Global model agnostic interpretation

Throughout we assume the variable of interest is with realisations , the trained machine learning model is referred to as

# Feature function form

….

## Illustrative models

Linear model with no interaction.

Linear model with interaction.

In both cases there may be correlation between x\_1 and x\_2

## Partial dependency plot

The partial dependency plot illustrates the marginal effect of variable x\_1 on the output of the model f. The **partial dependency function** is calculated by averaging the predicted output of f for X\_1=x\_1 with respect to the distribution of X\_2:

The **partial dependency plot** graphs across various values of X\_1.

Consider the linear model with no interaction effects , estimated via least squares on some dataset large enough we can ignore that our model is an estimate, here the partial dependency function is:

The resulting partial dependency plot (calculated via varying x\_1 across a range of appropriately chosen values, e.g. all unique values in the observed data) will be a graph of a straight line with slope and intercept which only depends on the other features through their expected values.

For the linear model with an interaction effect we get:

If x\_1 and x\_2 are correlated this doesn’t seem entirely correct, as in reality the effect of the interaction depends on the value of both x\_1 and x\_2 which vary together (e.g. the effect is stronger when x\_1 is large since x\_2 is likely large) which gets ignored here.

The PDP is extrapolating the effect of x\_1 observed around E\_X(f(x,..)) to the entire support of X?

## M-plots

Since the PDP extrapolates lets use the conditional distribution.

For the linear model with no interaction

If x1 and x2 are correlated, then this is not going to be much different to the original function – i.e. we haven’t isolated the effect of x\_2 at all!

For the linear model with interaction

How can this go wrong?

Consider …

## Accumulated local effects

Motivation: notice that for the linear model with no interaction

If we asked, “what would be the contribution of X\_1 if we change from x\_ref to x\_1?”

Sum up all the changes…

Now notice that:

By reformulating the conditional expectation seen in the M-plot in terms of rate of chance with respect to x\_1 only we remove the effect of x\_2 that “contaminates” the M-plot.

thus difference between any two points on the graph correspond to integral of change in function as x\_1 changes – if we change from x\_ref to x1 then the function changes by .

For the ?? model