Paper presentation at ACML 2022

DALE: Differential Accumulated Local Effects for efficient and accurate global explanations

Vasilis Gkolemis[1,2] Theodore Dalamagas[1] Christos Diou[2]

[1]ATHENA Research and Innovation Center

[2]Harokopio University of Athens

December 2022

Feature Effect

$y = f(x_s) \rightarrow$ plot showing the effect of x_s on the output y

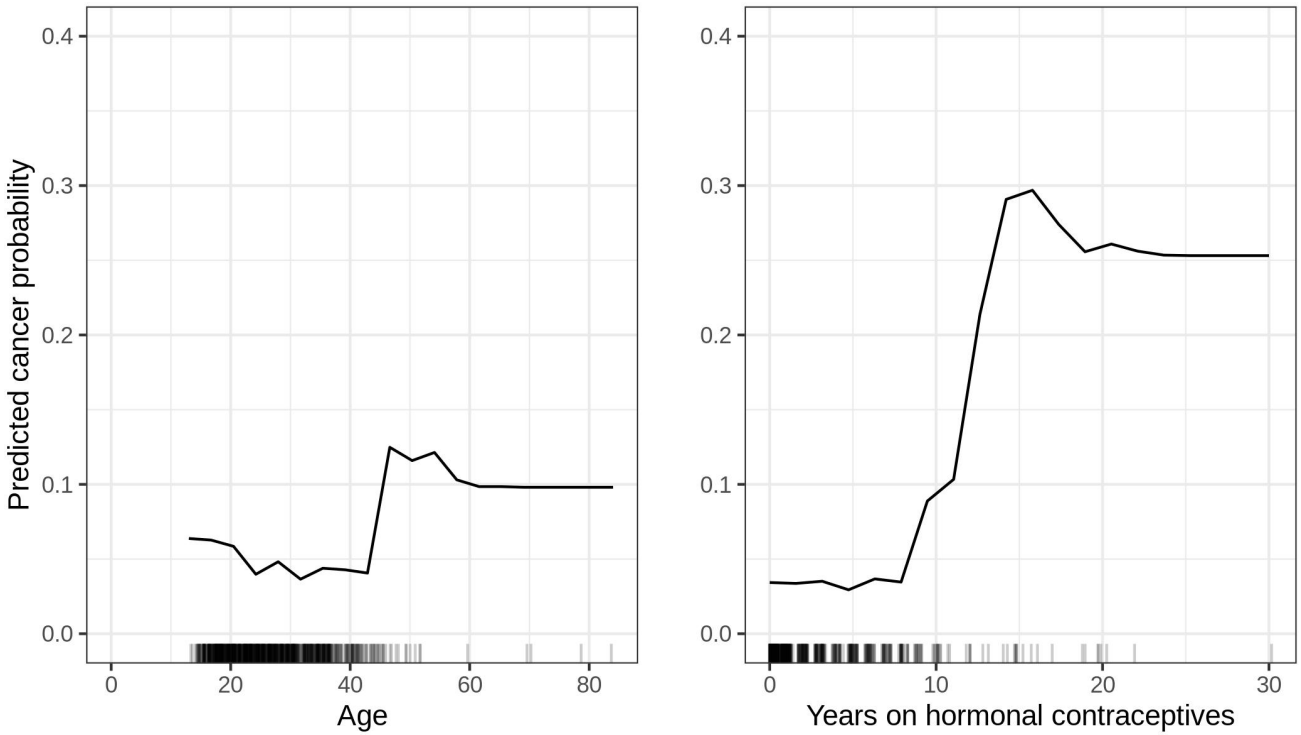


Figure: Image taken from Interpretable ML book.

Feature Effect is simple and intuitive.

Feature Effect Methods

- $x_s \rightarrow$ feature of interest, $\boldsymbol{x_c} \rightarrow$ other features
- FE methods take $(f, \mathcal{D}, s)$ and return $y = f_{<\texttt{name}>}(x_s)$
- PDP
  - Expected outcome over $\boldsymbol{x_c}$: $f(x_s) = \mathbb{E}_{\boldsymbol{x_c}}[f(x_s, \boldsymbol{x_c})]$
  - **Unrealistic instances**
- MPlot
  - Expected outcome over $\boldsymbol{x_c}|x_s$: $f(x_s) = \mathbb{E}_{\boldsymbol{x_c}|x_s}[f(x_s, \boldsymbol{x_c})]$
  - **Aggregated effects**
- ALE
  - $f(x_s) = \int_{x_{min}}^{x_s} \mathbb{E}_{\boldsymbol{x_c}|z}[\frac{\partial f}{\partial x_s}(z, \boldsymbol{x_c})]\partial z$
  - **Resolves both failure modes**

PDP vs MPlot vs ALE

DALE - Differential ALE

DALE, from the dataset $\mathcal{D} = \{\boldsymbol{x}^i, y^i\}_{i=1}^N$

$$f(x_s) = \Delta x \sum_{k}^{k_x} \underbrace{\frac{1}{|\mathcal{S}_k|} \sum_{i:\boldsymbol{x}^i \in \mathcal{S}_k} \underbrace{[\frac{\partial f}{\partial x_s}(x_s^i, \boldsymbol{x_c^i})]}_{\text{point effect}}}_{\text{bin effect}}$$

- only change point effect computation
- Fast $\rightarrow$ use of auto-differentiation, all derivatives in a single pass
- Versatile $\rightarrow$ point effects computed once, change bins without cost
- Secure $\rightarrow$ does not create artificial instances

For differentiable models, DALE resolves ALE weaknesses

DALE is faster and versatile

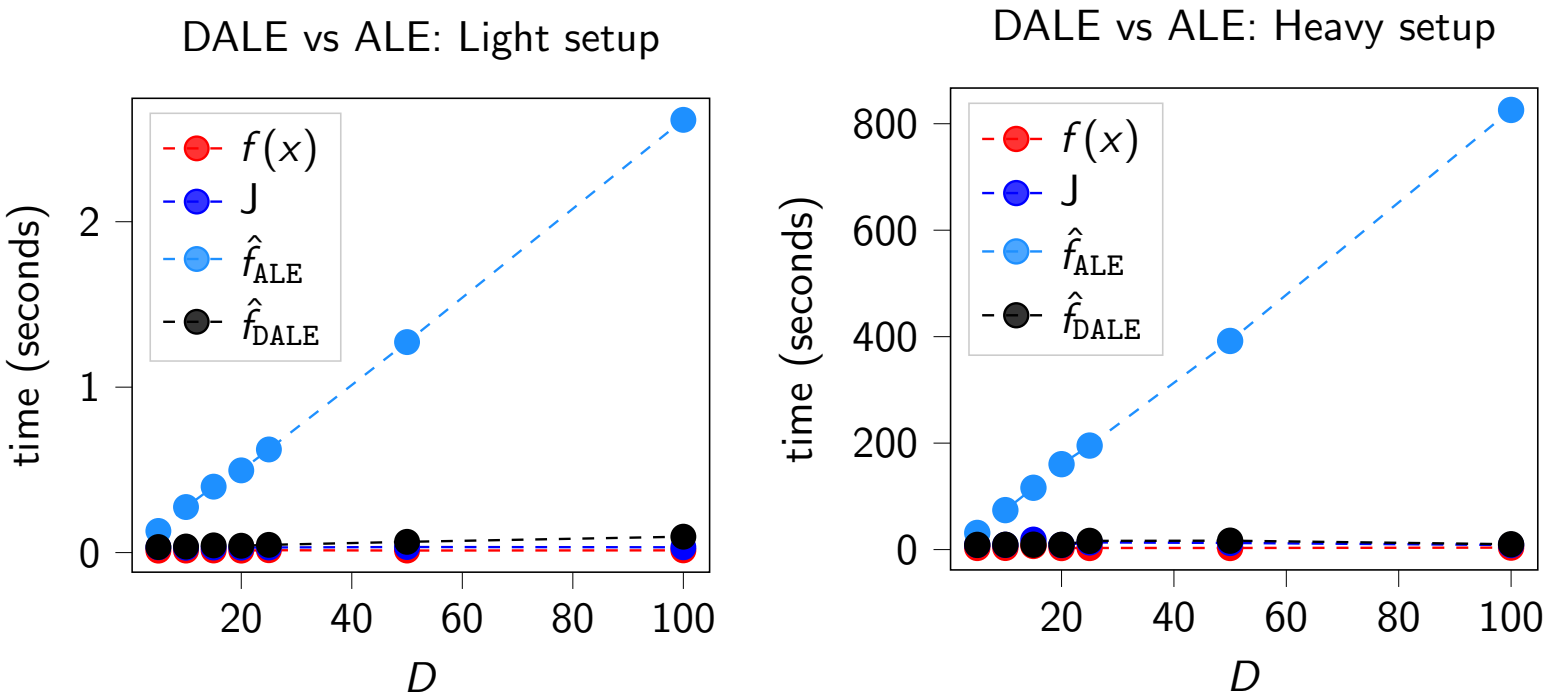
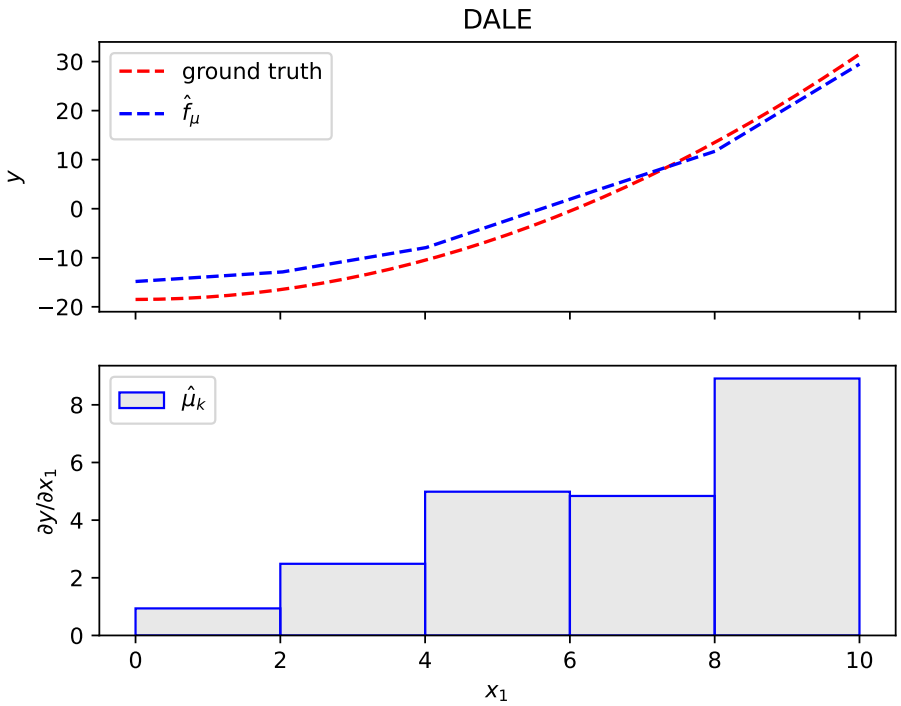
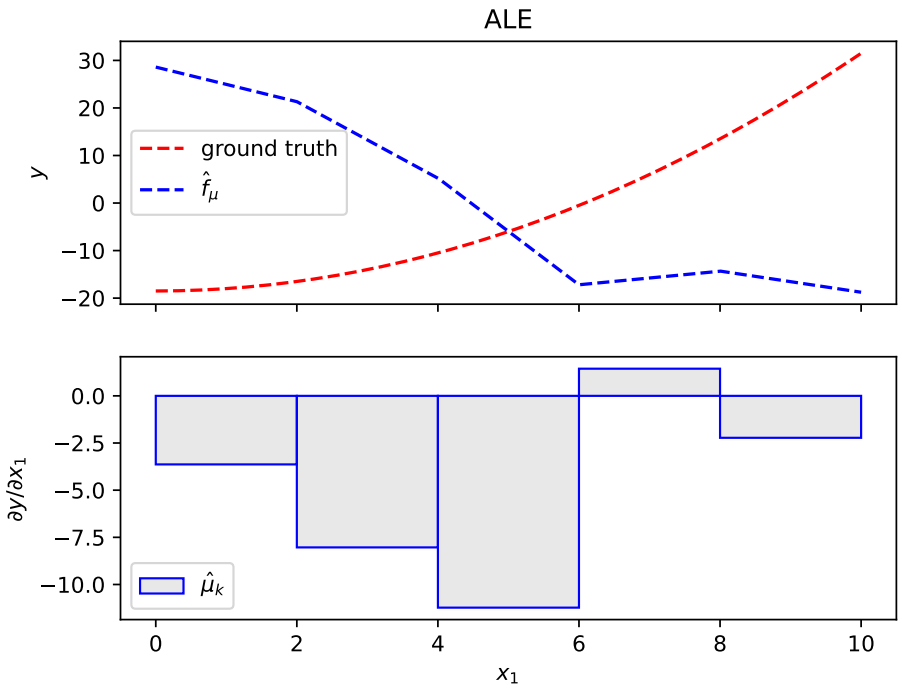


Figure: Light setup; small dataset ($N = 10^2$ instances), light f. Heavy setup; big dataset ($N = 10^5$ instances), heavy f

DALE considerably accelerates the estimation

DALE vs ALE - 5 Bins





- DALE: on-distribution, robust bin effect → good estimation
- ALE: completely OOD, robust bin effect → poor estimation