Interpretable ML

PDP, ICE, ALE

Partial Dependence Plots

Plotting feature effects in “black box” learning methods (Friedman 2001)

$$\tilde{f}(x) = \frac{1}{n}\sum_{i=1}^{n} f(x, x_{iC})$$

General idea

- Compute $\tilde{f}(x)$ over the range of x while averaging the effects of the remaining predictors x_C
- Generate artificial datasets by fixing x-values for all cases
  - Regression: Averaging over $f(x, x_{iC})$ for each value of x
  - Classification: Averaging over p or logit(p) for each value of x

Partial Dependence Plots

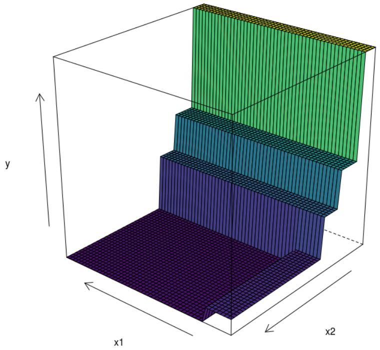
Constructing PDPs

1. Choose a range of values $\{x_{11}, x_{12}, \dots, x_{1k}\}$ of x_1
2. For each $i \in \{1, 2, \dots, k\}$
    1. Generate an artificial dataset by fixing x_1 to x_{1i} for all cases
    2. Compute predictions for all cases using the prediction model (e.g. RF)
    3. Average the predictions over all cases
3. Plot the obtained average predictions against x_{1i} for $i = 1, 2, \dots, k$

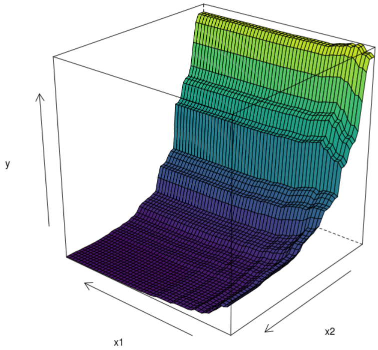
Partial Dependence Plots

Figure: Partial dependence plots

(a) CART



(b) Random Forest



Individual Conditional Expectation

ICE plots (Goldstein et al. 2014)

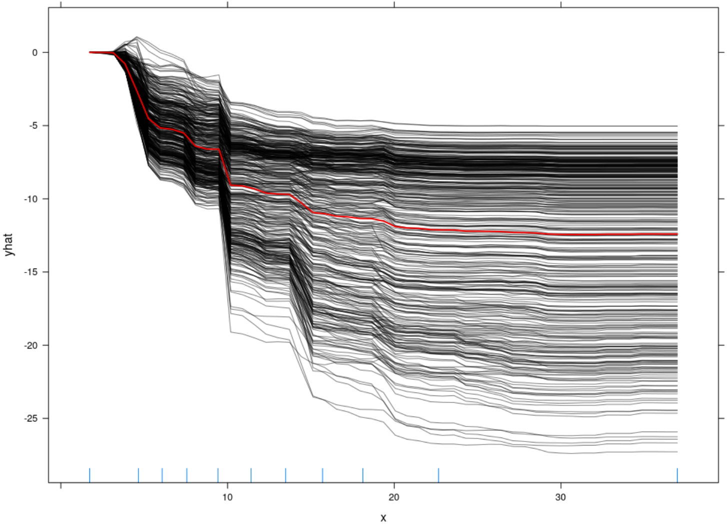
- Individual PDPs for all cases w/o final averaging
- One line represents the predictions for one case over the range of x
- Can uncover heterogeneous effects that are driven by interactions

Centered ICE plots

- Adjusts for different individual baselines
- Shows differences in prediction relative to anchor (e.g., x_{min})

Individual Conditional Expectation

Figure: Centered ICE plot



Accumulated Local Effects

ALE plots (Apley 2016)

- With correlated features, PDPs can (artificially) construct very unlikely combinations
- ALE solution:
  1. Use only cases with (similar) x-values within a given interval
  2. Calculate differences in predictions between upper and lower limit of this interval

$$\hat{\tilde{f}}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i:x_j^{(i)} \in N_j(k)} \left[f(z_{k,j}, x_{\setminus j}^{(i)}) - f(z_{k-1,j}, x_{\setminus j}^{(i)}) \right]$$

$\rightarrow$ Differences in predictions in interval $z_{k,j}$, $z_{k-1,j}$ for cases in neighborhood $N_j(k)$ accumulated up to interval k_j

Accumulated Local Effects

Figure: Comparison of feature effect plots

(a) ALE

(b) PDP