Machine learning: The basics

Russell J. Funk

Carlson School of Management
University of Minnesota

April 17, 2020

Roadmap

- Motivation
- Concepts
- Computation
- Organizational research
- Frontiers

Motivation

Why machine learning?

Recent years have witnessed an explosion in the availability of data for research.

- social media (Park et al, 2018)
- electronic sensors (Pentland, 2014; Kabo et al., 2015)
- administrative records (Kossinets & Watts, 2009; Landon et al., 2012)
- government documents (Kosack et al., 2018)

In addition to greater availability, data is also bigger, in two senses.

- in the rows (i.e., more observations)
- in the columns (i.e., more variables)

While more and richer data is great, it strains the analytical capacities of researchers.

What is machine learning?

- Machine learning is a set of techniques that allow computers to learn from experience.
- Typically, when we say "experience," what we mean is data.

Machine learning and the growth of data

- On the one hand, machine learning has benefitted from growth of data because many algorithms require very large amounts of data to perform well.
- On the other hand, the growth of data has benefitted from machine learning because machine learning facilitates processing of very large amounts of data.

Concepts

There are two main branches of machine learning

Supervised learning

- Given training data of paired inputs (X) and outputs (Y) to learn a function that predicts Y from X in previously unseen data.
  - When Y is continuous, we're doing regression.
  - When Y is categorical, we're doing classification.
- Training data is both what gives supervised learning its name and what sets supervised learning apart from other approaches to machine learning.

Approaches and algorithms

- linear regression
- logistic regression
- k-nearest neighbors
- naive Bayes
- neural networks

Examples

- image classification (Does this picture contain a cat?)
- recommendation (Would this person like this movie?)
- prediction (Is this person likely to default on their loan?)

There are two main branches of machine learning

Unsupervised learning

- Given data on inputs X, learn a function that summarizes or characterizes meaningful patterns in those data.
- Unlike supervised learning, unsupervised learning does not rely on training data.

Approaches and algortihms

- k means
- hierarchical clustering
- principal component analysis
- singular value decomposition
- multidimensional scaling
- topic models

Examples

- clustering (Are there similar groups of music fans based on tastes?)
- dimension reduction (Can a few dimensions characterize neighborhood disadvantage?)
- community detection (Does a social network cluster into meaningful groups?)

Machine learning researchers distinguish between a few categories of data

Training data

- Data used to create a model (e.g., labeled examples).

Validation data

- Data used to help tune model parameters (helps to avoid overfitting the training data).

Test data

- Data used to evaluate the performance of the model (e.g., held out from the training data).

Computation

What's the connection to computational social science?

At a surface level. . .

- Machine learning is, of course, very computationally intensive.
- All about using computers to advance scientific discovery.

At a deeper level. . .

- Machine learning embodies the inductive orientation of computational social science.
- All about letting the data speak, and helping us to see hidden structure.
- Helpful for understanding complex interactions among our observations and variables.

Organizational research

Machine learning and organizational research

- To some degree, organizational research has used machine learning for a long time.
- However, we are starting to see some broader uses as well.

Prediction and model development

- Supervised learning is particularly valuable anytime our primary focus is on prediction over theory testing.
- Along those lines, we're seeing machine learning used for things like matching.
- Machine learning is also useful when we may have complex relationships among our variables (e.g., many interactions).

Data cleaning and coding

- Machine learning methods have also been increasingly used in organizational research for "backstage" work.
- For example, we can train models to help us with data cleaning (e.g., de-duplication).
- We can also use machine learning for coding observations on scales that would not be possible for humans.

Hidden structure

- Probably one of the oldest uses of machine learning in organizational research is for finding hidden structure in our date (e.g., PCA, factor analysis).

Frontiers

Where is machine learning going next?

My $0.02

Deep learning

- So far, most applications of deep learning have been in industry.
- We'll likely start to see more applications in science, especially for prediction tasks (e.g., matching, instruments).

Theory building

- We're seeing increasing adoption of machine learning among traditionally qualitative, inductive researchers.
- There is growing interest in using machine learning for theory building.

Opening the black box

- Many advanced machine learning models perform well, but what happens under the hood is a black box (to mix two metaphors).
- That may limit their value from a scientific perspective.

Prediction

- My hunch is that we'll see new standards and expectations for matching over the next few years.

Appendix