Model Card for CB SH Lasso

The model is a penalized GLM with hierarchical constraints.

Model Details

Model Description

This section briefly describes the implementation of the generalized Lasso model with strong heredity constraints for four-way interactions, using a coordinate descent algorithm. To ensure numerical stability, each coefficient update step compares the new loss to the previous one. If there is no improvement, the previous estimate is kept. This guarantees that the algorithm is non-increasing and converges. As stated in the paper, the coefficient for the four-way interactions is updated entirely at once using the continuous Bernoulli Lasso. All other coefficients are updated component-wise. These univariate optimization problems are efficiently solved using standard optimization packages such as optimize.

- **Developed by:** Tatyana Krivobokova, Razvan-Andrei Morariu, Gianluca Finocchio
- **Model type:** Supervised learning: Hierarchical generalized linear model
- **Language:** R
- **License:** MIT

Uses

Direct Use

The model can be applied to any dataset that conforms to the structure described in Section 3 of the Supplementary material. of the paper . Check `Data_Analysis.R` for an example of how to run the model.

Out-of-Scope Use

The model should not be applied to data where the response cannot be reasonably modeled by the continuous Bernoulli distribution, and it is intended for datasets with four interacting factors. Moreover, the data matrix X should follow the form described in Section 3 of the Supplementary material.

Bias, Risks, and Limitations

From a technical perspective, the model is specifically tailored for data where hierarchical structure is expected. No significant sociotechnical biases or limitations have been identified at this stage. However, users should exercise caution and evaluate the model's suitability for their specific context.

Recommendations

Users should be aware of the model's assumptions regarding the response distribution and the hierarchical constraints. No further recommendations are available at this time.

How to Get Started with the Model

To get started, run `Data_Analysis.R` from Analysis folder to see how the model works.

Training Details

Training Data

The data is available at Data. There are 3960 values of the yield and each of 120 descriptors are available, which result from measurements under all possible $22 \times 15 \times 4 \times 3 = 3960$ combinations of reaction components.

Training Procedure

All the details can be found in the paper and the Supplementary material.

Speeds The running times for our model were measured on an **Intel® Core™ Ultra 7 155U 1.70 GHz** processor. All computations were performed in **R** (version 4.4.2, released on 2024-10-31, UCRT), without parallelization.

Training the hierarchical model from scratch on the full dataset takes between **1 and 15 hours**, depending on the amount of regularization and the convergence tolerance, which varies between 5×10^{-2} and 1×10^{-3}. Higher regularization levels lead to faster convergence due to the model's hierarchical constraints.

Training the model on the full dataset using the final regularization parameter selected by **AIC** takes approximately **3 hours** — this corresponds to the execution time of the `Data_Analysis.R`.

Evaluation

Testing Data, Factors & Metrics

Testing Data The data is available at Data. Details of the train/test split are given in the first section of the paper.

Factors The study comprises 3960 reactions and includes the following factors:

- **Isoxazole Additives**: 22 levels (originally 23; one was excluded from analysis)
- **Aromatic Halides**: 15 levels
- **Palladium Catalyst Ligands**: 4 levels
- **Bases**: 3 levels

These factors define the full factorial design used to generate the dataset.

Metrics The model is evaluated using **Root Mean Squared Error (RMSE)** and **R^2 score**. However, we note that the primary goal of the model is **accurate estimation and interpretability**, rather than maximizing predictive performance.

Results

See Section 2 of the paper and folder Results.

Summary The results provide insight into how interactions between different factor levels influence the yield of the reactions.

Environmental Impact

- **Hardware Type:** Personal laptop (Intel® Core™ Ultra 7 155U, 1.70 GHz)
- **Hours used:** Approximately 1–15 hours per full training run (depending on regularization)
- **Cloud Provider:** Not applicable (local computation)
- **Compute Region:** Not applicable (local computation)
- **Carbon Emitted:** Not formally estimated; expected to be minimal

Model Card Contact

For questions or feedback regarding this model, please contact:

- **Tatyana Krivobokova** — tatyana.krivobokova@univie.ac.at
- **Razvan-Andrei Morariu** — razvan-andrei.morariu@univie.ac.at