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.

2. Related Work

3. Probabilistic ALE plots

3.1. The ALE method

In this section, we introduce the reader to the feature effect method ALE. Given that $f : \mathbb{R}^S \to \mathbb{R}$ is known, we can measure the *local* 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* feature effect at x_s as an expectation over the distribution of the unkown (latent) features $\mathbb{E}_{p(\mathbf{x}_c;x_s)}[f_s(x_s, \mathbf{x}_c)]$. Afterwards, ALE measures the *global* effect at x_s as an accumulation of the expected *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 et. al 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, i.e. $\hat{\mu}_{s,k} = \frac{1}{|\mathcal{S}_k|}\sum_{i:\mathbf{x}^{(i)}\in\mathcal{S}_k}[f_s(\mathbf{x}^{(i)})]$. The global effect at x_s is then estimated through:

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

where Δx is the bin size and k_{x_s} the bin-index of the value x_s.

3.2. ALE plots with uncertainty quantification

ALE directly defines the local effect as an expectation over the unknown features and the global effect as integral over the expectations. Therefore, both local and global effects are modeled as simple variables. In constrast, for integrating ALE plots in a probabilistic framework, we model the local effect as a random variable, dependent on the random variable $\mathbf{X}_c$ representing the latent features:

$$Y_s^{\text{local}} \sim p(y_s^{\text{local}}; x_s) = \int_{\mathbf{x}_c} \delta(y_s^{\text{local}} - f_s(\mathbf{x}))p(\mathbf{x}_c; x_s)\partial \mathbf{x}_c \tag{3}$$

Following the core idea of ALE, we define the global feature effect as the random variable Y_s which is the integration of the local effects:

$$Y_s \sim p(y_s; x_s) = \int_{x_{s,min}}^{x_s} p(y_s^{\text{local}}; x_s = z)\partial z \tag{4}$$

Through eqs. (3), (4) we formulate the ALE in a fully probabilistic manner. We notice, that the definition of ALE as given by Apley and Zhu (2020) is simply the expected value of Y_s, i.e. $f_{\texttt{ALE}}(x_s) = \mathbb{E}[Y_s; x_s]$. For simpler notation we denote $\mathbb{E}[Y_s; x_s]$ as $\mu_s(x_s)$. For measuring the uncertainty, it is also important to capture the variance of the global effect. Notating the variance of the feature effect $\text{Var}[Y_s; x_s]$ as σ_s^2, we can compute it through:

$$\sigma_s^2(x_s) = \int_{x_{s,min}}^{x_s} \int_{\mathbf{x}_c} p(\mathbf{x}_c|x_s)(f_s(\mathbf{x}) - \mu_s(x_s))^2 \partial \mathbf{x}_c (\partial z)^2 \tag{5}$$

As before, it is infeasible to compute the the variance $\sigma_s^2(x_s)$ in closed-form. Therefore, we reside on estimating them from the available examples of the training set. The estimation of the variance can also done through the available points:

$$\hat{\sigma}_s^2(x_s) = (\Delta x)^2 \sum_{k=1}^{k_{x_s}} \hat{\sigma}_{s,k}^2 = (\Delta x)^2 \sum_{k=1}^{k_{x_s}} \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^{(i)} \in \mathcal{S}_k} (f_s(\mathbf{x}^{(i)}) - \hat{\mu}_{s,k})^2 \tag{6}$$

3.3. Unsupervised metric for assessing the quality of ALE plots

3.4. ALE plots with variable-size bins

3.5. Variable-size bins as an optimization problem

4. Experiments

4.1. Synthetic Data sets

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`.

An example of citation Zhou and Washio (2009).

Figure 1: A spiral.

4.2. Real Data sets

5. Conclusion

5.1. Subsection Title

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. Derivations

Derivations

$$Y_s^{\text{local}} \sim p(y_s^{\text{local}}; x_s) = \int p(y_s^{\text{local}}|\mathbf{x}_c; x_s)p(\mathbf{x}_c; x_s)\partial\mathbf{x}_c$$
$$= \int_{\mathbf{x}_c} \delta(y_s^{\text{local}} - f_s(\mathbf{x}))p(\mathbf{x}_c; x_s)\partial\mathbf{x}_c = \int_{\mathbf{x}_c} \mathbb{1}(y_s^{\text{local}} = f_s(\mathbf{x}))p(\mathbf{x}_c; x_s)\partial\mathbf{x}_c \quad (7)$$

Some statistics for the local feature effect:

$$\mathbb{E}[Y_s^{\text{local}}; x_s] = \int_{\mathbf{x}_c} p(\mathbf{x}_c|x_s)f_s(\mathbf{x})\partial\mathbf{x}_c \quad (8)$$

$$\text{Var}[Y_s^{\text{local}}; x_s] = \int_{\mathbf{x}_c} p(\mathbf{x}_c|x_s)(f_s(\mathbf{x}) - \mathbb{E}[Y_s^{\text{local}}; x_s])^2\partial\mathbf{x}_c \quad (9)$$

Some statistics for the global feature effect:

$$\mathbb{E}[Y_s; x_s] = \int_{x_{s,min}}^{x_s} \mathbb{E}(y_s^{\text{local}}; z)\partial z \quad (10)$$

$$\text{Var}[Y_s; x_s] = \int_{x_{s,min}}^{x_s} \text{Var}[Y_s^{\text{local}}; x_s](\partial z)^2 \quad (11)$$

Appendix B. Second Appendix

This is the second appendix.