Hybrid Methods: Black and White Box Models in Safety-Critical Domain

I. Desired Properties

1) High trackability: If a false prediction occurs, the reason for this failure can be tracked and fixed.
2) High interpretability: The human experts could understand the model.
3) Awareness of its limitations. Black models usually performs well "locally" where there is a lot of training data. Thus, if the input is within the space that the model is not welled trained, the model must be aware of it.

II. Conformal Prediction

In this section, we present a brief introduction about the theory of conformal prediction [1]. The conformal prediction framework was firstly proposed by Vovk et al. [1], [2], and since then, a lot of extensions (e.g., [3], [4]) have been proposed. Consider regression data $Z^1, Z^2, \ldots, Z^n$, where each $Z^i = (X^i, Y^i)$is a random variable in $\mathbb{R}^d \times \mathbb{R}$, comprised of a response variable Y^i and a d-dimensional vector of features $(X_1^i, X_2^i, \ldots, X_d^i)$. Suppose the regression function is

$$\mu(x) = \mathbb{E}(Y|X = x), \quad x \in \mathbb{R}^d.$$

The conformal prediction uses past experience (e.g., $Z^1, Z^2, \ldots, Z^n$) to form prediction interval $\Gamma^\epsilon(X^{n+1})$ with precise levels of confidence $1 - \epsilon$ in new predictions.

The conformal prediction intervals are guaranteed to deliver proper *finite-sample coverage* on the assumption that $Z^1, Z^2, \ldots, Z^n$ are independent and identically distributed, with no knowledge of the data distribution and the regression function $\mu(x)$, i.e.,

$$\mathbb{P}(Y \in \Gamma^\epsilon(X^{n+1})) \geq 1 - \epsilon \tag{1}$$

Let $\mathcal{I}$ denotes a sub set of index of regression samples and $R_{y,i} = |Y^i - \mu(X^i)|$ for $i \in \mathcal{I}$ denotes the fitted residual of the regression data. We define *nonconformity measure* as

$$A(\mathcal{I}, Z^i) = R_{y,i}$$

to represent how sample Z^i is different from the examples in $\mathcal{I}$. Given a nominal miscoverage level ϵ, then

$$\Gamma^\epsilon(X) = [\hat{\mu}(X) - d, \hat{\mu}(x) + d]$$

$$\text{, where } d = \text{the kth smallest value in } \{R_{y,i} : i \in \mathcal{I}\} \text{ and } k = \lceil (|\mathcal{I}| + 1) \times (1 - \epsilon) \rceil$$

The details of the conformal prediction algorithm is presented in Algorithm

Algorithm 1: Conformal Prediction

Input : Data $Z^1, Z^2, \ldots, Z^n$, miscoverage level ϵ, regression fit algorithm $\mathcal{A}$
Output: Prediction band

Randomly split $\{1, 2, \ldots, n\}$ into two subsets $\mathcal{I}_1$ and $\mathcal{I}_2$
$\hat{\mu} = \mathcal{A}(\{X^i, Y^i : i \in \mathcal{I}_1\})$
$R_{y,i} = |Y^i - \hat{\mu}(X^i)|$ $i \in \mathcal{I}_2$
$d =$ the kth smallest value in $\{R_{y,i} : i \in \mathcal{I}\}$ and $k = \lceil (|\mathcal{I}| + 1) \times (1 - \epsilon) \rceil$
return $[\hat{\mu}(X) - d, \hat{\mu}(x) + d]$

References

[1] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

[2] V. Vovk, I. Nouretdinov, and A. Gammerman, "On-line predictive linear regression," *Ann. Statist.*, vol. 37, no. 3, pp. 1566–1590, 06 2009. [Online]. Available: https://doi.org/10.1214/08-AOS622

[3] J. Lei, J. Robins, and L. Wasserman, "Distribution-free prediction sets," *Journal of the American Statistical Association*, vol. 108, no. 501, pp. 278–287, 2013. [Online]. Available: https://doi.org/10.1080/01621459.2012.751873

[4] J. Lei and L. Wasserman, "Distributionfree prediction bands for nonparametric regression," vol. 76, 01 2014.