Machine Learning Guidelines for Natural Resource Management Practitioners

Shih-Ni Prim and Natalie Nelson

2024-02-15

Contents

Chapter 1

Motivation

As machine learning (ML) has become a powerful tool, it is noted by some that ML has not been widely used in environmental studies. This booklet is meant to provide a concise guide for natural resource management practitioners. This book serves as a staring point rather than a comprehensive resource, so that practitioners can have a basic understanding of how ML works and how to utilize it to analyze data and answer research questions. When appropriate, we provide case studies and R code as well as other online resources to help the readers on the journey of gaining one powerful tool that seems to be omnipresent in the research world.

Chapter 2

Introduction

What is machine learning? Essentially, machine learning teaches computer models to look for patterns or make predictions. This might sound like magic or it might seem complicated, but you can think of machine learning models as finding underlying formulas that the data come from. To solve for such formula, many, many mathematical calculations are involved. As we human beings are prone to mistakes, as long as we can identify a framework, we can give the framework and data to a computer model. It is best at repeating meticulous calculations to find a best guess based on our believes of the system and the data we observed.

2.1 Supervised Learning

2.2 Unsupervised Learning

Chapter 3

Data

3.1 What to do with data?

One could argue that data is the single most important ingredient when it comes to machine learning models or any type of analysis. As one might say, junk in, junk out.

3.2 Data Requirement

Chapter 4

Evaluation

4.1 Training vs Testing

We should first address the concepts of training and testing.

4.2 Metrics

4.2.1 Continuous Responses

4.2.2 Discrete Responses

4.3 Cross Validation

One very standard way of evaluation is k-fold cross validation, commonly with $k = 5$ or $k = 10$. The idea is simple. Divide the data into k groups. Each time, choose $k - 1$ groups for training, fit the model on the last group, which is the test data, and calculate the desired metrics, such as MSE.

In this way, although less data is used for training, the metrics are more accurate, because now we are not using the same data points for training and testing. Using metrics from cross validation for model selection can ensure that your model does not overfit, which means the model does well with training data but does not generalize well on new data.

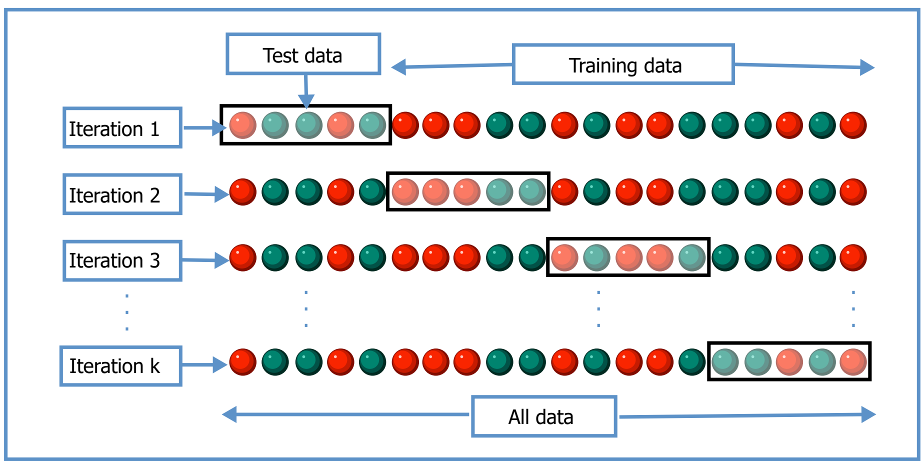

Figure 4.1: Image Source: https://en.wikipedia.org/wiki/Cross-validation_(statistics)

Chapter 5

Machine Learning Methods

Here we provide a list of commonly used machine learning methods and some brief discussion.

5.1 Random Forest

Chapter 6

Presentation

It is also important to present the results in a way that aids rather than impede communication.

6.1 Table

6.2 Figure

Chapter 7

Ethical Considerations

7.1 Reproducibility

To allow for others to reproduce your work, it is important to provide enough details in terms of methods, data processing, code implementation, etc. It is also encouraged to have all the code and data available online in a repository. If parts or all of the data should not be shared publicly, it helps to provide a simulated data set.

7.2 Decision making

Since research related to environmental sciences and natural resources could likely affect decision making, we will now address some topics.

7.2.1 Uncertain qualification

While all ML models can provide point estimates, not all can quantify uncertainty. It is, however, important to show how confident the model is about the estimates. If the conclusion of the study could affect an important policy change, it is crucial to present a full picture of the findings, which include uncertainty quantification. It is wildly different whether the model is 20% or 95% confident about its answer, for instance.

7.2.2 Interpretability

While some models, such as neural networks, are highly efficient, they are more like black boxes and do not lend easily to interpretablity. In the case of neural networks, even if you are able to find all the weights in the hidden layers, there is really no way to interpret them. To ensure that the model arrives at a reasonable conclusion, you might consider using a model that is more interpretable, such as linear regression or tree-based methods. This way, experts with domain knowledge can examine whether the conclusion makes sense. In other words, the findings from ML models can add to researchers' understanding of the field rather than throwing out an answer that is not easily interpreted.

Chapter 8

Appendix

8.1 Do's and Don'ts