Introduction to Online Machine Learning : Simplified

BIG DATA (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/BIG-DATA/) BUSINESS ANALYTICS (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/BUSINESS-ANALYTICS/) PYTHON (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/PYTHON-2/) R (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/R/)

Data is being generated in huge quantities everywhere. Twitter generates 12 + TB of data every day, Facebook generates 25 + TB of data everyday and Google generates much more than these quantities everyday. Given that such data is being produced everyday, we need to build tools to handle data with high

1. Volume : High volume of data are stored today for any industry. Conventional models on such huge data are infeasible.

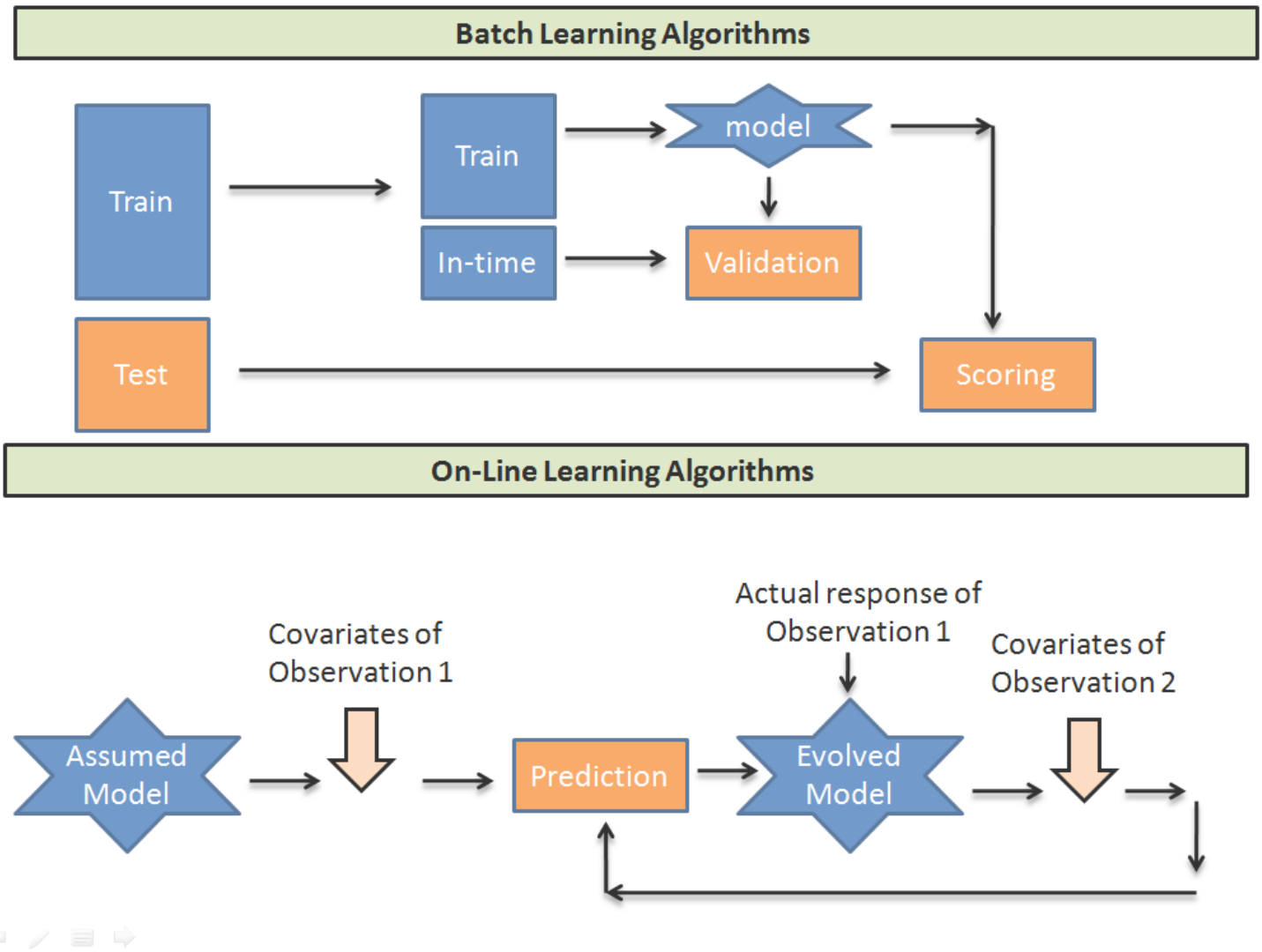
2. Velocity : Data come at high speed and demand quicker learning algorithms.

3. Variety : Different sources of data have different structures. All these data contribute to prediction. A good algorithm can take in such variety of data.

A simple predictive algorithm like Random Forest on about 50 thousand data points and 100 dimensions take 10 minutes to execute on a 12 GB RAM machine. Problems with hundreds of millions of observation is simply impossible to solve using such machines. Hence, we are left with only two options : Use a stronger machine or change the way a predictive algorithm works. First option is not always feasible. In this article we will learn about On-line Learning algorithms which are meant to handle data with such high Volume and Velocity with limited performance machines.

How does On-line learning differ from batch learning algorithms?

If you are a starter in the analytics industry, all you would have probably heard of will fall under batch learning category. Let's try to visualize how the working of the two differ from each other.



(https://www.analyticsvidhya.com/blog/wp-content/uploads/2015/01/schematics.png)

Batch learning algorithms take batches of training data to train a model. Then predicts the test sample using the found relationship. Whereas, On-line learning algorithms take an initial guess model and then picks up one-one observation from the training population and recalibrates the weights on each input parameter. Here are a few trade-offs in using the two algorithms.

- **Computationally much faster and more space efficient.** In the online model, you are allowed to make exactly one pass on your data, so these algorithms are typically much faster than their batch learning equivalents, since most batch learning algorithms are multi-pass. Also, since you can't reconsider your previous examples, you typically do not store them for access later in the learning procedure, meaning that you tend to use a smaller memory footprint.
- **Usually easier to implement.** Since the online model makes one pass over the data, we end up processing one example at a time, sequentially, as they come in from the stream. This usually simplifies the algorithm, if you're doing so from scratch.
- **More difficult to maintain in production.** Deploying online algorithms in production typically requires that you have something constantly passing datapoints to your algorithm. If your data

changes and your feature selectors are no longer producing useful output, or if there is major network latency between the servers of your feature selectors, or one of those servers goes down, or really, any number of other things, your learner tanks and your output is garbage. Making sure all of this is running ok can be a trial.

- **More difficult to evaluate online.** In online learning, we can't hold out a "test" set for evaluation because we're making no distributional assumptions — if we picked a set to evaluate, we would be assuming that the test set is representative of the data we're operating on, and that is a distributional assumption. Since, in the most general case, there's no way to get a representative set that characterizes your data, your only option (again, in the most general case) is to simply look at how well the algorithm has been doing recently.
- **Usually more difficult to get "right".** As we saw in the last point, online evaluation of the learner is hard. For similar reasons, it can be very hard to get the algorithm to behave "correctly" on an automatic basis. It can be hard to diagnose whether your algorithm or your infrastructure is misbehaving.

In cases where we deal with huge data, we are left with no choice but to use online learning algorithms. The only other option is to do a batch learning on a smaller sample.

Example Case to understand the concept

We want to predict the probability that it will rain today. We have a panel of 11 people who predict the class : Rain and non-rain on different parameters. We need to design an algorithm to predict the probability. Let us first initialize a few denotions.

i are individual predictors

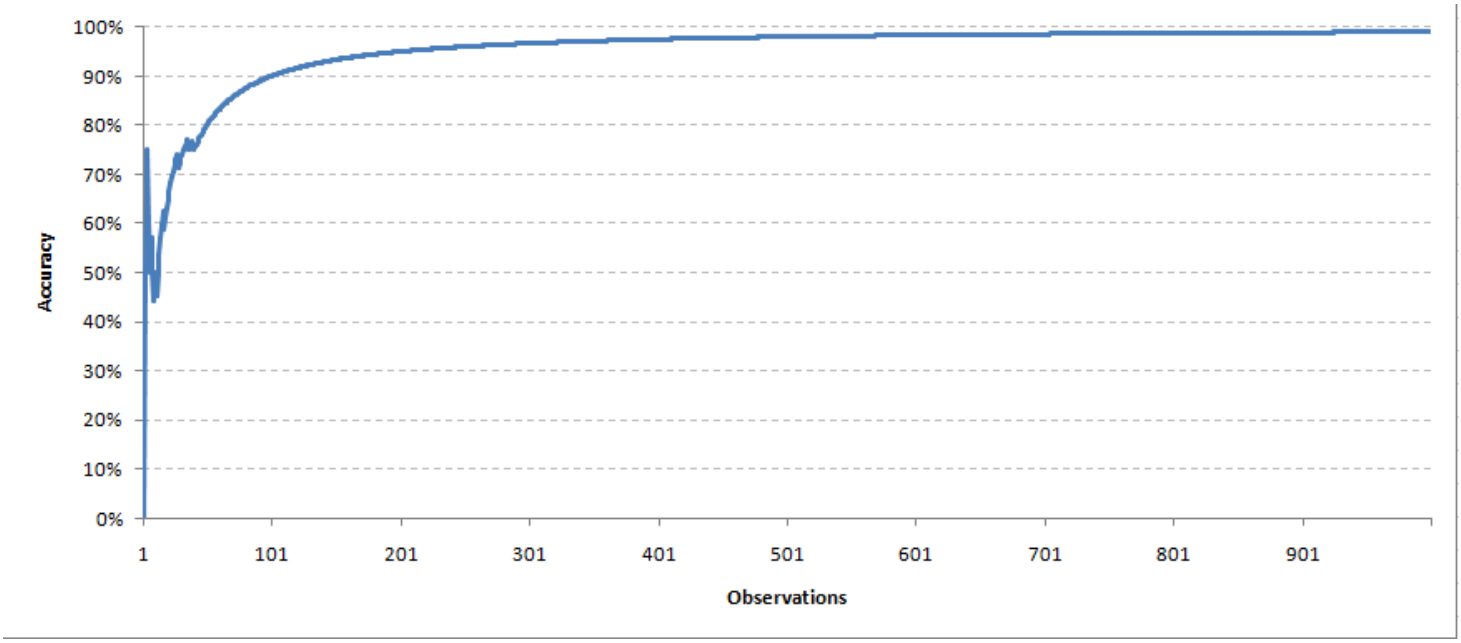
w(i) is the weight given to the i th predictor

Initial w(i) for i in [1,11] are all 1

We will predict that it will rain today if,

Sum(w(i) for all rain prediction) > Sum(w(i) for all non rain prediction)

Once, we have the actual response of the target variable, we now send a feedback on the weights of all the parameters. In this case we will take a very simple feedback mechanism. For every right prediction, we will keep the weight of the predictor same. While for every wrong prediction, we divide the weight of the predictor by 1.2 (learning rate). With time we expect the model to converge with a right set of parameters.We created a simulation with 1000 predictions done by each of the 11 predictors. Here is how our accuracy curve came out,



(https://www.analyticsvidhya.com/blog/wp-content/uploads/2015/01/accuracy.png)

Each observation was taken at a time to re adjust the weights. Same way we will make predictions for the future data points.

End Notes

Online learning algorithms are widely used by E-commerce and social networking industry. It is not only fast but also has the capability to capture any new trend visible in with time. A variety of feedback systems and converging algorithms are presently available which should be selected as per the requirements. In some of the following articles, we will also take up a few practical examples of Online learning algorithm applications.

Did you find the article useful? Have you used online learning algorithms before ? Share with us any such experiences. Do let us know your thoughts about this article in the box below.

If you like what you just read & want to continue your analytics learning, subscribe to our emails (http://feedburner.google.com/fb/a/mailverify?uri=analyticsvidhya), follow us on twitter (http://twitter.com/analyticsvidhya) or like our facebook page (http://facebook.com/analyticsvidhya).

Share this:

RELATED

Data Scientist Innoplexus Gurgaon/Pune (2-5 years of Experience) (https://www.analyticsvidhya.com/blog/2015/07/data-scientist-innoplexus-gurgaon-2-5-years-experience/)

Designation – Data Scientist Location – Gurgaon/Pune About employer – Innoplexus Job description: Responsibilities Design & develop algorithms and

In "Jobs"

(https://www.analyticsvidhya.com/blog/2015/12/year-review-analytics-vidhya-from-2015/)

Year in Review: Best of Analytics Vidhya from 2015 (https://www.analyticsvidhya.com/blog/2015/12/year-review-analytics-vidhya-from-2015/)

In "Business Analytics"

(https://www.analyticsvidhya.com/blog/2015/06/machine-learning-basics/)

Machine Learning basics for a newbie (https://www.analyticsvidhya.com/blog/2015/06/machine-learning-basics/)

In "Business Analytics"

TAGS: BATCH LEARNING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/BATCH-LEARNING/), FAST LEARNING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/FAST-LEARNING/), ON-LINE MACHINE LEARNING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/ON-LINE-MACHINE-LEARNING/), VOWPAL WABBIT (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/VOWPAL-WABBIT/)

Next Article

How to create Box-Plot chart in Qlikview? (https://www.analyticsvidhya.com/blog/2015/01/create-box-plot-qlikview/)