Week 5 - Profiles and Local Effects

Sunday, February 16, 2025 9:34 PM

Motivation

1. Partial dependence profiles (PDP), individual conditional expectation (ICE) plots, and accumulated local effects (ALE) are types of global explainability methods. They can be applied post-hoc to any model and try to reveal how the model relates a predictor variable/set of predictor variables with the response.

2. Partial dependence profiles were first introduced as a minor point in an influential paper about gradient boosting. ICE and ALE were introduced to address some limitations of PDP, mainly related to the fact that predictions surfaces can become complicated in high-dimensions and when many interactions are present.

Partial Dependence Profiles

1. Let f be the learned black box function. Let x be its D-dimensional input. The idea is to ask how f changes as we vary the d^th coordinate over its range. We need to do something with the remaining coordinates; the idea of PDP is to average across them.

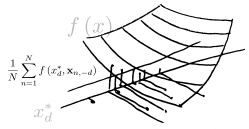
$$PDP_d(z) = \mathbb{E}_{P(x_{-d})}\left[f(x_d = z, x_{-d})\right]$$
$$\approx \frac{1}{N}\sum_{n=1}^{N} f(x_d = z, x_{n,-d})$$

The highlighted part refers to manually intervening on the d^th coordinate and setting it to z. If we evaluate this average across many different values of z, we get a PDP curve. Note that for each value z, we compute an average across the entire dataset -- this can get slow.

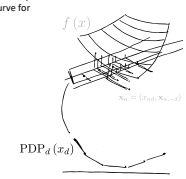
2. This definition generalizes to larger-dimensional subsets of variables, but we'll stick to one-dimensional curves for simplicity.

3. Here is some geometric intuition. We take the original samples, manually set all their d^th coordinates to z, evaluate the function f, and then average. We repeat this for many z's to trace out the partial dependence curve.

Evaluating PDP at a single value z = x*.



Repeating this at many z gives the PDP curve for coordinate d.



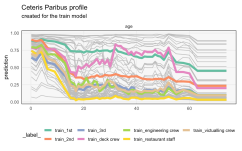
Individual Conditional Expectations

1. ICE plots consider each term in the PDP formula as its own function of z.

$$h_n(z) := f(x_d = z, x_{n,-d})$$

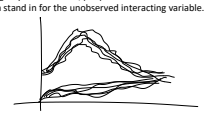
This has the nice "Ceteris Paribus" interpretation. For each sample, it traces out the curve where the d^th coordinate is set to different values z but all other coordinates are held fixed.

2. The key insight is that these curves can be meaningful on their own. The averaging step in PDP can mask some meaningful variation in model predictions. This is especially useful for discovering interactions. For example, here are ICE curves for a model trained to predict survival with the Titanic dataset. The age effect is very different for crew vs. passengers.



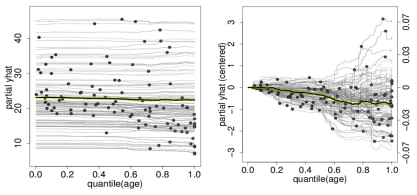
3. Two common tricks for ICE plots:

- We might suspect an interaction with an unmeasured variable. This can sometimes be seen as groups of curves with different shapes, but where we don't have a variable in the data that splits those groups. In this case, people often cluster the curves and use cluster membership as a stand in for the unobserved interacting variable.



- Often the curves are offset because of differences in the coordinates other than the d^th. To see whether they nonetheless have similar shapes, it's sometimes helpful to "center" the curves, so that they all start out equal to one another.

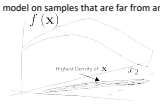
$$\tilde{h}_n(z) = f(x_d = z, x_{n,-d}) - f(x_d = 0, x_{n,-d})$$



Accumulated Local Effects

1. ALE plots are designed to resolve two fundamental issues with partial dependence profiles. The first issue is relatively obvious, the second is more subtle.

2. The obvious issue is that by manually setting the d^th coordinate of each sample to z, we might be querying the model on samples that are far from any samples that are observed.



This is not a purely theoretical concern. Imagine we fit a model of employee characteristics -> income, and now we want to ask about the age effect. (age, seniority) are highly correlated predictors. PDP will intervene on all samples' age regardless of seniority; e.g., we might query the model about 18 year old CEOs, even if we didn't see any in the training data.

3. A natural idea is to restrict the average to samples whose actual x[d] is close to the queried z (in the example above, only samples with age close to the currently queried age).

$$\frac{1}{|B_d(z)|}\sum_{n \in B_d(z)} f(x_d = z, x_{n,-d})$$

4. But this isn't enough! The issue is that it will pick on the effects of correlated features. For example, suppose:

- The input x is two dimensional.
- f is a linear model where the first coordinate has no effect.
- The two coordinates are correlated.

$\beta = (0, 1)$

$f(x) = \beta^T x = x_2$

$\text{Cor}(x_1, x_2)$ large

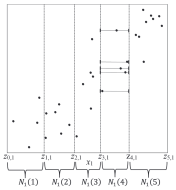
Then, the formula above will look like,

$$\frac{1}{|B_i(z)|}\sum_{n \in B_i(z)} f(z, x_{n2}) = \frac{1}{|B_i(z)|}\sum_{n \in B_i(z)} x_{n2}$$

Since x_1 and x_2 are correlated, this is large when z is large and small otherwise. So, we will think that there is an effect due to x_1, even though the associated model coefficient is 0.

5. Accumulated Local Effects plots deal with this by considering _differences_ in f at nearby values of z. Here, K indexes a fine grid along the dimension of interest, and z[k], z[k - 1] are neighboring grid points. The set B now includes all samples whose z value lie between those endpoints.

$$\sum_{k=1}^{K} \frac{1}{B_d(k)} \sum_{n \in B_d(k)} \left[f(z_k, x_{n,-d}) - f(z_{k-1}, x_{n,-d})\right]$$



6. Notice that this fixes the problem we identified above, because the effect of the second coordinate cancels:

$$\sum_{n \in B_d(k)} f(z_k, x_{n2}) - f(z_{k-1}, x_{n2})$$
$$= \sum_{n \in B_d(k)} x_{n2} - x_{n2} = 0$$

Effects and Partial Residual Plots

1. Fox and Weisberg (2018) focus on visualizing effects just in the case of linear regression models. Even here, it pays to be careful when thinking about the role of interactions.

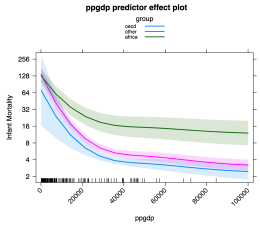
2. Suppose we are interested in the d^th predictor in a linear regression model. They partition the predictors into (z, v, u):

- z - the predictor we are interested in.
- v - the predictors that interact with z.
- u - other unrelated predictors.

3. They recommend identifying a grid of values across v and then tracing the function of z with v fixed to each of those grid points.

$$\left\{h_d(z, v)\right\}_{v \in G_v}$$
$$= \left\{\left(\begin{matrix} z \\ v \\ 0 \end{matrix}\right)^T \hat{\beta}\right\}_{v \in G_v}$$

In this example z = per person GDP and v = country group



4. They also suggest overlying the residuals on these effects plots.

$$\left(\begin{matrix} z = x_n \\ v \end{matrix}\right)^T \hat{\beta} + e(x_n)$$

(I don't like their notation for this... though I don't know if this is any better).

Regardless, the idea is simple: This plot can help identify systematic structure in the residuals. E.g., if the residuals are systematically different in some regions of the z-space, then we might have a mis-specified model. For example, in this toy example, the residuals suggest that the model is missing an interaction term between x2 and x3.