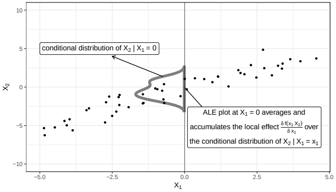
Interpretable Machine Learning

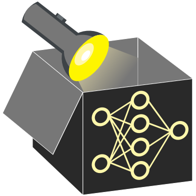
Accumulated Local Effect (ALE) plot



Learning goals

- PD plots and its extrapolation issue
- M plots and its omitted-variable bias
- Understand ALE plots

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

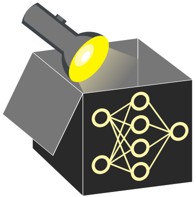


ALE plots use the idea of integrating partial derivatives. They do not suffer from the extrapolation issue of PD plots and the OVB issue of M plots when features are dependent.

Concept of ALE plots is based on

1. estimating local effects $\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}$ (via finite differences) evaluated at certain points $(x_S = z_S, \mathbf{x}_{-S})$

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

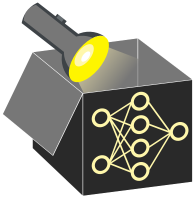


ALE plots use the idea of integrating partial derivatives. They do not suffer from the extrapolation issue of PD plots and the OVB issue of M plots when features are dependent.

Concept of ALE plots is based on

1. estimating local effects $\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}$ (via finite differences) evaluated at certain points $(x_S = z_S, \mathbf{x}_{-S})$
2. averaging local effects over conditional distribution $\mathbb{P}(\mathbf{x}_{-S} | x_S)$ similar to M plots
   $\Rightarrow$ Avoids extrapolation issue

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



ALE plots use the idea of integrating partial derivatives. They do not suffer from the extrapolation issue of PD plots and the OVB issue of M plots when features are dependent.

Concept of ALE plots is based on

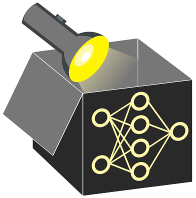
1. estimating local effects $\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}$ (via finite differences) evaluated at certain points $(x_S = z_S, \mathbf{x}_{-S})$
2. averaging local effects over conditional distribution $\mathbb{P}(\mathbf{x}_{-S} | x_S)$ similar to M plots
   $\Rightarrow$ Avoids extrapolation issue
3. integrating averaged local effects up to a specific value $x \sim \mathbb{P}(x_S)$
   $\Rightarrow$ Accumulates local effects to estimate global main effect of x_S
   $\Rightarrow$ Avoids OVB issue as other unwanted main effects were removed in (1)

FIRST ORDER ALE

- Let x_S be feature of interest with $z_0 = \min(x_S)$ and $\mathbf{x}_{-S}$ all other features (complement of S)
- Uncentered first order ALE $\tilde{f}_{S,ALE}(x)$ at feature value $x \sim \mathbb{P}(x_S)$ is defined as:

$$\tilde{f}_{S,ALE}(x) = \underbrace{\int_{z_0}^{x}}_{(3)} \underbrace{\mathbb{E}_{\mathbf{x}_{-S}|x_S}}_{(2)} \left(\underbrace{\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}}_{(1)} \middle| x_S = z_S \right) dz_S$$

FIRST ORDER ALE



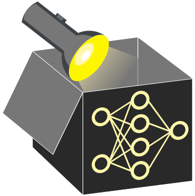
- Let x_S be feature of interest with $z_0 = \min(x_S)$ and $\mathbf{x}_{-S}$ all other features (complement of S)
- Uncentered first order ALE $\tilde{f}_{S,ALE}(x)$ at feature value $x \sim \mathbb{P}(x_S)$ is defined as:

$$\tilde{f}_{S,ALE}(x) = \underbrace{\int_{z_0}^{x}}_{(3)} \underbrace{\mathbb{E}_{\mathbf{x}_{-S}|x_S}}_{(2)} \left(\underbrace{\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}}_{(1)} \Bigg| x_S = z_S \right) dz_S$$

- Subtract average of uncentered ALE curve (constant) to obtain centered ALE curve $f_{S,ALE}(x)$ with zero mean regarding marginal distribution of feature of interest x_S:

$$f_{S,ALE}(x) = \tilde{f}_{S,ALE}(x) - \underbrace{\int_{-\infty}^{\infty} \tilde{f}_{S,ALE}(x_S)\, d\mathbb{P}(x_S)}_{:=constant}$$

ALE ESTIMATION

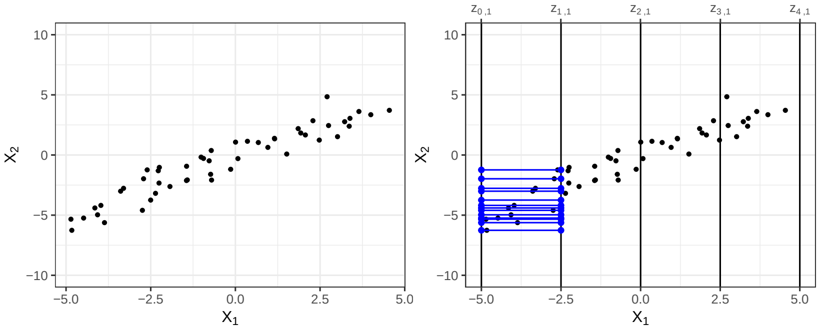


- Partial derivatives not useful for all models (e.g., tree-based methods)
- Approximate them by finite differences of predictions within K intervals for $\mathbf{x}_S$:

$$
\begin{aligned}
x \in [\min(\mathbf{x}_S), \max(\mathbf{x}_S)] \iff & \; x \in [z_{0,S}, z_{1,S}] \\
& \vee x \in \,]z_{1,S}, z_{2,S}] \\
& \dots \\
& \vee x \in \,]z_{K-1,S}, z_{K,S}]
\end{aligned}
$$

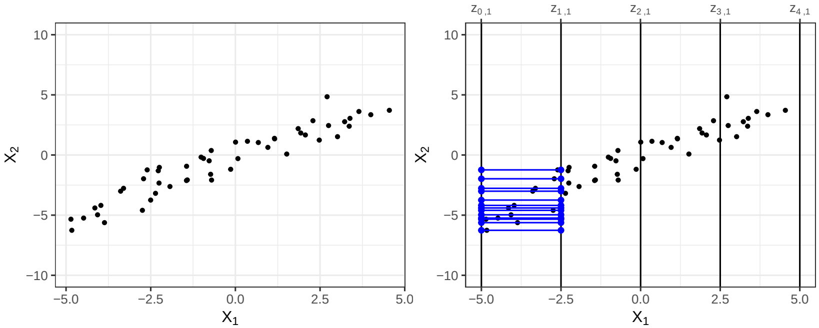
- Create K intervals for feature $\mathbf{x}_S$, e.g., using quantiles as interval bounds

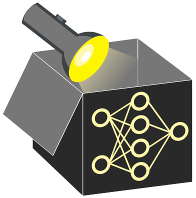
2-D ILLUSTRATION



- Divide feature of interest into intervals (vertical lines)
- For all points within an interval, compute **prediction difference** when we replace feature value with upper/lower interval bound (blue points) while keeping other feature values unchanged
- These **finite differences** (approximate local effect) are accumulated & centered $\Rightarrow$ ALE plot

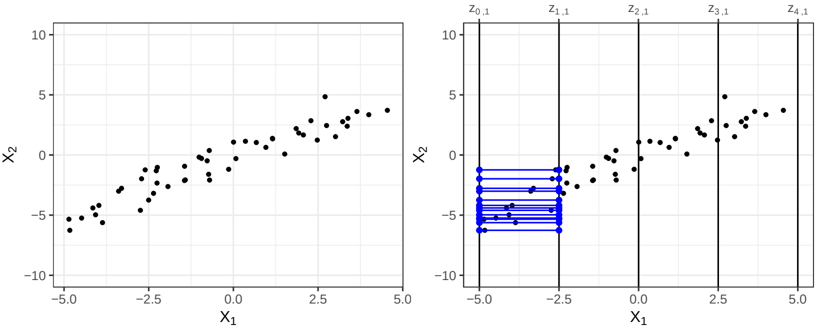
2-D ILLUSTRATION





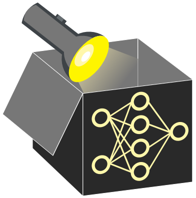
- For $\mathbf{x}^{(i)} = (x_S^{(i)}, \mathbf{x}_{-S}^{(i)})$, value $x_S^{(i)}$ is located within k-th interval of $\mathbf{x}_S$ $(x_S^{(i)} \in]z_{k-1,S}, z_{k,S}])$
- Replace $x_S^{(i)}$ by upper/lower interval bound while all other feature values $\mathbf{x}_{-S}^{(i)}$ are kept constant
- Finite differences correspond to $\hat{f}(z_{k,S}, \mathbf{x}_{-S}^{(i)}) - \hat{f}(z_{k-1,S}, \mathbf{x}_{-S}^{(i)})$

2-D ILLUSTRATION



- Estimate local effect of $\mathbf{x}_S$ within each interval by averaging all observation-wise finite differences $\hat{=}$ Approximation of inner integral that integrates over local effects w.r.t. $\mathbb{P}(\mathbf{x}_{-S}|z_S)$.
- Sum up local effects of all intervals up to point of interest $\hat{=}$ Estimates outer integral

ALE ESTIMATION: FORMULA



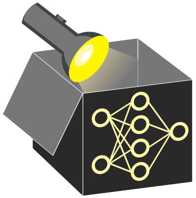
- Estimated uncentered first order ALE $\hat{\tilde{f}}_{S,ALE}(x)$ at point x:

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

- $k_S(x)$ denotes the interval index a feature value $x \in \mathbf{x}_S$ falls in
- $n_S(k)$ denotes the number of observations inside the k-th interval of $\mathbf{x}_S$
- Subtract average of estimated uncentered ALE to obtain centered ALE estimate:

$$\hat{f}_{S,ALE}(x) = \hat{\tilde{f}}_{S,ALE}(x) - \frac{1}{n} \sum_{i=1}^{n} \hat{\tilde{f}}_{S,ALE}(x_S^{(i)})$$

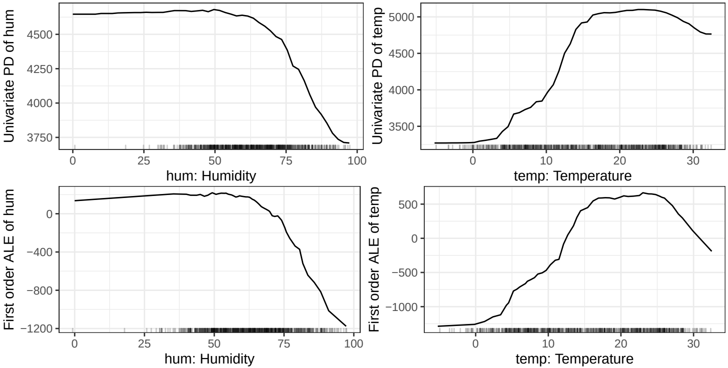
ALE ESTIMATION: ALGORITHM



1. Create K intervals for value range of $\mathbf{x}_S$
2. Repeat for each interval:
   - Replace observation's feature value $x_S^{(i)}$ with upper/lower interval bound for each observation inside k-th interval
   - Compute observation-wise finite difference inside k-th interval and average them to estimate interval-wise local effects
3. Accumulate interval-wise local effects up to value of interest x to estimate uncentered ALE and then center it

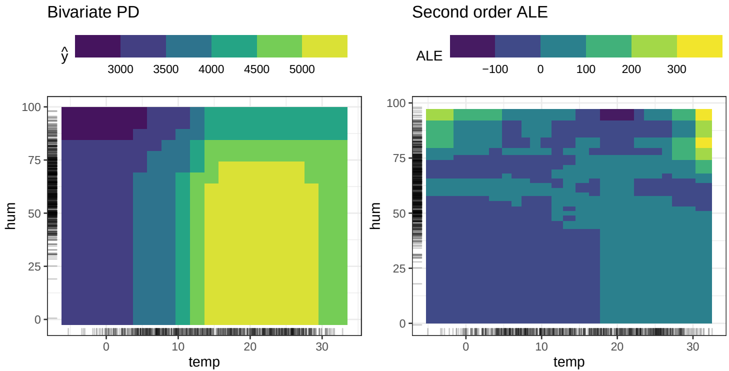
BIKE SHARING DATASET: FIRST ORDER ALE

Shape of PD plot (left) often looks similar to (centered) first order ALE plot (right) but on different y-axis scale. In case of correlated features, ALE might be better due to PD's extrapolation issue.

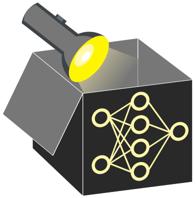


BIKE SHARING DATASET: SECOND ORDER ALE

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:

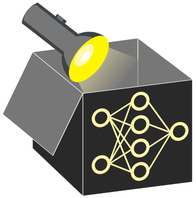
$$f_{S,PD}(x_S) = \mathbb{E}_{\mathbf{x}_{-S}}\left(\hat{f}(x_S, \mathbf{x}_{-S})\right)$$

ALE:

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

- Recall: PD directly averages predictions over marginal distribution of $\mathbf{x}_{-S}$
- Difference 1: ALE averages the
  - change of predictions (via partial derivatives approximated by finite differences)
  - over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S = z_S)$

PD VS. ALE



PD:

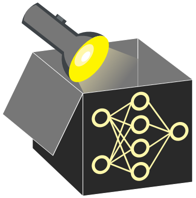
$$f_{S,PD}(x_S) = \mathbb{E}_{\mathbf{x}_{-S}}\left(\hat{f}(x_S, \mathbf{x}_{-S})\right)$$

ALE:

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

- Recall: PD directly averages predictions over marginal distribution of $\mathbf{x}_{-S}$
- Difference 1: ALE averages the
  - change of predictions (via partial derivatives approximated by finite differences)
  - over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S = z_S)$
- Difference 2: ALE integrates partial derivatives of feature S over z_S
  $\leadsto$ isolates effect of feature S and removes main effect of other dependent features

PD VS. ALE



PD:

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

ALE:

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

- Recall: PD directly averages predictions over marginal distribution of $\mathbf{x}_{-S}$
- Difference 1: ALE averages the
  - change of predictions (via partial derivatives approximated by finite differences)
  - over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S = z_S)$
- Difference 2: ALE integrates partial derivatives of feature S over z_S
  $\rightsquigarrow$ isolates effect of feature S and removes main effect of other dependent features
- Difference 3: ALE is centered so that $\mathbb{E}_{x_S}\left(f_{S,ALE}(x)\right) = 0$