A Machine Learning Primer for Natural Resource Management

Natalie Nelson, Shih-Ni Prim, et al.

2024-10-15

Contents

Chapter 1

Welcome

From satellites to field-deployable sensors, the entire agricultural and environmental landscape is increasingly monitored. Advances in sensing allow for a wide range of variables to be measured at unprecedented rates and scales. The large amounts of data produced from agricultural and environmental systems are paving the way for data-intensive methods like machine learning and Artificial Intelligence to support decision-making in natural resources management. But these opportunities also create new educational needs, particularly among applied scientists and engineers who may not have received formal training in "data science" methods (Saia et al., 2021), which is why we made this primer.

1.1 How to use this primer

This primer serves as a succinct guidebook on using machine learning in natural resources management. It is written as a starting point, and not a comprehensive resource.

All readers should start with chapter 2, which defines and describes "machine learning". If you are not interested in developing machine learning models, and instead want to understand how to interpret machine learning models or critique tools that use machine learning, continue reading through chapters 3 and 4. If you are interested in seeing an example workflow for developing a machine learning model, skip to chapter 5.

1.2 Learning how to code

For those of you who are interested in developing machine learning models, the material presented in this primer will help show you where to start. In chapter

5, we include an example using code prepared in R, an open source statistical software environment. To learn more about coding in R, we recommend R for Data Science.

1.3 Acknowledgements

Support for this primer was provided by the U.S. Department of Agriculture National Institute of Food and Agriculture grant number 2019-67021-29936.

1.4 Citation

Nelson, N.G., Prim. S., others? (2024), Machine Learning in Natural Resources Management: A Primer, Accessed online via https://nataliegnelson.github.io/ml_primer/.

1.5 License

Chapter 2

What is machine learning?

Machine learning teaches computer models to make predictions from patterns in data (ref). To find patterns in data, many, many mathematical calculations are involved. Computers are needed to develop machine learning models, and advances in computing gave rise to machine learning (ref).

2.1 How is machine learning useful for natural resources management?

Historically, environmental and agricultural systems were primarily modeled with "process-based" or "mechanistic" models, which are models that simulate key system processes based on their underlying physics, chemistry, or biology (ref). Models were often used to fill gaps in direct measurements of various environmental and agricultural phenomena. In the past, collecting any type of environmental or agricultural data was grueling, time consuming, and done by people. Today, some variables still require tough and tedious work to measure, but many variables are now readily measured with sensors and other automated instruments. As an example of how process-based models are used, let's consider the flow of water through an open channel. If you were estimating the velocity of water through an open channel using a process-based approach, you may use Manning's Equation, which simulates water velocity as a function of the channel's physical properties like its slope and cross sectional area. To apply Manning's Equation, you would have to take measurements of the channel's dimensions.

However, with machine learning, you do not have to simulate environmental and agricultural variables using underlying processes. Instead, you can create a model that makes predictions from patterns in data. Machine learning models require that you have some measurements of your response or target variable,

which is the variable you are seeking to estimate. In our current example, water velocity in the channel is the target variable. You also need measurements of predictor variables, or variables that you will use to predict your response.

As one hypothetical example, a machine learning model of water velocity in an open channel might make predictions based on images instead of physical channel properties. A camera could be situated towards the channel to take time-lapse photos every few minutes. A machine learning model could then be developed to predict water velocity based on the appearance of the water in the images. For a real-world example of this, check out Chapman et al. (2024), Stage and discharge prediction from documentary time-lapse imagery, in PLoS ONE.

To create such a machine learning model, water velocity measurements would be needed to create a training set, or data used to train or develop the machine learning model. The water velocity measurements would be collected at the same time as the images, so the images could then be directly related to water velocity measurements (i.e., when the image looks like this, the water velocity is that). The predictor variables would be derived from the images. A machine learning algorithm could be selected to search for patterns between the images – e.g., the shading/color of individual image pixels, relationships between neighboring pixels – and the water velocity measurements. Once these patterns are established, they can be used to estimate water velocity from new images as they are collected. Importantly, the image-based machine learning model knows nothing about the underlying physics controlling water velocity in the channel; it has simply learned that certain patterns in the images correspond to higher or lower water velocities.

Machine learning models can easily be expanded to include many different types of predictor variables. For example, the image-based machine learning model for water velocity could be further built out to include additional predictors like rainfall, irrigation, and time of year. The flexibility of machine learning models allows them to consider many diverse streams of information when learning patterns from which to make predictions. To see additional examples of machine learning models developed for natural resource management applications, see chapter 4.

2.2 Pros and cons of machine learning

Because they make predictions from correlative patterns between the predictor and response variables, machine learning models have parallels to simpler statistical models like linear regression (ref). However, machine learning models consist of hundreds, thousands, or millions of individual equations, while a linear regression model consists of only one equation. Because of their many equations, machine learning models are commonly considered “black boxes” since peering

into a machine learning model can feel like looking into a pitch black box – you don't precisely know what's inside (ref). Methods now exist to help us illuminate machine learning black boxes, and the field of "interpretable" or "explainable" machine learning has made great strides in support of better understanding the inner workings of these models (ref). Still, machine learning models are substantially more challenging to interpret than most other model alternatives.

While challenges to interpretation are a clear pitfall of machine learning models, their complicated structures allow for them to pick up on relatively subtle or nuanced relationships between datasets, making them effective predictive tools (ref). In many cases, machine learning models outperform process-based models (e.g., their predictions have greater accuracy than the predictions from other models) (ref). The ability to create strong predictions is arguably the hallmark strength of machine learning. The previously described flexibility of being able to include many different diverse data types (e.g., images, sensor data, weather station observations) as predictors is also a key strength. But, because they do not necessarily account for underlying processes, machine learning models are at risk of spurious predictions, particularly if developed irresponsibly. In chapter 3, we include questions you can ask to evaluate whether a model was developed responsibly, and assess whether it is vulnerable to making spurious predictions.

2.3 Broad classifications of machine learning models

In the previous example on image-based predictions of water velocity, we summarized the use of a supervised learning approach. Simply put, supervised learning is when there are true answers for the model, or there are measurements for the response variable (ref). In the water velocity model example, water velocity measurements were used during model development to facilitate the learning of patterns in the images that could specifically be used to predict water velocity. Once a supervised learning model is developed, the predictions from the model can be compared with the measurements to assess how well the model predicts.

Unsupervised learning, on the other hand, performs tasks that do not have answers for the model. For example, instead of using the images of the stream to estimate water velocity, an unsupervised machine learning model could be used to group or cluster images with shared similarities. Unsupervised learning is often used as a computer-assisted way of exploring patterns in data (ref). Sometimes, the groups or clusters uncovered by an unsupervised learning model can be assigned labels that are of use for other modeling or analysis efforts.

2.4 An example of supervised and unsupervised learning: The history of the National Land Cover Database

The National Land Cover Database or Dataset (NLCD) is created by the U.S. Geological Survey and its partners to map land cover across the contiguous U.S. The data are updated every few years, creating a historical record of land cover change across the country over time. In the 1970s and 1980s, land cover mapping was performed by manually delineating areas from aerial photographs. In 1992, the NLCD premiered a land cover data product that was created from Landsat satellite imagery. Landsat imagery has pixels that are 30 meters by 30 meters. To create the 1992 product, NLCD imagery was clustered into 100 groups using an unsupervised learning approach, and people evaluated the 100 groups and manually assigned them to land cover categories (e.g., forest, developed area). Later, to create the 2001 NLCD product, a supervised learning approach was used in which areas of known land cover were used to train a model to predict land cover categories based on Landsat imagery. Read more about the history of the NLCD program in Chapter 18 of The Nature of Geographic Information (DiBiase et al.)[https://www.e-education.psu.edu/natureofgeoinfo/].

2.5 Machine learning algorithms

Within the two broad classifications of machine learning (supervised and unsupervised), there is an overwhelming number of machine learning algorithms, or specific computer model frameworks for implementing machine learning. The sheer number of machine learning algorithms attests to its power and versatility. This primer does not catalog different machine learning algorithms. As a starting point, you can see some algorithm examples at the resources below. Which machine learning algorithm should I use by SAS blogs Machine Learning Cheat Sheet by Datacamp

2.6 A deeper dive into machine learning

If you would like to learn more about machine learning, check out An Introduction to Statistical Learning by James et al. (2013). This book provides a wealth of information on machine learning models, and the authors assume the readers are mainly interested in applying, rather than deeply studying, machine learning models. The four premises offered by the authors, listed on pp 8-9, demonstrate the practical focus of the book: > Many statistical learning methods are relevant and useful in a wide range of academic and non-academic disciplines, beyond just the statistical sciences. > Statistical learning should not

be viewed as a series of black boxes. > While it is important to know what job is performed by each cog, it is not necessary to have the skills to construct the machine inside the box! > We presume that the reader is interested in applying statistical learning methods to real-world problems.

There are free PDF versions of the books available online, and examples are presented in both R and Python.