Machine Learning

Kristen Vaccaro
September 2021

Introduction

Machine learning (ML) is a branch of artificial intelligence that is closely related to the fields of statistics and data science. Machine learning algorithms determine which items are recommended to you when you shop online, which emails are filtered out of your inbox as spam, and which words your messaging app suggests when you begin typing.

Building these systems—or any machine learning system—requires *data*. There are many different kinds of data that machine learning algorithms may use; Table 1 provides examples of several common data types.

Data Type	Example
Binary	Is Mary Anne a UCSD student? (Yes)
Nominal	In which engineering department is Mary Anne a student? (CSE)
Ordinal	What size coffee does Mary Anne order? (medium)
Discrete	How many classes has Mary Anne taken at UCSD? (18)
Continuous	How tall is Mary Anne? (5.43 ft)

Table 1: Common data types

Two common ML tasks: classification & regression

1 *Classification:* Spam detection is an example of a classification task. A spam detection model takes an input that represents an email. Its output is typically binary—spam or not spam. More generally, the job of a classification model is to assign categories to its inputs. There may be only two categories (binary classification) or there may be multiple categories (multiclass/multinomial classification).

2 *Regression:* A model that predicts the cost of rent given the location and size of an apartment is performing a regression task. More generally, the job of a regression model is to assign a numerical output to each of its inputs. Other examples would include predicting a person's weight based on their age and height, predicting a student's exam score based on their homework scores, etc.

Example: Rent Prediction

The inputs to an ML model are typically called *features*, and the outputs are often called *labels* or *targets*. Table 2 provides an example of a dataset that could be used to train a regression model for predicting the cost of rent. It is not always clear which features should be used to represent a particular input; this choice may require insight from subject matter experts.

Input		Output
Size	On/Off Campus	Rent
275	1	900
720	0	1800
950	1	2110
1000	0	3000
1200	0	3800

Table 2: Dataset for predicting rent (target) based on apartment size and location (features)

A variety of algorithms exist for constructing machine learning models, but the general process works as follows. We use a labeled dataset (like the one represented in Table 2) to *train* a machine learning model. Then this model can take new inputs and produce outputs. Figure 1 illustrates this process.

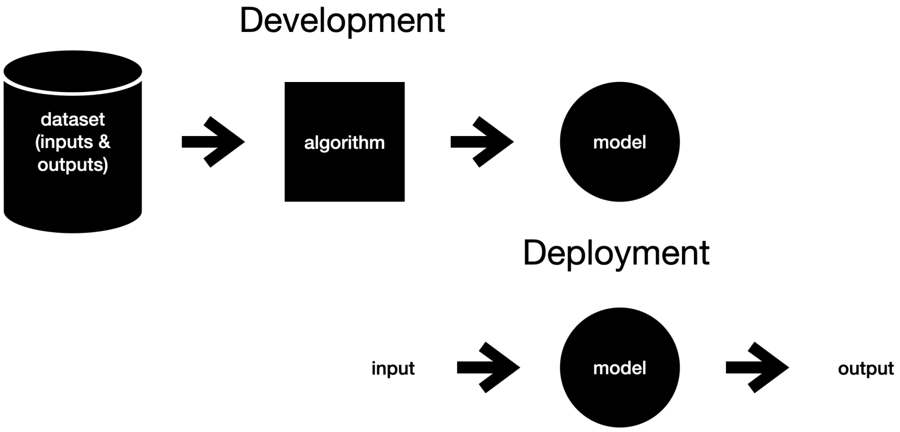

Figure 1: Machine learning algorithms can produce a model given a labeled dataset. This model can then take new inputs and produce corresponding outputs.

Other resources

- Google's Machine Learning Glossary:
  https://developers.google.com/machine-learning/glossary
- ML Cheatsheet Glossary:
  https://ml-cheatsheet.readthedocs.io/en/latest/glossary.html