Linear Forest

Generating Linear Relationships with Random Linear Forests

In our package, we introduce a powerful tool called "random_linear_forest" that offers a unique approach to generating linear relationships between two variables while accounting for the influence of covariates. This innovative method combines the principles of random forests with linear tree modeling, resulting in a versatile and robust framework for understanding the intricate interplay between variables and their covariates.

Understanding the Concept:

At its core, the "random_linear_forest" algorithm leverages the strength of ensemble learning techniques seen in random forests, while focusing on modeling linear relationships. It builds on the existing Linear tree methods by capturing a more nuanced effects of covariates without overfitting.

Key Features:

Ensemble of Linear Tree Models: The algorithm creates an ensemble of linear tree models, each trained on a random subset of the dataset. By combining the predictions from multiple models, we gain a more stable and accurate representation of the linear relationships, reducing the risk of overfitting.

Covariate Integration: The inclusion of covariates is a distinguishing feature of our "random_linear_forest" package. Covariates play a crucial role in real-world scenarios, affecting the strength and nature of the linear relationship. Our algorithm takes these covariates into account, enabling a more comprehensive analysis that better reflects the underlying data dynamics.

Nonlinear Effects: Despite the emphasis on linear relationships, the ensemble nature of the algorithm can capture certain nonlinear effects that might be missed by traditional linear models. This additional flexibility enhances our ability to unveil hidden relationships within the data.

Applications:

The main use we found for this tecnique is estimating the impacts certain covariate factors have on the linear relationship Log Cases and Log concentration tend to have. We found that the generic factors (population, location, ect) and true covariates (Flow, PMMoV, BCov) can effect the relationship.

```
form <- conf_case ~ N1 + N2 | . - N1 - N2

forest_model <- random_linear_forest(data = na.roughfix(model_data),
                                     num_tree = 20,
                                     model_formula = form,
                                     max_depth = 3,
                                     verbose = FALSE)

forest_model
```

```
## [[1]]
## conf_case ~ N1 + N2 | regions + date + pop + PMMoV + flow + conductivity +
##     temperature + ph + tests
## <environment: 0x000001a782c5ecb0>
##
## [[2]]
## [1] "size of data: 9113"
##
## [[3]]
## [1] "Number of trees: 20"
##
## [[4]]
## [1] "Mean of squared residuals: 0.892557403879219"
##
## [[5]]
## [1] "% Var explained: 59.775460662509"
##
## attr(,"class")
## [1] "summary.random_linear_forest"
```

These functions are used in this analysis Random Linear Forst