Interpretable machine learning practices for species distribution models

Abstract

todo

## SDMs are an ML problem

* probabilistic classifiers
* 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
* ROC and PR
* review recommendations about number of background points – lead to classification extremely unbalanced
* elements from Becker et al. on how to validate with positive only

## Cross-validation and model evaluation

* review different folding techniques
* unambiguous recommendation about MCC over everything else
* BUT examples with different acceptable biases for different uses

Although there are recommendations about the optimal classifier to select, the reality of *why* we predict the distribution of the species can lead to justified departures from this standard, by selecting a model that is not optimal; which is to say, there may be models with the same value of MCC that have different suitability depending on the intent behind the modeling process. For example, consider the following confusion matrices:

They have the exact same accuracy (0.91) and approximately the same MCC (0.83 for and , and 0.82 for ). Yet they represent drastically different types of errors, with being biased towards false negatives, towards false positives, and being equally biased towards false positives/negatives. Depending on the question, the classifiers that these matrices represent would differ in their suitability. To monitor an invasive species, where missing a detection event is costly, would be more appropriate: its prediction that a species is absent are far more reliable. Conversely, to monitor a threatened or a rare species, where sampling sites outside of its distribution is a misallocation of effort, the classifier represented by would be more reliable: its positive recommendations are more trustworthy.

This simplistic example highlights how even within a family of models with similar predictive performance, there can be variation in their suitability for a specific task. The point of cross-validation is to obtain the best model, but “best” is a relative notion that must be framed by domain knowledge about how the prediction is ultimately turned into actions.

## Hyper-parameters and self-tuning models

* optimal threshold for rangemap
* 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

* Shapley values + LIME
* overview of literature on policy based on ML
* framework: MTL declaration
* examples of mapping – see the here be dragons paper

## Enabling what-if scenarios for predictive models

* counterfactuals theory
* example of guiding actions: how can we flip a prediction
* types of model that support this

## Perspective: what SDM can contribute

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

## Conclusions