Full Title of Article

Author Name1 ABC@SAMPLE.COM
Address 1

Author Name2 XYZ@SAMPLE.COM
Address 2

Editors: Emtiyaz Khan and Mehmet Gonen

Abstract

This is the abstract for this article.
Keywords: List of keywords separated by semicolon.

1. Introduction

Main contents here.

The decision about the bin size has major influence in the feature effect plot. Therefore, it is very important (a) to inform the user to what extend they should trust the ALE plot and, consequently, decide the bin size with the optimal effect. In the end, we will see that it i

2. Related Work

3. Probabilistic formulation of ALE plots

3.1. ALE background

This section introduces the reader to the ALE formulation of feature effect. Given that $f : \mathbb{R}^S \to \mathbb{R}$ is known, we can measure the effect of the s-th feature at a specific point $\mathbf{x} = (\mathbf{x}_c, x_s)$ of the input space $\mathcal{X}$ as $f_s(\mathbf{x}) = {}^{\partial f(\mathbf{x})}/_{\partial x_s}$. ALE models the *local* effect at x_s as the expected effect over the distribution of the unkown (latent) features $\mathbf{X}_c$, i.e. $\mathbb{E}_{p(\mathbf{x}_c;x_s)}[f_s(x_s, \mathbf{x}_c)]$. Afterwards, the *global* effect at x_s is the accumulation of the *local* effects:

$$f_{\texttt{ALE}}(x_s) = \int_{x_{s,\min}}^{x_s} \mathbb{E}_{p(\mathbf{x}_c;x_s=z)}[f_s(x_s, \mathbf{x}_c)]\partial z \tag{1}$$

In real cases, it is infeasible to compute eq. (1) analytically. Therefore, we reside on estimating the effect from the training set. Let's denote the available dataset as $\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N$, where $\mathbf{x}^{(i)}$ is the i-th feature vector of the training data and $y^{(i)}$ is the i-th label. Apley and Zhu (2020) proposed spliting the axis into K equal-sized bins, find the set of points that lie in each bin, i.e. $\mathcal{S}_k = \{\mathbf{x}^{(i)} : x_s^{(i)} \in [z_{k-1}, z_k)\}$ and, finally, find the local effect at each bin as the mean value of the population $\hat{\mu}_{s,k}$. The global effect is then estimated through the following formula:

$$\hat{f}_{\texttt{ALE}}(x_s) = \Delta x \sum_{k=1}^{k_{x_s}} \hat{\mu}_{s,k} = \sum_{k=1}^{k_{x_s}} \frac{1}{|\mathcal{S}_k|} \sum_{i:x_s^{(i)} \in \mathcal{S}_k} [f(z_k, \mathbf{x}_c^{(i)}) - f(z_{k-1}, \mathbf{x}_c^{(i)})] \tag{2}$$

where Δx is the bin size and k_{x_s} the index of the bin that includes x_s.

3.2. Quantification of the uncertainty of the estimation

We propose an alternative estimation of the feature effect inside each bin. The alternative estimation is better because it (a) separates the decision about the bin limits from the local effect of each point, (b) secures from out-of-distribution sampling and (c) provides better variance estimation inside the bin. The alternative estimation is:

$$\tilde{f}_{\texttt{ALE}}(x_s) = \Delta x \sum_{k=1}^{k_{x_s}} \tilde{\mu}_{s,k} = \Delta x \sum_{k=1}^{k_{x_s}} \frac{1}{|\mathcal{S}_k|} \sum_{i:x_s^{(i)} \in \mathcal{S}_k} f_s(\mathbf{x}^{(i)}) \tag{3}$$

Our goal is to inform the user to what extent they should trust the ALE plot. For this purpose, we introduce two metrics; the variance and the standard error of the estimation. The variance of the estimation is:

$$\tilde{\sigma}_s^2(x_s) = (\Delta x)^2 \sum_{k=1}^{k_{x_s}} \tilde{\sigma}_{s,k}^2 = (\Delta x)^2 \sum_{k=1}^{k_{x_s}} \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^{(i)} \in \mathcal{S}_k} (f_s(\mathbf{x}^{(i)}) - \tilde{\mu}_{s,k})^2 \tag{4}$$

The standard error of the estimation is:

$$\text{SE}(x_s) = \Delta x \sum_{k=1}^{k_{x_s}} (\frac{\hat{\sigma}_{s,k}^2}{|\mathcal{S}_k|})^{\frac{1}{2}} \tag{5}$$

The two metrics, (4) and (5), should be trusted only in case they respect some contstraints. Under these constraints, the user can be confident that the correct feature effect is inside two or three standard errors errors around the mean estimation.

Our computations are based on the hypothesis that inside all bins **the gradient wrt. to the feature of interest doesn't depend on the value feature of interest**. Unfortunately, we cannot when this is the case. We just know that as the bins grow larger, it is more possible to violated this hypothesis. In this case both the ALE effect and the standard error are wrong.

The standard error should be trusted when it is estimated by a large of data points. For example, in plots (c) and (d), there are bins with less than 10 points. Therefore, in these cases, we cannot trust the plot or the standard error.

3.3. Unsupervised metric for assessing the quality of ALE plots

Based on the observations of the previous chapter, we want to create a single metric for choosing the best feature effect plot. Our goal is to minimize the variance, given that we have points inside eac bin

We propose the minimization of the accumulated standard deviation (or accumulated variance). For ALE with K bins, let's notate as:

- dx^K the length of each bin
- p_i^K the number of the training points inside the $i-th$ bin
- σ_i^K the std of the local effects of the training points inside the $i-th$ bin

We will minimize:

$$K_{min} = \text{argmin}_K \quad [dx^K \sum_i^K \sigma_i^K * (1 - d_i^K)] \tag{6}$$

$$\text{s.t. } p_i^K \geq \texttt{min\_points\_per\_bin} \ \forall i \tag{7}$$

where $d_i^K = 0.2 * \frac{p_i^K}{N} \in [0, 0.1]$ works as a 'discount', favoring the creation of bigger bins in cases of similar standard deviation.

3.4. ALE plots with variable-size bins

4. Experiments

4.1. Synthetic Data sets

1. One example to show that (a) variance and (b) standard error are good estimates to what extend to trust the ALE plot.

2. One example to show the importance of variable size bins

4.2. Real Data sets

5. Conclusion

A figure in Fig. 1. Please use high quality graphics for your camera-ready submission – if you can use a vector graphics format such as `.eps` or `.pdf`.

Figure 1: A spiral.

An example of citation Zhou and Washio (2009).

References

Daniel W. Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 82(4):1059–1086, 2020. ISSN 14679868. doi: 10.1111/rssb.12377.

Zhi-Hua Zhou and Takashi Washio, editors. *Advances in Machine Learning, First Asian Conference on Machine Learning, ACML 2009, Nanjing, China, November 2-4, 2009. Proceedings*, volume 5828 of *Lecture Notes in Computer Science*, 2009. Springer. ISBN 978-3-642-05223-1. doi: 10.1007/978-3-642-05224-8. URL http://dx.doi.org/10.1007/978-3-642-05224-8.

Appendix A. Symbols

Spaces and points:

- $\mathcal{X} \subseteq \mathbb{R}^S$: the input space
- $\mathbf{x} = (x_s, \mathbf{x}_c) \in \mathcal{X}$
- $x_s \in \mathbb{R}$
- $\mathbf{x}_c \in \mathbb{R}^{S-1}$
-

Functions:

- $f : \mathbb{R}^S \to \mathbb{R}$: black-box function
- $f_s(\mathbf{x}) = {}^{\partial f(\mathbf{x})}/_{\partial x_s} : \mathbb{R}^S \to \mathbb{R}$: effect at point $\mathbf{x}$
-

ALE:

- ${}^{f(z_{k_x}) - f(z_{k_x-1})}/_{\Delta x}$: effect of x_s at $\mathbf{x}$
- $\mathbb{E}_{p(\mathbf{x}_c; x_s)}[f_s(x_s, \mathbf{x}_c)]$: local effect of x_s at x_s
- $f_{\texttt{ALE}}(x_s) : \mathbb{R} \to \mathbb{R}$: global effect of x_s at x_s
- $\hat{f}_{\texttt{ALE}}(x_s) : \mathbb{R} \to \mathbb{R}$: *estimator* of global effect of x_s at x_s

Appendix B. Derivations

Derivations

$$
\begin{aligned}
Y_s^{\text{local}} \sim p(y_s^{\text{local}}; x_s) &= \int p(y_s^{\text{local}} | \mathbf{x}_c; x_s) p(\mathbf{x}_c; x_s) \partial \mathbf{x}_c \\
&= \int_{\mathbf{x}_c} \delta(y_s^{\text{local}} - f_s(\mathbf{x})) p(\mathbf{x}_c; x_s) \partial \mathbf{x}_c = \int_{\mathbf{x}_c} \mathbb{1}(y_s^{\text{local}} = f_s(\mathbf{x})) p(\mathbf{x}_c; x_s) \partial \mathbf{x}_c \quad (8)
\end{aligned}
$$

Some statistics for the local feature effect:

$$
\mathbb{E}[Y_s^{\text{local}}; x_s] = \int_{\mathbf{x}_c} p(\mathbf{x}_c | x_s) f_s(\mathbf{x}) \partial \mathbf{x}_c \quad (9)
$$

$$
\text{Var}[Y_s^{\text{local}}; x_s] = \int_{\mathbf{x}_c} p(\mathbf{x}_c | x_s) (f_s(\mathbf{x}) - \mathbb{E}[Y_s^{\text{local}}; x_s])^2 \partial \mathbf{x}_c \quad (10)
$$

Some statistics for the global feature effect:

$$
\mathbb{E}[Y_s; x_s] = \int_{x_{s,min}}^{x_s} \mathbb{E}(y_s^{\text{local}}; z) \partial z \quad (11)
$$

$$
\text{Var}[Y_s; x_s] = \int_{x_{s,min}}^{x_s} \text{Var}[Y_s^{\text{local}}; x_s] (\partial z)^2 \quad (12)
$$

Appendix C. Second Appendix

This is the second appendix.