General Introduction to Nonlinear Regression Models [1]

Trung Do
Mentor: Prof. Maryclare Griffin

University of Massachusetts Amherst

June 16, 2023

Presentation Outline

Machine Learning models interpretability

- In the data science field (particularly in biomedicine), there is a growing need of interpretation methods for ML models.
- Definition of model *interpretability*: the ability to explain a model in understandable terms to a human.
- Example of interpretable models are linear models. These linear models are provided with statistical significant measures such as p-values, posterior inclusion probabilities (PIPs), or Bayes factors.
- Linear models are limited/infeasible in modeling complex phenomena (responses' variation might be dominantly influenced by nonlinear interactions).

Machine Learning models interpretability

- Nonlinear ML models can overcome these limitation. Trade-off: lack of transparency in explaining the model learning process and decision-making behaviors.

2 classical strategies in achieving interpretability in machine learning methods:

- Intrinsic interpretability: Limiting the architecture used in the ML methods to simple structures (example, customized NN architecture based on biological system through experimental validation). Depends largely on reliable domain knowledge of the problem to come up with these architectures, thus challenging to generalize to other practical scientific problems.

Machine Learning models interpretability

- Auxiliary, or *post hoc* interpretability: assess the feature importance after the model has been trained. Two strategies in achieving interpretability of machine learning methods:
  1. Salience methods: essentially calculates the gradient of a loss function with respect to each of the feature. In practice, this method is shown to be highly unreliable in data with noise.
  2. "Sensitivity score": measure the variable importance by measuring the amount of predictive accuracy lost when that feature is perturbed (examples of this method: information criterion, distributional centrality measures, Shapley Additive Explanations (SHAP), and knockoffs)

Limitations of these methods

Limitation of these methods: only optimized for either (1) a global scale where the task is to rank the features based on their contribution to the overall variation of the given population, or (2) the local scale where the task is to understand how important a feature is to a particular individual in the dataset.

GOALS methods

- This paper presents the "GlObal And Local Score" (GOALS) operator, built off of the distributional centrality literature to provide simultaneous assessment of local and global importance for features.
- This paper uses the Gaussian process regression model, but the method is generally applicable to other models. The only requirements are (1) access to the fitted model and (2) the ability to generate out-of-sample predictions.
- Goals: show the flexibility and the efficient utility of GOALS over state-of-the-art variable importance strategies.

Simple, standard linear model

Given a dataset of N data points. Let $\mathbf{y}$ be an N-dimensional vector, $\mathbf{X}$ be an $N \times J$ matrix of feature columns, and $\boldsymbol{\beta}$ be the J-dimensional coefficient vector. Consider the simple, standard linear model:

$$\mathbf{y} = \boldsymbol{f} + \epsilon, \quad \boldsymbol{f} = \mathbf{X}\boldsymbol{\beta}, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}) \tag{1}$$

where $\boldsymbol{f}$ is the function to be estimated, and is assumed to be a linear combination of $\mathbf{X}$ and $\boldsymbol{\beta}$. ϵ is the normally distributed error term with mean 0 and scaled variance term τ^2, $\mathbf{I}$ is the $N \times N$ identity matrix.

Simple, standard linear model, continue

$$\mathbf{y} = \boldsymbol{f} + \epsilon, \quad \boldsymbol{f} = \mathbf{X}\boldsymbol{\beta}, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I})$$

- For convenience, y's are mean centered and standardized.
- Thus, $\mathbb{V}[\mathbf{y}] = 1$ and $\mathbb{V}[\mathbf{X}\boldsymbol{\beta}] = h^2$ is the variance explained by the features. The remaining variance is $\tau^2 = (1 - h^2)$
- In complex real world problems, y's might have a nonlinear relationship with the features, so using only linear (additive) relationship to describe the responses is restrictive. One way to overcome this limitation is perform model inference in high-dimensional function space.

"Weight-space" nonlinear model

- This study takes a nonparametric approach and conduct inference in reproducing kernel Hilbert space (RKHS) by specifying a Gaussian process (GP) prior:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k_\theta(\mathbf{x}, \mathbf{x}')) \tag{2}$$

where $f(\bullet)$ is defined by its mean function $m(\bullet)$, consider fixed at 0, and positive definite covariance function $k(\bullet, \bullet)$. When conditioning on a finite set of observations, the GP prior becomes a multivariate normal distribution.

"Weight-space" nonlinear model, continue

Thus we can define the "weight-space" nonlinear model for complex features:

$$\mathbf{y} = \boldsymbol{f} + \epsilon, \quad \boldsymbol{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}), \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I}) \tag{3}$$

where $\mathbf{f} = [f(x_1), f(x_2), ..., f(x_N)]$ is an N-dimensional normally distributed random variable with mean vector $\mathbf{0}$ and the covariance matrix $\mathbf{K}$ with each element given by $k_{ij} = k(x_i, x_j)$, x_i and x_j are the i-th and j-th observations, respectively.

- Without loss of generality, we use a radial basis function: $k_{ij} = \exp\{-\theta \|x_i - x_j\|^2\}$
- The GP regression model is the generalization of the linear mixed model which uses a nonlinear covariance $\mathbf{K}$ to account for the nonlinearity interaction between observations.

References

[1] E. T. Winn-Nuñez, M. Griffin, and L. Crawford, *A simple approach for local and global variable importance in nonlinear regression models*, 2023. arXiv: 2302.02024 [stat.ME].