ALE

ALE solves the problem of mixing effects from different features. As with the function $M(x_1)$, ALE uses the conditional distribution to average over other features, but instead of averaging the predictions directly, it averages ***differences in predictions*** to block the effect of correlated features. The ALE function is defined as follows:

$$
\begin{aligned}
\mathrm{ALE}(x_1) &= \int_{\min(x_1)}^{x_1} \mathbb{E}\left[\frac{\partial f(X_1, X_2)}{\partial X_1} \middle| X_1 = z_1\right] dz_1 - c_1 \\
&= \underbrace{\int_{\min(x_1)}^{x_1} \int p(x_2|z_1) \frac{\partial f(z_1, x_2)}{\partial z_1} dx_2 dz_1}_{\text{uncentered ALE}} - c_1,
\end{aligned}
$$

where the constant c_1 is chosen such that the resulting ALE values are independent of the point $\min(x_1)$ and have zero mean over the distribution $p(x_1)$.

常量c的选择是为了使ALE的值与点min(x₁)无关，并且在p(x₁)的分布上有0平均值、

The term $\frac{\partial f(x_1, x_2)}{\partial x_1}$ is called the ***local effect*** of x_1 on f. Averaging the local effect over the conditional distribution $p(x_2|x_1)$ allows us to isolate the effect of x_1 from the effects of other correlated features avoiding the issue of M plots which directly average the predictor f. Finally, note that the local effects are integrated over the range of x_1, this corresponds to the ***accumulated*** in ALE. This is done as a means of visualizing the ***global*** effect of the feature by "piecing together" the calculated local effects.

In practice, we calculate the local effects by finite differences so the predictor f need not be differentiable. Thus, to estimate the ALE from data, we compute the following:

$$
\widehat{\mathrm{ALE}}(x_1) = \underbrace{\sum_{k=1}^{k(x_1)} \frac{1}{n(k)} \sum_{i: x_1^{(i)} \in N(k)} \left[f(z_k, x_{\setminus 1}^{(i)}) - f(z_{k-1}, x_{\setminus 1}^{(i)})\right]}_{\text{uncentered ALE}} - c_1.
$$

Here $z_0, z_1, \dots$ is a sufficiently fine grid of the feature x_1 (typically quantiles so that each resulting interval contains a similar number of points), $N(k)$ denotes the interval $[z_{k-1}, z_k)$, $n(k)$ denotes the number of points falling into interval $N(k)$ and $k(x_1)$ denotes the index of the interval into which x_1 falls into, i.e. $x_1 \in [z_{k(x_1)-1}, z_{k(x_1)})$.

n(k)代表在N(k)区间里的样本数、

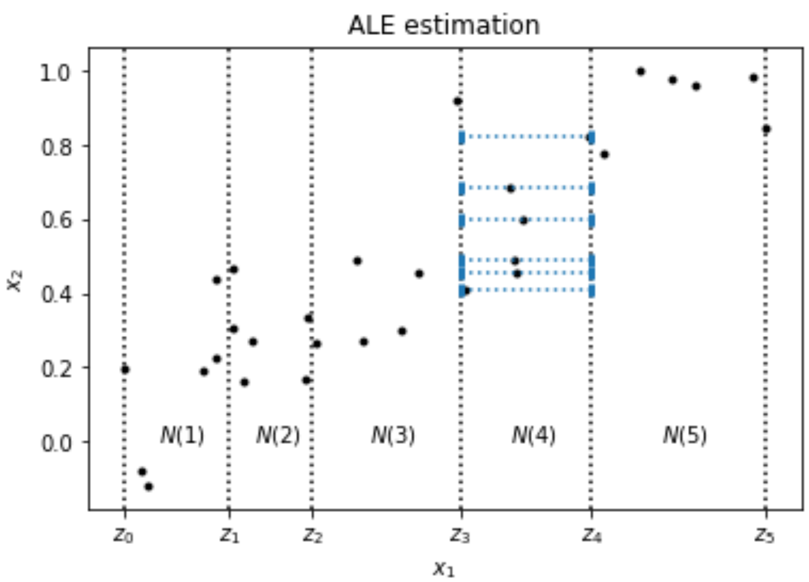
k(x₁)代表x₁ feature落入的区间的索引.

Finally, the notation $f(z_k, x_{\setminus 1}^{(i)})$ means that for instance i we replace x_1 with the value of the right interval end-point z_k (likewise for the left interval end-point using z_{k-1}), leaving the rest of the features unchanged, and evaluate the difference of predictions at these points.

区间的每个样本都用端点值替换它们在feature x₁的值，计算差异.

The following plot illustrates the ALE estimation process. We have subdivided the feature range of x_1 into 5 bins with roughly the same number of points indexed by $N(k)$. Focusing on bin $N(4)$, for each point falling into this bin, we replace their x_1 feature value by the left and right end-points of the interval, z_3 and z_4. Then we evaluate the difference of the predictions of these points and calculate the average by dividing by the number of points in this interval $n(4)$. We do this for every interval and sum up (accumulate) the results. Finally, to calculate the constant c_1, we subtract the expectation over $p(x_1)$ of the calculated uncentered ALE so that the resulting ALE values have mean zero over the distribution $p(x_1)$.

c₁的计算：实际上c₁即是所有ALE值的期望，当所有ALE计算好后，减去它们的平均值就得到一个0平均的中心化分布

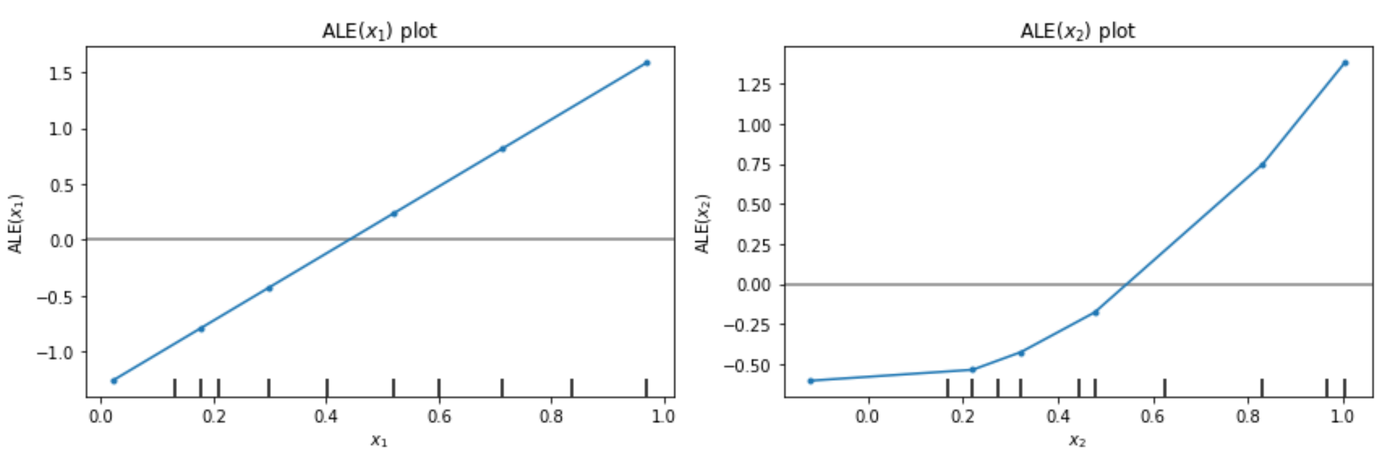


We show the results of ALE calculation for a model $f(x_1, x_2) = 3x_1 + 2x_2^2$. The resulting plots correctly recover the linear effect of x_1 and the quadratic effect of x_2 on f. Note that the ALE is estimated for each interval edge and linearly interpolated in between, for real applications it is important to have a sufficiently fine grid but also one that has enough points into each interval for accurate estimates. The x-axis also shows feature deciles of the feature to help judge in which parts of the feature

deciles 十分位数

space the ALE plot is interpolating more and the estimate might be less trustworthy.

The value of $\mathrm{ALE}(x_i)$ is the main effect of feature x_i as compared to the average prediction for the data. For example, the value of $\mathrm{ALE}(x_1) = 0.75$ at $x_1 = 0.7$, if we sample data from the joint distribution $p(x_1, x_2)$ (i.e. realistic data points) and $x_1 = 0.7$, then we would expect the first order effect of feature x_1 to be 0.75 higher than the ***average*** first order effect of this feature. Seeing that the $\mathrm{ALE}(x_1)$ plot crosses zero at $x_1 \approx 0.45$, realistic data points with $x_1 \approx 0.45$ will have effect on f similar to the average first order effect of x_1. For realistic data points with smaller x_1, the effect will become negative with respect to the average effect.



→ 中心化后的 ALE

→ x 值其实是 percentile

Because the model $f(x_1, x_2) = 3x_1 + 2x_2^2$ is explicit and differentiable, we can calculate the ALE functions analytically which gives us even more insight. The partial derivatives are given by $(3, 4x_2)$. Assuming that the conditional distributions $p(x_2|x_1)$ and $p(x_1|x_2)$ are uniform, the expectations over the conditional distributions are equal to the partial derivatives. Next, we integrate over the range of the features to obtain the ***uncentered*** ALE functions:

$$\mathrm{ALE}_u(x_1) = \int_{\min(x_1)}^{x_1} 3dz_1 = 3x_1 - 3\min(x_1)$$
$$\mathrm{ALE}_u(x_2) = \int_{\min(x_2)}^{x_2} 4z_2dz_2 = 2x_2^2 - 2\min(x_2)^2.$$

Finally, to obtaine the ALE functions, we center by setting $c_i = \mathbb{E}(\mathrm{ALE}_u(x_i))$ where the expectation is over the marginal distribution $p(x_i)$:

$$\mathrm{ALE}(x_1) = 3x_1 - 3\min(x_1) - \mathbb{E}(3x_1 - 3\min(x_1)) = 3x_1 - 3\mathbb{E}(x_1)$$
$$\mathrm{ALE}(x_2) = 2x_2^2 - 2\min(x_2)^2 - \mathbb{E}(2x_2^2 - 2\min(x_2)^2) = 2x_2^2 - 2\mathbb{E}(x_2^2).$$

This calculation verifies that the ALE curves are the desired feature effects (linear for x_1 and quadratic for x_2) relative to the mean feature effects across the dataset. In fact if f is additive in the individual features like our toy model, then the ALE main effects recover the correct additive components (Apley and Zhu (2016): https://arxiv.org/abs/1612.08468).

Furthermore, for additive models we have the decomposition $f(x) = \mathbb{E}(f(x)) + \sum_{i=1}^{d} \mathrm{ALE}(x_i)$, here the first term which is the average prediction across the dataset X can be thought of as zeroth order effects.