CHAPTER 15

Batch vs. Online Learning

Data is a vital component for building learning models. There are two design choices for how data is used in the modeling pipeline. The first is to build your learning model with data at rest (batch learning), and the other is when the data is flowing in streams into the learning algorithm (online learning). This flow can be as individual sample points in your dataset, or it can be in small batch sizes. Let's briefly discuss these concepts.

Batch Learning

In batch learning the machine learning model is trained using the entire dataset that is available at a certain point in time. Once we have a model that performs well on the test set, the model is shipped for production and thus learning ends. This process is also called *offline learning*. If in the process of time, new data becomes available, and there is need to update the model based on the new data, the model is trained from scratch all over again using both the previous data samples and the new data samples.

This pipeline is further illustrated in Figure 15-1.

E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*,
https://doi.org/10.1007/978-1-4842-4470-8_15

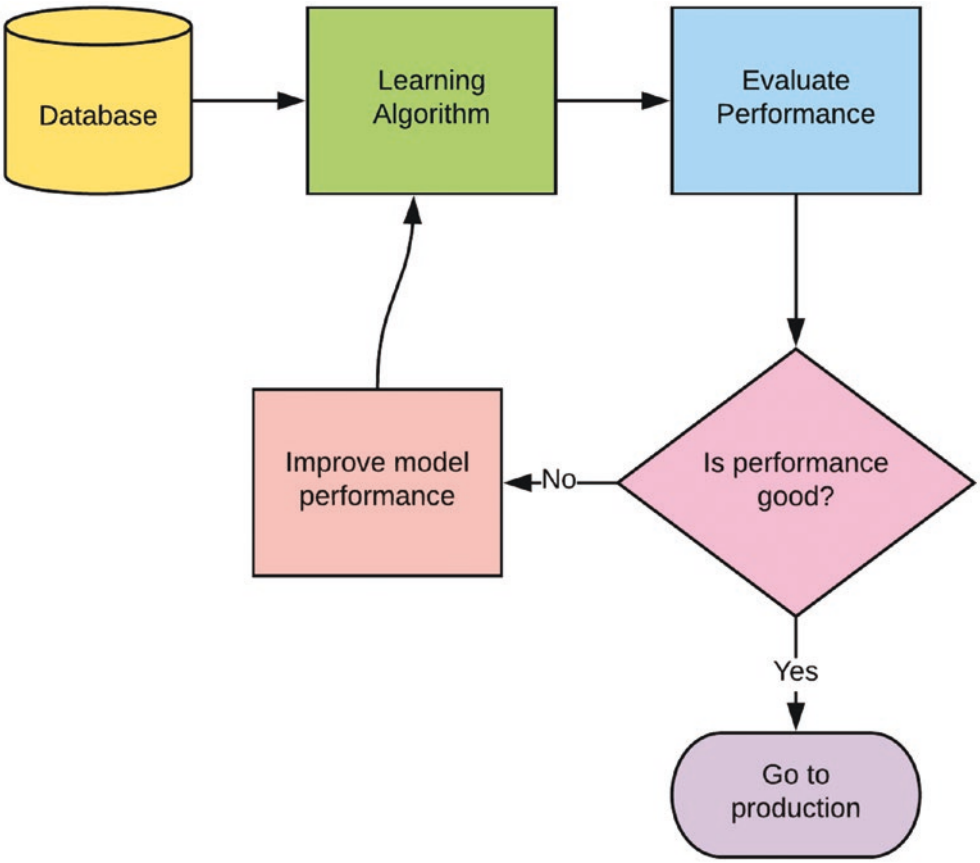

Figure 15-1. *Batch learning*

In a situation where there is a need to train the model with data that is generated continuously from the source, batch learning becomes inappropriate to deal with that situation. In such a circumstance, we want to be able to update our learning model on the go, based on the new data samples that are available.

Online Learning

In online learning, data *streams* (either individually or in mini-batches) into the learning algorithm and updates the model. Online learning is ideal in situations where data is generated continuously in time, and we need to use real-time data samples to build a prediction model. A typical example of this case is in stock market prediction.

Online learning is illustrated in Figure 15-2.