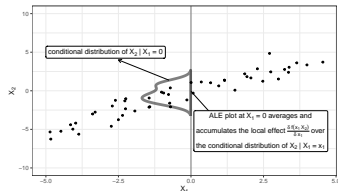


Interpretable Machine Learning

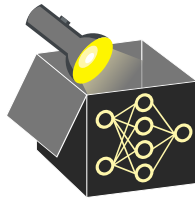
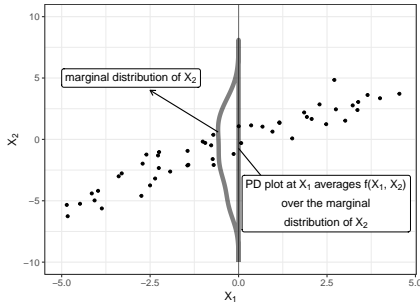
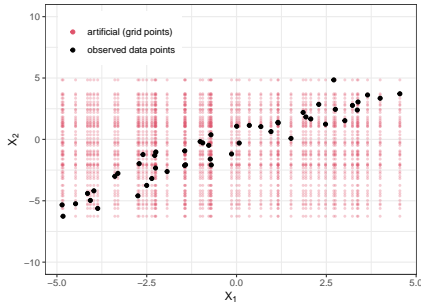
Accumulated Local Effect (ALE): Introduction



Learning goals

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

MOTIVATION - CORRELATED FEATURES

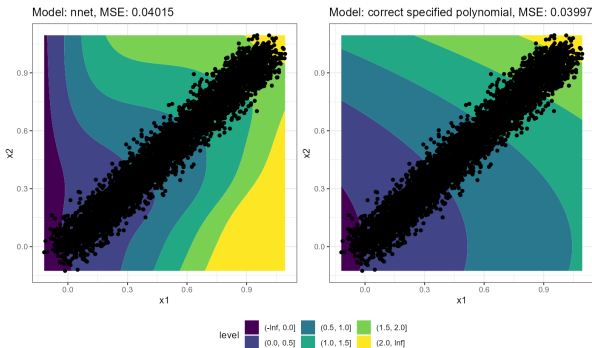


- PD plots **average over predictions** of artificial points that are out of distribution/unlikely (red)
⇒ Can lead to misleading / biased interpretations, especially if model also contains interactions
- Not wanted if interest is to interpret effects within data distribution

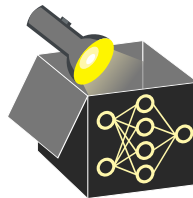
MOTIVATION - CORRELATED FEATURES

Example: Fit a NN to 5000 simulated data points with $x \sim Unif(0, 1)$, $\epsilon \sim N(0, 0.2)$ and

$$y = x_1 + x_2^2 + \epsilon, \text{ where } x_1 = x + \epsilon_1, x_2 = x + \epsilon_2 \text{ and } \epsilon_1, \epsilon_2 \sim N(0, 0.05).$$



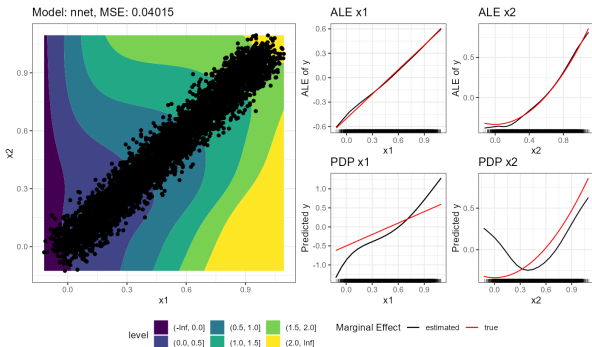
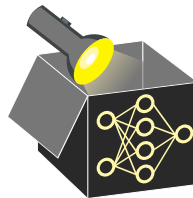
- Test error (MSE) of NN is comparable to other models
- NN contains interactions (see complex pred. surface)



MOTIVATION - CORRELATED FEATURES

Example: Fit a NN to 5000 simulated data points with $x \sim Unif(0, 1)$, $\epsilon \sim N(0, 0.2)$ and

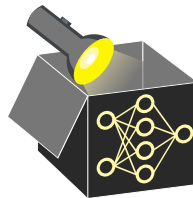
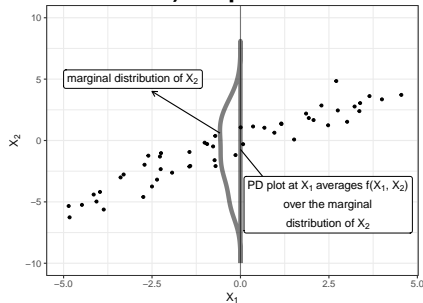
$$y = x_1 + x_2^2 + \epsilon, \text{ where } x_1 = x + \epsilon_1, x_2 = x + \epsilon_2 \text{ and } \epsilon_1, \epsilon_2 \sim N(0, 0.05).$$



- Test error (MSE) of NN is comparable to other models
- NN contains interactions (see complex pred. surface)
- ALE in line with ground truth
- PDP does not reflect ground truth effects of DGP well
⇒ Due to interactions and averaging of points outside data distribution

M PLOT VS. PD PLOT

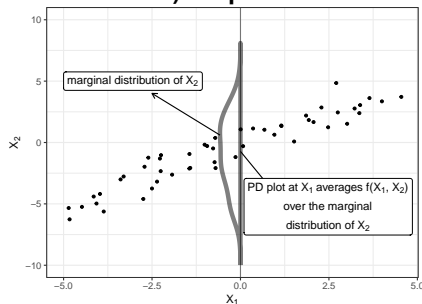
a) PD plot



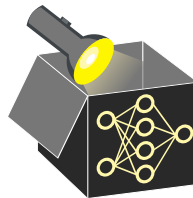
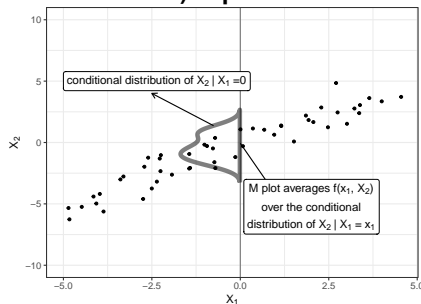
a) PD plot $\mathbb{E}_{\mathbf{x}_2} \left(\hat{f}(x_1, \mathbf{x}_2) \right)$ is estimated by $\hat{f}_{1,PD}(x_1) = \frac{1}{n} \sum_{i=1}^n \hat{f}(x_1, \mathbf{x}_2^{(i)})$

M PLOT VS. PD PLOT

a) PD plot



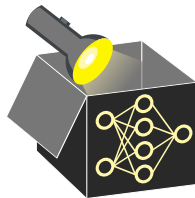
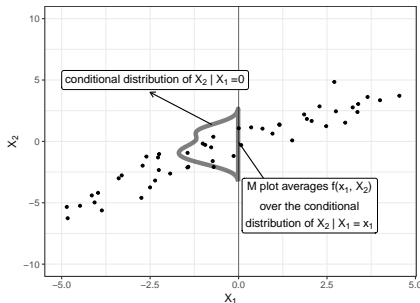
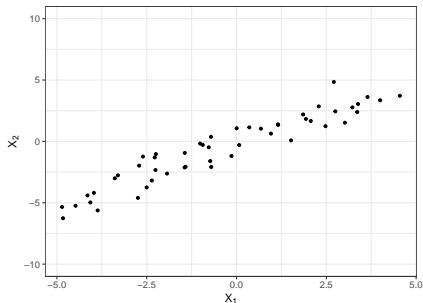
b) M plot



a) PD plot $\mathbb{E}_{\mathbf{x}_2} \left(\hat{f}(x_1, \mathbf{x}_2) \right)$ is estimated by $\hat{f}_{1,PD}(x_1) = \frac{1}{n} \sum_{i=1}^n \hat{f}(x_1, \mathbf{x}_2^{(i)})$

b) M plot $\mathbb{E}_{\mathbf{x}_2 | \mathbf{x}_1} \left(\hat{f}(x_1, \mathbf{x}_2) \middle| \mathbf{x}_1 \right)$ is estimated by $\hat{f}_{1,M}(x_1) = \frac{1}{|N(x_1)|} \sum_{i \in N(x_1)} \hat{f}(x_1, \mathbf{x}_2^{(i)})$,
where index set $N(x_1) = \{i : x_1^{(i)} \in [x_1 - \epsilon, x_1 + \epsilon]\}$ refers to observations with feature value close to x_1 .

M PLOT VS. PD PLOT



- M plots average predictions over conditional distribution (e.g., $\mathbb{P}(\mathbf{x}_2 | x_1)$)
⇒ Averaging predictions close to data distribution avoid extrapolation issues
- **But:** M plots suffer from omitted-variable bias (OVB)
 - They contain effects of other dependent features
 - Useless in assessing a feature's marginal effect if feature dependencies are present

M PLOT VS. PD PLOT - OVB EXAMPLE

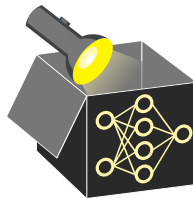
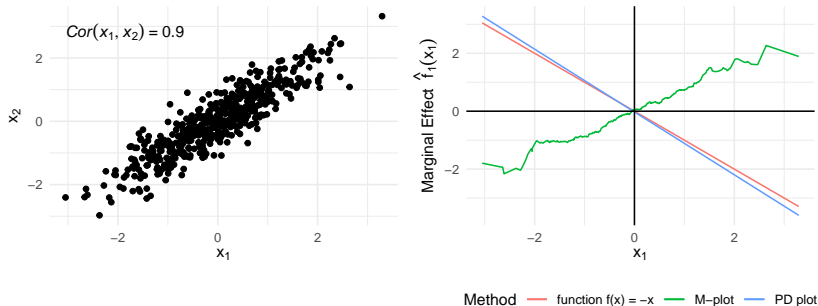


Illustration: Fit LM on 500 i.i.d. observations with features $x_1, x_2 \sim N(0, 1)$, $Cor(x_1, x_2) = 0.9$ and

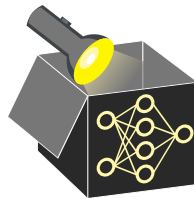
$$y = -x_1 + 2 \cdot x_2 + \epsilon, \quad \epsilon \sim N(0, 1).$$

Results: M plot of x_1 also includes marginal effect of all other dependent features (here: x_2)

IDEA: INTEGRATING PARTIAL DERIVATIVES

Idea: To remove unwanted effects of other features, take partial derivatives (local effects) of prediction function w.r.t. feature of interest and integrate (accumulate) them w.r.t. the same feature

- ⇒ Computing the partial derivative of \hat{f} w.r.t. \mathbf{x}_j removes other main effects
- ⇒ Integrating again w.r.t. \mathbf{x}_j recovers the original main effect of \mathbf{x}_j



IDEA: INTEGRATING PARTIAL DERIVATIVES

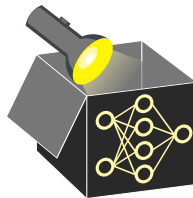
Idea: To remove unwanted effects of other features, take partial derivatives (local effects) of prediction function w.r.t. feature of interest and integrate (accumulate) them w.r.t. the same feature

- ⇒ Computing the partial derivative of \hat{f} w.r.t. \mathbf{x}_j removes other main effects
- ⇒ Integrating again w.r.t. \mathbf{x}_j recovers the original main effect of \mathbf{x}_j

Example:

- Consider an additive prediction function:

$$\hat{f}(x_1, x_2) = 2x_1 + 2x_2 - 4x_1x_2$$



IDEA: INTEGRATING PARTIAL DERIVATIVES

Idea: To remove unwanted effects of other features, take partial derivatives (local effects) of prediction function w.r.t. feature of interest and integrate (accumulate) them w.r.t. the same feature

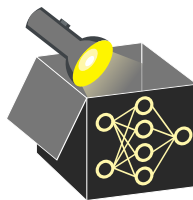
- ⇒ Computing the partial derivative of \hat{f} w.r.t. \mathbf{x}_j removes other main effects
- ⇒ Integrating again w.r.t. \mathbf{x}_j recovers the original main effect of \mathbf{x}_j

Example:

- Consider an additive prediction function:

$$\hat{f}(x_1, x_2) = 2x_1 + 2x_2 - 4x_1x_2$$

- Partial derivative of \hat{f} w.r.t. x_1 : $\frac{\partial \hat{f}(x_1, x_2)}{\partial x_1} = 2 - 4x_2$



IDEA: INTEGRATING PARTIAL DERIVATIVES

Idea: To remove unwanted effects of other features, take partial derivatives (local effects) of prediction function w.r.t. feature of interest and integrate (accumulate) them w.r.t. the same feature

- ⇒ Computing the partial derivative of \hat{f} w.r.t. \mathbf{x}_j removes other main effects
- ⇒ Integrating again w.r.t. \mathbf{x}_j recovers the original main effect of \mathbf{x}_j

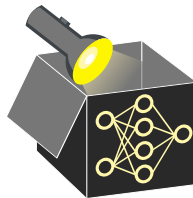
Example:

- Consider an additive prediction function:

$$\hat{f}(x_1, x_2) = 2x_1 + 2x_2 - 4x_1x_2$$

- Partial derivative of \hat{f} w.r.t. x_1 : $\frac{\partial \hat{f}(x_1, x_2)}{\partial x_1} = 2 - 4x_2$
- Integral of partial derivative ($z_0 = \min(x_1)$):

$$\int_{z_0}^x \frac{\partial \hat{f}(x_1, x_2)}{\partial x_1} dx_1 = [2x_1 - 4x_1x_2]_{z_0}^x$$



IDEA: INTEGRATING PARTIAL DERIVATIVES

Idea: To remove unwanted effects of other features, take partial derivatives (local effects) of prediction function w.r.t. feature of interest and integrate (accumulate) them w.r.t. the same feature

- ⇒ Computing the partial derivative of \hat{f} w.r.t. \mathbf{x}_j removes other main effects
- ⇒ Integrating again w.r.t. \mathbf{x}_j recovers the original main effect of \mathbf{x}_j

Example:

- Consider an additive prediction function:

$$\hat{f}(x_1, x_2) = 2x_1 + 2x_2 - 4x_1x_2$$

- Partial derivative of \hat{f} w.r.t. x_1 : $\frac{\partial \hat{f}(x_1, x_2)}{\partial x_1} = 2 - 4x_2$
- Integral of partial derivative ($z_0 = \min(x_1)$):

$$\int_{z_0}^x \frac{\partial \hat{f}(x_1, x_2)}{\partial x_1} dx_1 = [2x_1 - 4x_1x_2]_{z_0}^x$$

- We removed the main effect of x_2 , which was our goal

