1 Partial dependence

Let $F : R^p \to R$ denote the prediction function that maps the p-dimensional feature vector $\mathbf{x} = (x_1, \ldots, x_p)$ to its prediction. Furthermore, let $F_s(\mathbf{x}_s) = E_{\mathbf{x}_{\setminus s}}(F(\mathbf{x}_s, \mathbf{x}_{\setminus s}))$ be the partial dependence function of F on the feature subset $\mathbf{x}_s$, where $s \subseteq \{1, \ldots, p\}$, as introduced in [1]. Here, the expectation runs over the joint marginal distribution of features $\mathbf{x}_{\setminus s}$ not in $\mathbf{x}_s$.

Given data, $F_s(\mathbf{x}_s)$ can be estimated by the empirical partial dependence function

$$\hat{F}_s(\mathbf{x}_s) = \frac{1}{n} \sum_{i=1}^{n} F(\mathbf{x}_s, \mathbf{x}_{i\setminus s}),$$

where $\mathbf{x}_{i\setminus s}$, $i = 1, \ldots, n$, are the observed values of $\mathbf{x}_{\setminus s}$. Its disaggregated version is called *individual conditional expectation* (ICE), see [2].

2 Interaction statistics

2.1 Overall interaction strength

In [3], Friedman and Popescu introduced different statistics to measure interaction strength. Closely following their notation, we will summarize the main ideas.

If there are no interactions involving x_j, we can decompose the prediction function F into the sum of the partial dependence F_j on x_j and the partial dependence $F_{\setminus j}$ on all other features $\mathbf{x}_{\setminus j}$, i.e.,

$$F(\mathbf{x}) = F_j(x_j) + F_{\setminus j}(\mathbf{x}_{\setminus j}).$$

Correspondingly, Friedman and Popescu's statistic of overall interaction strength is given by

$$H_j^2 = \frac{\frac{1}{n} \sum_{i=1}^{n} \left[F(\mathbf{x}_i) - \hat{F}_j(x_{ij}) - \hat{F}_{\setminus j}(\mathbf{x}_{i\setminus j})\right]^2}{\frac{1}{n} \sum_{i=1}^{n} \left[F(\mathbf{x}_i)\right]^2}.$$

Remarks

1. Partial dependence functions (and F) are all centered to mean 0.
2. Partial dependence functions (and F) are evaluated over the data distribution. This is different to partial dependence plots, where one uses a fixed grid.
3. Weighted versions follow by replacing all arithmetic means by corresponding weighted means.
4. Multivariate predictions can be treated in a component-wise manner.
5. Due to (typically undesired) extrapolation effects of partial dependence functions, depending on the model, values above 1 may occur.

6. $H_j^2 = 0$ means there are no interactions associated with x_j. The higher the value, the more prediction variability comes from interactions with x_j.

7. Since the denominator is the same for all features, the values of the test statistics can be compared across features.

2.2 Pairwise interaction strength

Again following [3], if there are no interaction effects between features x_j and x_k, their two-dimensional partial dependence function F_{jk} can be written as the sum of the univariate partial dependencies, i.e.,

$$F_{jk}(x_j, x_k) = F_j(x_j) + F_k(x_k).$$

Correspondingly, Friedman and Popescu's statistic of pairwise interaction strength is defined as

$$H_{jk}^2 = \frac{A_{jk}}{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{F}_{jk}(x_{ij}, x_{ik})\right]^2}$$

where

$$A_{jk} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{F}_{jk}(x_{ij}, x_{ik}) - \hat{F}_j(x_{ij}) - \hat{F}_k(x_{ik})\right]^2.$$

Remarks

1. Remarks 1–5 of H_j^2 also apply here.

2. $H_{jk}^2 = 0$ means there are no interaction effects between x_j and x_k. The larger the value, the more of the joint effect of the two features comes from the interaction.

3. Since the denominator differs between variable pairs, unlike H_j, this test statistic is difficult to compare between variable pairs. If both main effects are very weak, a negligible interaction can get a high H_{jk}^2. Therefore, [3] suggests to calculate H_{jk}^2 only for *important* variables.

Modification: To be better able to compare pairwise interaction strength across variable pairs, and to overcome the problem mentioned in the last remark, we suggest as alternative the unnormalized test statistic on the scale of the predictions, i.e., $\sqrt{A_{jk}}$. Furthermore, we do pairwise calculations not for the most *important* features but rather for those features with *strongest overall interactions*.

2.3 Three-way interactions

[3] also describes a test statistic to measure three-way interactions: in case there are no three-way interactions between features x_j, x_k and x_l, their three-dimensional partial dependence function F_{jkl} can be decomposed into lower order terms:

$$F_{jkl}(x_j, x_k, x_l) = B_{jkl} - C_{jkl}$$

with

$$B_{jkl} = F_{jk}(x_j, x_k) + F_{jl}(x_j, x_l) + F_{kl}(x_k, x_l)$$

and

$$C_{jkl} = F_j(x_j) + F_k(x_k) + F_l(x_l).$$

The squared and scaled difference between the two sides of the equation leads to the statistic

$$H^2_{jkl} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{F}_{jkl}(x_{ij}, x_{ik}, x_{il}) - B^{(i)}_{jkl} + C^{(i)}_{jkl}\right]^2}{\frac{1}{n}\sum_{i=1}^{n}\hat{F}_{jkl}(x_{ij}, x_{ik}, x_{il})^2},$$

where

$$B^{(i)}_{jkl} = \hat{F}_{jk}(x_{ij}, x_{ik}) + \hat{F}_{jl}(x_{ij}, x_{il}) + \hat{F}_{kl}(x_{ik}, x_{il})$$

and

$$C^{(i)}_{jkl} = \hat{F}_j(x_{ij}) + \hat{F}_k(x_{ik}) + \hat{F}_l(x_{il}).$$

Similar remarks as for H^2_{jk} apply.

2.4 Total interaction strength of all variables together

If the model is additive in all features (no interactions), then

$$F(\mathbf{x}) = \sum_{j}^{p} F_j(x_j),$$

i.e., the (centered) predictions can be written as the sum of the (centered) main effects. To measure the relative amount of variability unexplained by all main effects, we can therefore study the test statistic of total interaction strength

$$H^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[F(\mathbf{x}_i) - \sum_{j=1}^{p}\hat{F}_j(x_{ij})\right]^2}{\frac{1}{n}\sum_{i=1}^{n}\left[F(\mathbf{x}_i)\right]^2}.$$

A value of 0 means there are no interaction effects at all. Due to (typically undesired) extrapolation effects of partial dependence functions, depending on the model, values above 1 may occur.

In [4], $1-H^2$ is called *additivity index*. A similar measure using accumulated local effects is discussed in [5].

2.5 Workflow

Calculation of all H_j^2 requires $O(n^2p)$ predictions, while calculating of all pairwise H_{jk} requires $O(n^2p^2)$ predictions. Therefore, we suggest to reduce the workflow in two important ways:

- Evaluate the statistics only on a subset of the data, e.g., on $n' = 300$ observations.
- Calculate H_j^2 for all features. Then, select a small number $m = O(\sqrt{p})$ of features with highest H_j^2 and do pairwise calculations only on this subset.

This leads to a total number of $O(n'^2p))$ predictions. If also three-way interactions are to be studied, m should be of the order $p^{1/3}$.

3 Variable importance

[6] proposed the standard deviation of the partial dependence function as a measure of variable importance.

Since the partial dependence function suppresses interaction effects, we propose a different measure in the spirit of the interaction statistics above: If x_j has no effects, the (centered) prediction function F equals the (centered) partial dependence $F_{\setminus j}$ on all other features $\mathbf{x}_{\setminus j}$, i.e.,

$$F(\mathbf{x}) = F_{\setminus j}(\mathbf{x}_{\setminus j}).$$

Therefore, the following measure of variable importance follows:

$$\text{Imp}_j = \frac{\frac{1}{n}\sum_{i=1}^n \left[F(\mathbf{x}_i) - \hat{F}_{\setminus j}(\mathbf{x}_{i\setminus j})\right]^2}{\frac{1}{n}\sum_{i=1}^n \left[F(\mathbf{x}_i)\right]^2}.$$

It differs from H_j^2 only by not subtracting the main effect of the j-th feature in the numerator. It can be read as the proportion of prediction variability unexplained by all other features. As such, it measures variable importance of the j-th feature, including its interaction effects.

4 Limitation

1. H-statistics are based on partial dependence estimates and are thus as good or bad as these. One of their problems is that the model is applied to unseen/impossible feature combinations. In extreme cases, H-statistics intended to be in the range between 0 and 1 can become larger than 1. Accumulated local effects (ALE) [7] mend above problem of partial dependence estimates. They, however, depend on the notion of "closeness", which is highly non-trivial in higher dimension and for discrete features.

2. Due to their computational complexity, H-statistics are usually evaluated on relatively small subsets of the training (or validation/test) data. Consequently, the estimates are typically not very robust. To get more robust results, increase the sample size.

References

[1] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *Ann. Statist.*, vol. 29, pp. 1189–1232, 10 2001.

[2] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015.

[3] J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 916–954, 2008.

[4] A. Żółkowski, M. Krzyziński, and P. Fijałkowski, "Methods for extraction of interactions from predictive models," Undergraduate thesis, Warsaw University of Technology, 2023.

[5] C. Molnar, G. Casalicchio, and B. Bischl, "Quantifying model complexity via functional decomposition for better post-hoc interpretability," in *Machine Learning and Knowledge Discovery in Databases* (P. Cellier and K. Driessens, eds.), pp. 193–204, Springer International Publishing, 2020.

[6] B. M. Greenwell, B. C. Boehmke, and A. J. McCarthy, "A simple and effective model-based variable importance measure," *ArXiv*, 2018.

[7] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," 2016.