Interpretable machine learning practices for species distribution models

Abstract

todo

## SDMs are an ML problem

* probabilistic classifiers
* optimal threshold for rangemap
* see Beery for the same argument aimed at ML prac
* can do ML even without doing DL
* lots of practices from ML already, we suggest a broader guideline for what/how to adopt
* in each section, introduce SDM concepts and map to ML vocabulary
* conclude each section on recommendations + nuance: when departing from ML best prac is acceptable

## Exploratory data analysis and the risk of data leakage

* data transformation issues + can lead to leakage
* overview of data transformation practices in the literature
* problem when changing the scale of prediction

## Imbalanced classification and the generation of background points

* classification: balance v. imbalance
* accuracy paradox + colinearity + issues with the status of negatives
* review recommendations about number of background points – lead to classification extremely unbalanced

## Cross-validation and model evaluation

* elements from Becker et al. on how to validate with positive only
* review different folding techniques
* unambiguous recommendation about MCC over everything else
* BUT examples with different acceptable biases for different uses

## Hyper-parameters and self-tuning models

* show example of learning the threshold (can even be done with BIOCLIM)
* how to tune many parameters at once?
* overview of tech solutions, *e.g.* wrapping models in a tuning model

## Interpretability of predictions

## Enabling what-if scenarios for predictive models

## Perspective: what SDM can contribute

* JSDM
* nested models 1 (tick require deer require vegetation)
* nested models 2 (abundance require presence)

## Conclusions