathbb{P}(\mathbf{x}_{-S}|x_S)$ similar to M plots
   $\Rightarrow$ Avoids extrapolation (unlike PD plots)

ACCUMULATED LOCAL EFFECTS (ALE)

▶ "Apley, Zhu" 2020

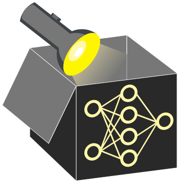


ALE plots estimate marginal effect of a feature by accumulating its local effects (integrating partial derivatives), evaluated in regions supported by the data.

Computation Steps:

1. **Estimate local effects** $\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}$ (via finite differences)
   $\Rightarrow$ Removes unwanted main effects of other features $\mathbf{x}_{-S}$ (unlike M plots)
2. **Average local effects** over conditional distr. $\mathbb{P}(\mathbf{x}_{-S}|x_S)$ similar to M plots
   $\Rightarrow$ Avoids extrapolation (unlike PD plots)
3. **Accumulate:** Integrate averaged local effects up to a specific $x \in \mathcal{X}_S$
   $\Rightarrow$ Reconstructs main effect of x_S

FIRST ORDER ALE FUNCTION

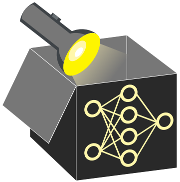


Uncentered ALE Function evaluated at $x \in \mathcal{X}_S$ (domain of feature x_S):

$$\tilde{f}_{S,\text{ALE}}(x) = \underbrace{\int_{z_0}^{x}}_{\text{(3)}} \underbrace{\mathbb{E}_{\mathbf{x}_{-S}|x_S=z_S}}_{\text{(2) average locally}} \left(\underbrace{\frac{\partial \hat{f}(z_S, \mathbf{x}_{-S})}{\partial z_S}}_{\text{(1) local effect}} \right) dz_S = \int_{z_0}^{x} \int \frac{\partial \hat{f}(z_S, \mathbf{x}_{-S})}{\partial z_S} d\mathbb{P}(\mathbf{x}_{-S} \mid z_S) dz_S$$

- x_S is feature of interest, with minimum value $z_0 = \min(x_S)$
- z_S is integration variable ranging over $\mathcal{X}_S$, used to evaluate local effects
- $\mathbf{x}_{-S}$ denotes all other features (complement of S)

FIRST ORDER ALE FUNCTION



Uncentered ALE Function evaluated at $x \in \mathcal{X}_S$ (domain of feature x_S):

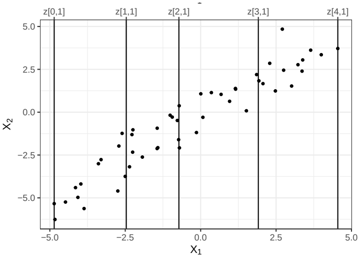
$$\tilde{f}_{S,\text{ALE}}(x) = \underbrace{\int_{z_0}^{x}}_{(3)} \underbrace{\mathbb{E}_{\mathbf{x}_{-S}|x_S=z_S}}_{\text{(2) average locally}} \left(\underbrace{\frac{\partial \hat{f}(z_S, \mathbf{x}_{-S})}{\partial z_S}}_{\text{(1) local effect}} \right) dz_S = \int_{z_0}^{x} \int \frac{\partial \hat{f}(z_S, \mathbf{x}_{-S})}{\partial z_S} d\mathbb{P}(\mathbf{x}_{-S} \mid z_S) dz_S$$

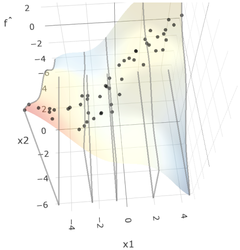
- x_S is feature of interest, with minimum value $z_0 = \min(x_S)$
- z_S is integration variable ranging over $\mathcal{X}_S$, used to evaluate local effects
- $\mathbf{x}_{-S}$ denotes all other features (complement of S)

Centering (to ensure identifiability):

$$f_{S,\text{ALE}}(x) = \tilde{f}_{S,\text{ALE}}(x) - \underbrace{\int \tilde{f}_{S,\text{ALE}}(x_S)\, d\mathbb{P}(x_S)}_{\text{constant shift to mean zero}}$$

ALE ESTIMATION: ILLUSTRATION



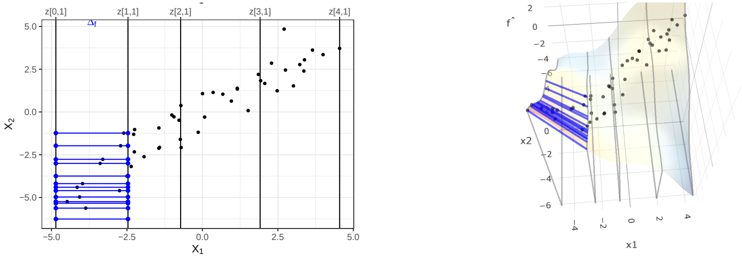


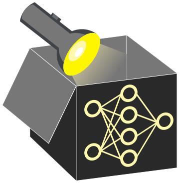
- **Motivation:** Partial derivatives are not well-defined for all models (e.g., tree-based methods). $\Rightarrow$ Use finite differences within intervals instead.
- Partition the feature range of x_S into K intervals (vertical lines)
  - Define intervals:

$$x_S \in [\min(x_S), \max(x_S)] \Rightarrow x_S \in [z_0, z_{1,S}] \cup [z_{1,S}, z_{2,S}] \cup \cdots \cup [z_{K-1,S}, z_{K,S}]$$

  - *Equidistant*: preserves resolution
  - *Quantile-based*: balances sample size per interval

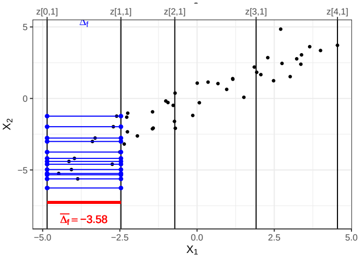
ALE ESTIMATION: ILLUSTRATION

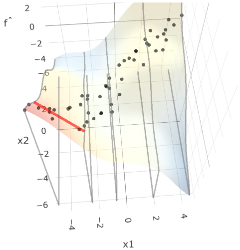


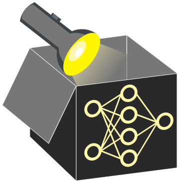


- For each observation in k-th interval, i.e., $\{i : x_S^{(i)} \in [z_{k-1,S}, z_{k,S}]\}$:
  - Replace $x_S^{(i)}$ with upper/lower interval bounds, keeping $\mathbf{x}_{-S}^{(i)}$ fixed
  - Compute obs.-wise finite difference of i-th obs. in k-th interval
    $\rightsquigarrow \hat{f}(z_{k,S}, \mathbf{x}_{-S}^{(i)}) - \hat{f}(z_{k-1,S}, \mathbf{x}_{-S}^{(i)})$ (approximates local effect)

ALE ESTIMATION: ILLUSTRATION

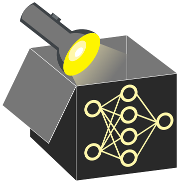






- For each observation in k-th interval, i.e., $\{i : x_S^{(i)} \in [z_{k-1,S}, z_{k,S}]\}$:
  - Replace $x_S^{(i)}$ with upper/lower interval bounds, keeping $\mathbf{x}_{-S}^{(i)}$ fixed
  - Compute obs.-wise finite difference of i-th obs. in k-th interval
    $\rightsquigarrow \hat{f}(z_{k,S}, \mathbf{x}_{-S}^{(i)}) - \hat{f}(z_{k-1,S}, \mathbf{x}_{-S}^{(i)})$ (approximates local effect)
- Average these finite differences over all observations in each interval
  $\rightsquigarrow$ Approximates **inner integral** $\mathbb{E}_{\mathbf{x}_{-S}|x_S=z_S}\left[\partial\hat{f}/\partial z_S\right]$
- Accumulate these averages from z_0 to the point of interest $x \in \mathcal{X}_S$
  $\rightsquigarrow$ Approximates **outer integral** over $z_S \in [z_0, x]$
  $\Rightarrow$ uncentered ALE function

ALE ESTIMATION: FORMULA

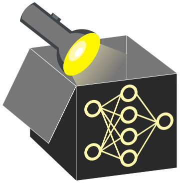


Estimated uncentered ALE: For a point $x \in \mathcal{X}_S$, define:

$$\hat{\tilde{f}}_{S,ALE}(x) = \sum_{k=1}^{k_S(x)} \frac{1}{n_S(k)} \sum_{i:\ x_S^{(i)} \in\ [z_{k-1,S}, z_{k,S}]} \left[\hat{f}(z_{k,S}, \mathbf{x}_{-S}^{(i)}) - \hat{f}(z_{k-1,S}, \mathbf{x}_{-S}^{(i)}) \right]$$

- $[z_{k-1,S}, z_{k,S}]$: k-th interval of feat. x_S with interval bounds $z_{k-1,S}$ and $z_{k,S}$
- $k_S(x)$: index of the interval in which x lies
- $n_S(k)$: number of observations in interval k
- $\mathbf{x}_{-S}^{(i)}$: all other features held fixed for i-th observation

ALE ESTIMATION: FORMULA



Estimated uncentered ALE: For a point $x \in \mathcal{X}_S$, define:

$$\hat{\hat{f}}_{S,ALE}(x) = \sum_{k=1}^{k_S(x)} \frac{1}{n_S(k)} \sum_{i:\ x_S^{(i)} \in\ [z_{k-1,S}, z_{k,S}]} \left[\hat{f}(z_{k,S}, \mathbf{x}_{-S}^{(i)}) - \hat{f}(z_{k-1,S}, \mathbf{x}_{-S}^{(i)})\right]$$

- $[z_{k-1,S}, z_{k,S}]$: k-th interval of feat. x_S with interval bounds $z_{k-1,S}$ and $z_{k,S}$
- $k_S(x)$: index of the interval in which x lies
- $n_S(k)$: number of observations in interval k
- $\mathbf{x}_{-S}^{(i)}$: all other features held fixed for i-th observation

Centering: Ensure identifiability by subtracting mean uncentered ALE (c):

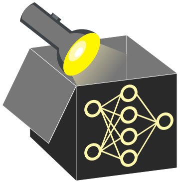
$$\hat{f}_{S,\text{ALE}}(x) = \hat{\hat{f}}_{S,\text{ALE}}(x) \ - \ c, \qquad c = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \hat{\hat{f}}_{S,\text{ALE}}\left(x_S^{(i)}\right).$$

Efficient centering (used in implementations): Use weighted trapezoidal averaging of interval-wise boundary values (avoids redundant re-evaluation at all n points):

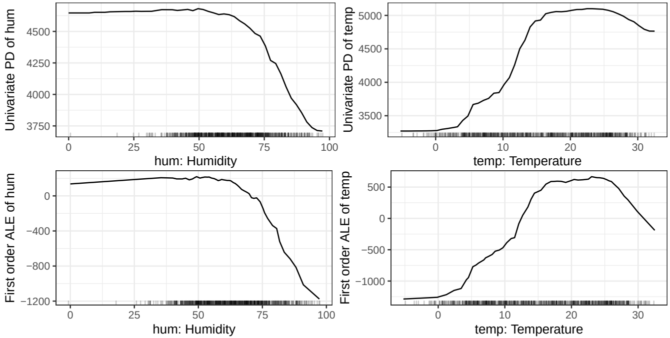
$$c = \textstyle\sum_{k=1}^{K} \tfrac{1}{2} \cdot \left(\hat{\hat{f}}_{S,\text{ALE}}(z_{k-1,S}) + \hat{\hat{f}}_{S,\text{ALE}}(z_{k,S})\right) \cdot \tfrac{n_S(k)}{n}$$

Plotting: Visualize pairs $\left(z_{k,S}, \hat{f}_{S,\text{ALE}}(z_{k,S})\right)$ for all interval boundaries $z_{k,S}$.

BIKE SHARING DATASET: FIRST ORDER ALE

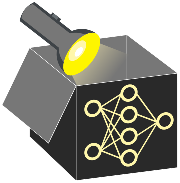


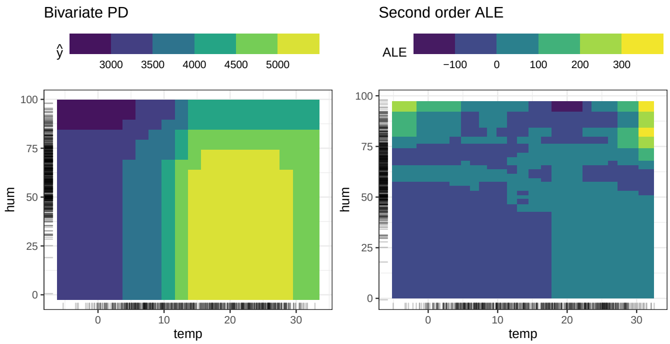
- **Visual comparison:** PD plot (top) vs. First-order ALE plot (bottom)
- **Shape:** Similar trends in both plots; y-axis scale differs due to centering
- **Interpretation:** ALE accounts for feature dependencies and avoids extrapolation into unsupported regions
  $\rightsquigarrow$ PD reflects model behavior in entire feature space ("true to the model")
  $\rightsquigarrow$ ALE focuses on effects in data-supported regions ("true to the data")



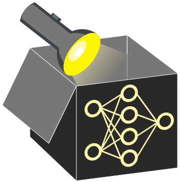
BIKE SHARING DATASET: SECOND ORDER ALE

Unlike bivariate PD plots, 2nd-order ALE plots only estimate pure interaction between two features (1st-order effects are not included).





PD VS. ALE



PD:

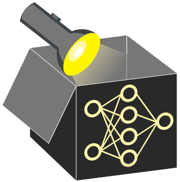
$$f_{S,PD}(x_S) = \mathbb{E}_{\mathbf{x}_{-S}}\left(\hat{f}(x_S, \mathbf{x}_{-S})\right)$$

ALE:

$$f_{S,\text{ALE}}(x) = \int_{z_0}^{x} \mathbb{E}_{\mathbf{x}_{-S}|x_S=z_S}\left(\frac{\partial \hat{f}(z_S, \mathbf{x}_{-S})}{\partial z_S}\right) dz - \text{const}$$

- Recall: PD directly averages predictions over marginal distribution of $\mathbf{x}_{-S}$
- ALE is faster: $O(2 \cdot n)$ model calls vs. $O(n \cdot g)$ for PD with g grid points
- Difference 1: ALE averages
  - prediction changes (via partial derivatives, estimated by finite differences)
  - over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S = z_S)$

PD VS. ALE



PD:

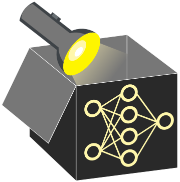
$$f_{S,PD}(x_S) = \mathbb{E}_{\mathbf{x}_{-S}}\left(\hat{f}(x_S, \mathbf{x}_{-S})\right)$$

ALE:

$$f_{S,\text{ALE}}(x) = \int_{z_0}^{x} \mathbb{E}_{\mathbf{x}_{-S}|x_S=z_S}\left(\frac{\partial \hat{f}(z_S, \mathbf{x}_{-S})}{\partial z_S}\right) dz - \text{const}$$

- Recall: PD directly averages predictions over marginal distribution of $\mathbf{x}_{-S}$
- ALE is faster: $O(2 \cdot n)$ model calls vs. $O(n \cdot g)$ for PD with g grid points
- Difference 1: ALE averages the
  - prediction changes (via partial derivatives, estimated by finite differences)
  - over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S = z_S)$
- Difference 2: ALE integrates these partial deriv. over $z_S \in [z_0, x] \subseteq \mathcal{X}_S$
  $\leadsto$ isolates effect of x_S and removes main effect of other dependent feat.

PD VS. ALE



PD:

$$f_{S,PD}(x_S) = \mathbb{E}_{\mathbf{x}_{-S}}\left(\hat{f}(x_S, \mathbf{x}_{-S})\right)$$

ALE:

$$f_{S,\text{ALE}}(x) = \int_{z_0}^{x} \mathbb{E}_{\mathbf{x}_{-S}|x_S=z_S}\left(\frac{\partial \hat{f}(z_S, \mathbf{x}_{-S})}{\partial z_S}\right) dz - \int \tilde{f}_{S,\text{ALE}}(x_S)\, d\mathbb{P}(x_S)$$

- Recall: PD directly averages predictions over marginal distribution of $\mathbf{x}_{-S}$
- ALE is faster: $O(2 \cdot n)$ model calls vs. $O(n \cdot g)$ for PD with g grid points
- Difference 1: ALE averages the
  - prediction changes (via partial derivatives, estimated by finite differences)
  - over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S = z_S)$
- Difference 2: ALE integrates these partial deriv. over $z_S \in [z_0, x] \subseteq \mathcal{X}_S$ $\rightsquigarrow$ isolates effect of x_S and removes main effect of other dependent feat.
- Difference 3: ALE is centered so that $\mathbb{E}_{x_S}\left(f_{S,ALE}(x)\right) = 0$