Creating an Open-Source Ecosystem for Trustworthy AI in Julia

This supporting document provides brief descriptions of the various packages that form part of our Trustworthy AI ecosystem for Julia. We will use an extended version of this notebook for demo purposes (may take a few minutes to load).

1 ConformalPrediction.jl

ConformalPrediction.jl is a package for Predictive Uncertainty Quantification through Conformal Prediction for Machine Learning models trained in MLJ.jl.

> Conformal prediction (a.k.a. conformal inference) is a user-friendly paradigm for creating statistically rigorous uncertainty sets/intervals for the predictions of such models.
>
> — Angelopoulos and Bates (2021)

Intuitively, CP works under the premise of turning heuristic notions of uncertainty into rigorous uncertainty estimates through repeated sampling or the use of dedicated calibration data. Figure 1a demonstrates this notion for the regression case.

2 LaplaceRedux.jl

LaplaceRedux.jl is a package that facilitates Effortless Bayesian Deep Learning through Laplace Approximation for Deep Neural Networks built in `Flux.jl`. It implements the ideas presented in Daxberger et al. (2021). Figure 1b shows a Bayesian Prediction interval for a Deep Neural Network with Laplace Approximation that was built and trained in `LaplaceRedux.jl`.

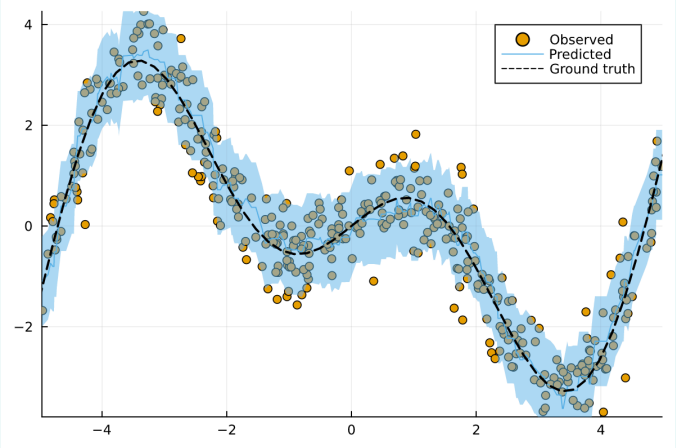
3 CounterfactualExplanations.jl

CounterfactualExplanations.jl is a package for generating Counterfactual Explanations and Algorithmic Recourse in Julia. In Figure 2a we have generated a Counterfactual Explanation for turning a cat into a dog (a toy example): as the tail length decreases and the height increases, the cat traverses through the feature space across the decision boundary of the underlying classifier. Figure 2b applies the same underlying principles to MNIST data: it demonstrates which pixels need to be perturbed in order for the underlying image classifier to predict 'four' (4) instead of 'nine' (9).

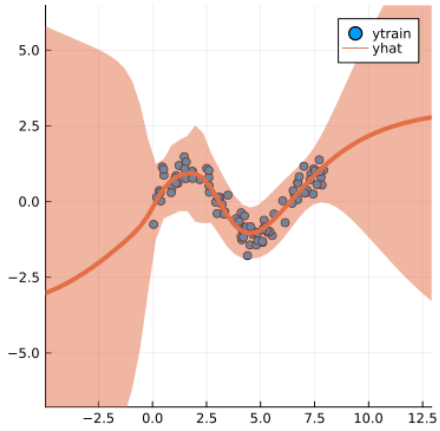
References

Angelopoulos, Anastasios N., and Stephen Bates. 2021. "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification." https://arxiv.org/abs/2107.07511.

Daxberger, Erik, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. 2021. "Laplace Redux-Effortless Bayesian Deep Learning." *Advances in Neural Information Processing Systems* 34.

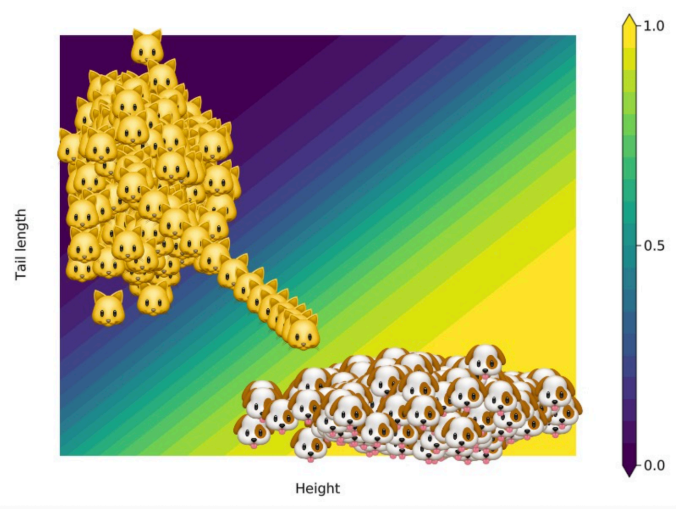


(a) Conformal Prediction interval for a Nearest Neighbour Regression model. Source: our blog post.

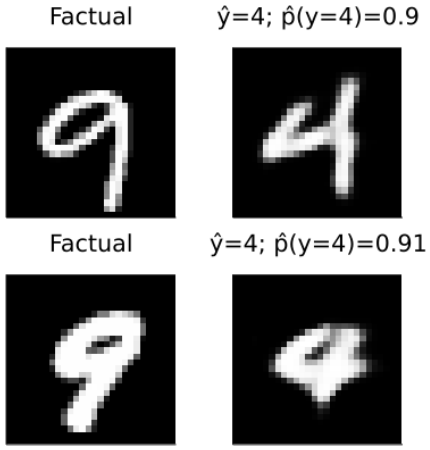


(b) Bayesian Prediction interval for a Deep Neural Network with Laplace Approximation. Source: package repository.

Figure 1: Frequentist and Bayesian approaches to Predictive Uncertainty Quantification.



(a) Turning a cat into a dog. Source: package repository.



(b) Turning a nine (9) into a four (4). Source: package repository.

Figure 2: Counterfactual Explanations in action.