Imperial College London

MSc Machine Learning & Data Science

Applicable Mathematics

Ethics of Data Science – Part II

Measuring feature effects in ML models: PDPs and ICEs.

Dr. Chris Anagnostopoulos, Hon. Senior Lecturer

Table of Contents

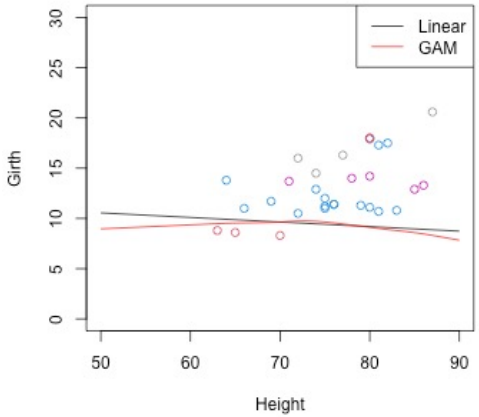
ICE is a form of local explanation

How will the prediction vary when you change one feature?

Girth Height Volume

8.3 70 10.3

50, 51, …, 89, 90



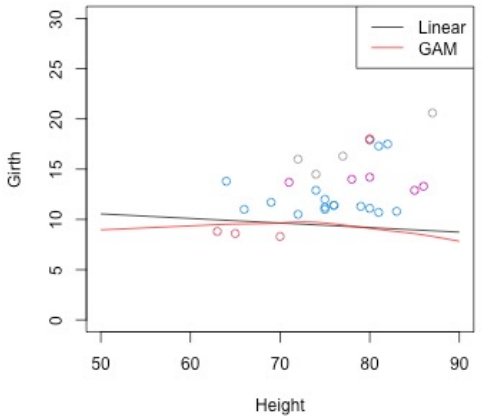
ICE is a form of local explanation

```
newdata = data.frame(Height=50:90, Volume=trees$Volume[1])

predict(lm(Girth ~ Height + Volume, data=trees), newdata=newdata)
predict(gam(Girth ~ s(Height) + Volume, data=trees), newdata=newdata)
```

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \beta_2 v_1$$

$$\hat{g}(x) = \hat{\beta}_0 + \hat{s}_1(x) + \beta_2 v_1$$



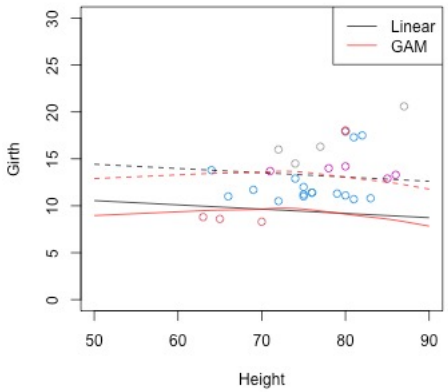
ICE is a form of local explanation

```
newdata = data.frame(Height=50:90, Volume=mean(trees$Volume))

predict(lm(Girth ~ Height + Volume, data=trees), newdata=newdata)
predict(gam(Girth ~ s(Height) + Volume, data=trees), newdata=newdata)
```

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \beta_2 v_1$$

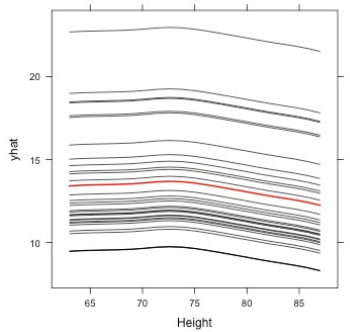
$$\hat{g}(x) = \hat{\beta}_0 + \hat{s}_1(x) + \beta_2 v_1$$



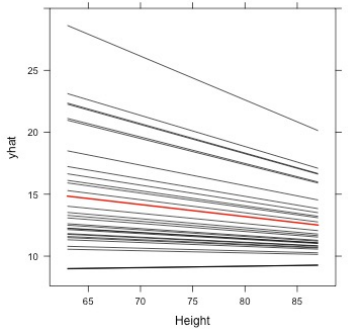
ICE curves can reveal linearity / additivity

```
library(pdp)
m_int = lm(Girth ~ Height + Volume + I(Height*Volume), data=trees)
partial(m_gam, pred.var = 'Height', ice=TRUE, plot=TRUE)
partial(m_int, pred.var = 'Height', ice=TRUE, plot=TRUE)
```

Additive, non-linear

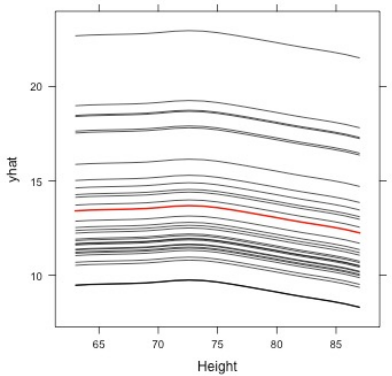


Linear, non-additive



- Linear ICE curves suggest linearity in that parameter
- ”Parallel” (shifted up or down) ICE curves suggest additivity

Partial Dependence Plots are an average over ICEs



Averaging the ICEs over the observed data except for one variable yields a PDP:

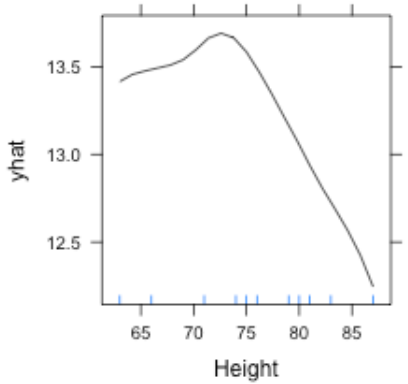
$$\hat{f}(x) = \frac{1}{n} \sum_{i=1:n} f(x, z_i, w_i, \dots)$$

In the linear and additive case, this is the same as an ICE where other variables are assigned their average values:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1:n} (\beta_0 + \beta_1 x + \beta_2 v_i) = \beta_0 + \beta_1 x + \beta_2 \frac{1}{n} \sum_{i=1}^{n} v_i$$

Partial Dependence Plots

- A flat PDP suggests that the feature has no impact on the curve (for example, an estimate of variation along the PDP curve –variance, max minus min- can measure feature importance).
- It can sometimes be the case that a feature is not informative on average but interacts with other features. This would translate to a flat PDP, but ICE plots with variable shapes.

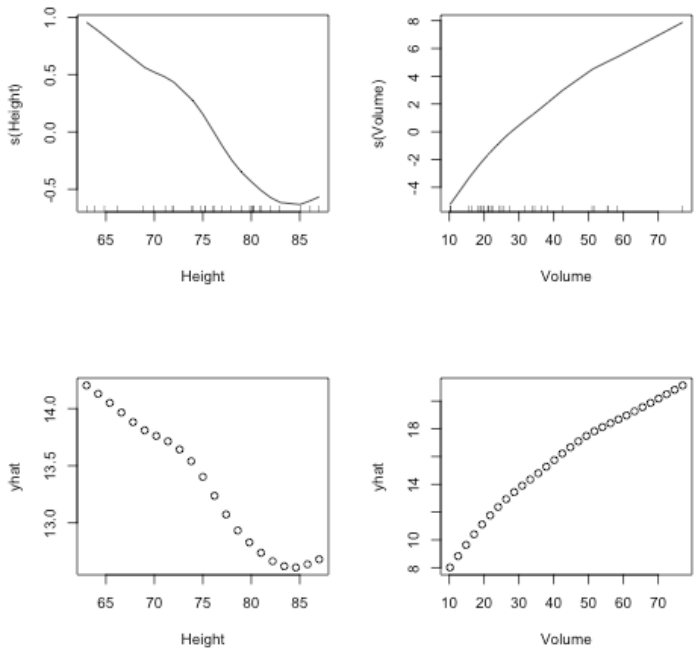


It is good practice to show the density of points on the x-axis to avoid over-interpreting out-of-sample (known as “rug”).

The R package *pdp* uses max/min and decile representation. Other packages in R/Python use a histogram representation.

Partial Dependence Plots

- In GAMs, each smooth term, due to additivity, is identically shaped to the respective ICE, or indeed the PDP, but with different vertical shift.
- It is worth thinking about the meaning of “0” in a GAM smooth plot, and that might vary across different implementations.



Partial Dependence Plots

- PDPs are ”model-based counterfactuals” – though might not represent real-world causality.
- However, some of these counterfactuals are describing combination of features that would not occur in real life. For example, predicting the girth for a tree of average volume and an extreme height.

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1:n} (\beta_0 + \beta_1 x + \beta_2 v_i) = \beta_0 + \beta_1 x + \beta_2 \frac{1}{n} \sum_{i=1}^{n} v_i$$

- This shortcoming is absent when the feature in question is uncorrelated with the rest – unlikely to hold in practice.
- It also holds with ICE, though other features are there fixed to a specific known value, without averaging – easier to explain.

Partial Dependence Plots

- PDPs are ”model-based counterfactuals” – though might not represent real-world causality.
- However, some of these counterfactuals are describing combination of features that would not occur in real life. For example, predicting the girth for a tree of average volume and an extreme height.

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1:n} (\beta_0 + \beta_1 x + \beta_2 v_i) = \beta_0 + \beta_1 x + \beta_2 \frac{1}{n} \sum_{i=1}^{n} v_i$$

- This shortcoming is absent when the feature in question is uncorrelated with the rest – unlikely to hold in practice.
- It also holds with ICE, though other features are there fixed to a specific known value, without averaging – easier to explain.

Summary

- ICEs describe the way in which predictions vary across the range of a feature while holding other things constant.
- The relative shape of ICEs can reveal whether effects are additive vs whether there are interaction effects present in the data
- The average of these curves is known as a PDP and measures the overall average effect of the feature on the response.