The story of MNIST and the perils of engineered data

Part 2: A primer on IID, exchangeability, and Bayesian causality.

ROBERT OSAZUWA NESS

APR 10, 2021

4

Share

In the previous post, I referred to a paper by Bengio, Scholkopf, and others that cast the reliability problems of deep learning as IID problems.

> Machine learning often disregards information that animals use heavily: interventions in the world, domain shifts, temporal structure — by and large, we consider these factors a nuisance and try to engineer them away. In accordance with this, the majority of current successes of machine learning boil down to large scale pattern recognition on suitably collected ***independent and identically distributed*** (*IID*) data.

Statistical machine learning views data as a sequence of random variables. There is a process that generates that data, we'll call it the data generating process (DGP).

Let's talk about **assumptions**.

An animal either is a cat or it isn't. Socrates either was mortal or he wasn't.

A DPG either produces IID data or it doesn't.

The problem is we rarely know the structure and mechanics of the DGP. And

© 2025 Robert Osazuwa Ness · Privacy · Terms · Collection notice
Substack is the home for great culture

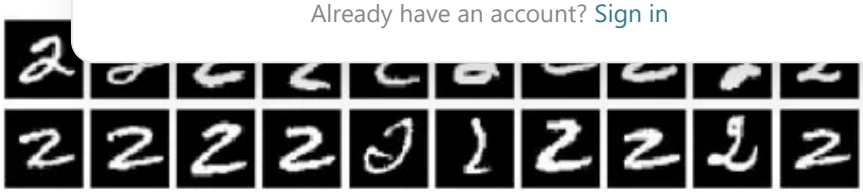
IID is nice because ata, but

Discover more from Altdeep.ai Newsletter
A semi-weekly newsletter to hone your skills in empirical modeling and statistical thinking.
Over 1,000 subscribers
Enter your email...
Subscribe
By subscribing, I agree to Substack's Terms of Use, and acknowledge its Information Collection Notice and Privacy Policy.
Already have an account? Sign in

they are all generat stead of
1000. If each point ave to
model that would-b

Statisticians are int ngs like
sampling theory, pa -values
flow directly from t

But in machine lear ted. For
example, MNIST is s. Each
digit image is paire



In this case, if MNIST were IID, it would mean every image and image label pair is independent of the others, and they all have the same distribution.

But is that true? In fact, the MNIST data was created from an earlier "NIST" dataset where the training data was Census Bureau employees, and the test data was high school students.

Yann LeCun (VP and Chief AI Scientist at Facebook) writes:

> The MNIST database was constructed from NIST's Special Database 3 and

> Special Database 1 which contain binary images of handwritten digits. NIST originally designated SD-3 as their training set and SD-1 as their test set. However, SD-3 is much cleaner and easier to recognize than SD-1. The reason for this can be found on the fact that SD-3 was collected among Census Bureau employees, while SD-1 was collected among high-school students. Drawing sensible conclusions from learning experiments **requires that the result be independent of the choice of training set and test among the complete set of samples**. Therefore it was necessary to **build a new database by mixing NIST's datasets** [emphasis added].

The NIST data was by nature not IID:

- **Not identical**. You wouldn't expect adult professional bureaucrats to write the same way as high school children.
- **Not independent**. If you were trying to predict, based on NIST data, how a person would write the digit "2", would it help to know if the person were a high schooler or a bureaucrat? Yes. Therefore, there is a statistical dependence between the handwritten digits of high-schoolers (and between bureaucrats).

LeCun and his team tried to engineer away the weakness of the IID assumption by mixing up the datasets. The remixing makes the differences between high-schoolers and bureaucrats seem like natural variation in the way people write.

So why would this be a problem?

The goal of MNIST is to train algorithms that can recognize digits. But the ability to recognize digits is only useful if that ability generalizes beyond the data the algorithm is trained on — a digit recognizer that only works on high school students and bureaucrats wouldn't be that interesting.

But if the data can't generalize from bureaucrats to high-schoolers, why should it be able to generalize from bureaucrats *and* high-schoolers to college professors, or the elderly, or to people who were drilled in writing Chinese

characters (Chinese elementary education places heavy emphasis on developing calligraphic skills)?

A common belief amongst the machine learning community is that more data solves the problem of generalization. Sometimes. That is only true if the data varies in a way that matches the natural diversity of the problem space — such that getting more data means that eventually, you get examples from even the rarest cases.

Unfortunately, the processes that collect the data often exclude vast swaths of the problem space. If the MNIST dataset had 70 million images instead of 70 thousand, it still would only be high-schoolers and bureaucrats.

Whether or not this is a huge deal depends on the stakes. Different problem spaces have different stakes. A document scanner erroneously recognizing an oddly styled character is probably not as bad as a self-driving car erroneously recognizing a pedestrian just because she's wearing an odd outfit.

If you liked this article, please share 🙏.

Discussion about this post

Comments Restacks

Write a comment...