In Machine Learning, What is Better: More Data or better Algorithms

Previous post
Next post

Tags: Big Data Hype, Data Quality, IMDb, Machine Learning, Quora, Xavier Amatriain

Gross over-generalization of “more data gives better results” is misguiding. Here we explain, in which scenario more data or more features are helpful and which are not. Also, how the choice of the algorithm affects the end result.

comments

By Xavier Amatriain (VP of Engineering at Quora).

"In machine learning, is more data always better than better algorithms?" No. There are times when more data helps, there are times when it doesn't.

Probably one of the most famous quotes defending the power of data is that of Google's Research Director Peter Norvig claiming that "We don't have better algorithms. We just have more data.". This quote is usually linked to the article on "The Unreasonable Effectiveness of Data", co-authored by Norvig himself (you should probably be able to find the pdf on the web although the original is behind the IEEE paywall). The last nail on the coffin of better models is when Norvig is misquoted as saying that "All models are wrong, and you don't need them anyway" (read here for the author's own clarifications on how he was misquoted).

The effect that Norvig et. al were referring to in their article, had already been captured years before in the famous paper by Microsoft Researchers Banko and Brill [2001]" Scaling to Very Very Large Corpora for Natural Language Disambiguation". In that paper, the authors included the plot below.

That figure shows that, for the given problem, very different algorithms perform virtually the same. however, adding more examples (words) to the training set monotonically increases the accuracy of the model.

So, case closed, you might think. Well… not so fast. The reality is that both Norvig's assertions and Banko and Brill's paper are right… in a context. But, they are now and again misquoted in contexts that are completely different than the original ones. But, in order to understand why, we need to get slightly technical. (I don't plan on giving a full machine learning tutorial in this post. If you don't understand what I explain below, read my answer to How do I learn machine learning?)

Variance or Bias?

The basic idea is that there are two possible (and almost opposite) reasons a model might not perform well.

In the first case, we might have a model that is too complicated for the amount of data we have. This situation, known as *high variance*, leads to model overfitting. We know that we are facing a high variance issue when the training error is much lower than the test error. High variance problems can be addressed by reducing the number of features, and… yes, by increasing the number of data points. So, what kind of models were Banko & Brill's, and Norvig dealing with? Yes, you got it right: high variance. In both cases, the authors were working on language models in which roughly every word in the vocabulary makes a feature. These are models with many features as compared to the training examples. Therefore, they are likely to overfit. And, yes, in this case adding more examples will help.

But, in the opposite case, we might have a model that is too simple to explain the data we have. In that case, known as *high bias*, adding more data will not help. See below a plot of a real production system at Netflix and its performance as we add more training examples.

So, no, **more data does not always help**. As we have just seen there can be many cases in which adding more examples to our training set will not improve the model performance.

More features to the rescue

If you are with me so far, and you have done your homework in understanding high variance and high bias problems, you might be thinking that I have deliberately left something out of the discussion. Yes, high bias models will not benefit from more training examples, but they might very well benefit from more features. So, in the end, it is all about adding "more" data, right? Well, again, it depends.

Let's take the Netflix Prize, for example. Pretty early on in the game, there was a blog post by serial entrepreneur and Stanford professor Anand Rajaraman commenting on the use of extra features to solve the problem. The post explains how a team of students got an improvement on the prediction accuracy by adding content features from IMDb.

In retrospect, it is easy to criticize the post for making a gross over-generalization from a single data point. Even more, the follow-up post references SVD as one of the "complex" algorithms not worth trying because it limits the ability of scaling up to larger number of features. Clearly, Anand's students did not win the Netflix Prize, and they probably now realize that SVD did have a major role in the winning entry.

As a matter of fact, many teams showed later that adding content features from IMDB or the like to an optimized algorithm had little to no improvement. Some of the members of the Gravity team, one of the top contenders for the Prize, published a detailed paper in which they showed how those content-based features would add no improvement to the highly optimized collaborative filtering matrix factorization approach. The paper was entitled Recommending New Movies: Even a Few Ratings Are More Valuable Than Metadata.

To be fair, the title of the paper is also an over-generalization. Content-based features (or different features in general) might be able to improve accuracy in many cases. But, you get my point again: **More data does not always help**.

Better Data != More Data (Added this section *in response to a comment)*

It is important to point out that, in my opinion, better data is always better. There is no arguing against that. So any effort you can direct towards "improving" your data is always well invested. The issue is that better data does not mean **more** data. As a matter of fact, sometimes it might mean **less**!

Think of data cleansing or outlier removal as one trivial illustration of my point. But, there are many other examples that are more subtle. For example, I have seen people invest a lot of effort in implementing distributed Matrix Factorization when the truth is that they could have probably gotten by with sampling their data and gotten to very similar results. In fact, doing some form of smart sampling on your population the right way (e.g. using stratified sampling) can get you to better results than if you used the whole unfiltered data set.

The End of the Scientific Method?

Of course, whenever there is a heated debate about a possible paradigm change, there are people like Malcolm Gladwell or Chris Anderson that make a living out of heating it even more (don't get me wrong, I am a fan of both, and have read most of their books). In this case, Anderson picked on some of Norvig's comments, and misquoted them in an article entitled: The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.

The article explains several examples of how the abundance of data helps people and companies take decision without even having to understand the meaning of the data itself. As Norvig himself points out in his rebuttal, Anderson has a few points right, but goes above and beyond to try to make them. And the result is a set of false statements, starting from the title: the data deluge does not make the scientific method obsolete. I would argue it is rather the other way around.

Data Without a Sound Approach = Noise

So, am I trying to make the point that the Big Data revolution is only hype? No way. Having more data, both in terms of more examples or more features, is a blessing. The availability of data enables more and better insights and applications. More data indeed enables better approaches. More than that, it **requires** better approaches.

In summary, we should dismiss simplistic voices that proclaim the uselessness of theory or models, or the triumph of data over these. As much as data is needed, so are good models and theory that explains them. But, overall, what we need is good approaches that help us understand how to interpret data, models, and the limitations of both in order to produce the best possible output.

In other words, data is important. But, data without a sound approach becomes noise.

Original. Reposted with permission.

Bio: Xavier Amatriain, is a VP of Engineering at Quora, well known for his work on Recommender Systems and Machine Learning. He build teams and algorithms to solve hard problems with business impact. Previously, he was Research/Engineering Director at Netflix.

Related

- Debunking Big Data Myths. Again
- Interview: Josh Hemann, Activision on Why the Tolerance for Ambiguity is Vital
- Will Deep Learning take over Machine Learning, make other algorithms obsolete?

The Modeling Agency, LLC • 3 years ago

Great explanation to the original false dichotomy of data vs. method. I agree that it's strategic problem design from adequate assessment and project definition that will overcome deficiencies in either methods or data!

In other words, I'll beat the "data scientist's" fast race car (optimized with superior data and algorithms)... just by understanding the rules of the race. I'll be at the finish line with 'good enough' data and algorithms (can optimize on later pass) while the typical data scientist drives way off course... at high speed. :^}

1 • Reply • Share ›

Seremonia • a year ago

Having enthusiast on research to provide better and original algorithm. One of them is able to reduce dataset more compact so that we can have dataset from big number of rows into very small number of rows compared to number of rows from query.

Suppose you have dataset with 10 thousand rows for specified query with 5 thousand rows, then using dataset optimizer, we can reduce dataset into 5 thousand rows or even (very) less. This i may call sufficient dataset.

If you want to try it, you can send your dataset at least 10 thousand rows and your query (must be relevant to dataset) . Provide your download link or through this link https://www.kaggle.com/data... (there are limitation of how many request may be picked)

• Reply • Share ›

Zack Chase Lipton Mod • 3 years ago

HI Xavier,

This topic (the relative significance of data and algorithms) was addressed in my March article on KDnuggets http://www.kdnuggets.com/20....

My piece was prompted in response to this paper http://arxiv.org/pdf/1503.0... ("Do we need more training data?") as well as the Norvig piece.

Perhaps if you find this topic interesting, check them out.

• Reply • Share ›

Previous post
Next post

Top Stories Past 30 Days

Most Popular

1. **The Most in Demand Skills for Data Scientists**
2. **What is the Best Python IDE for Data Science?**

Most Shared

1. **The Most in Demand Skills for Data Scientists**
2. **What is the Best Python IDE for Data Science?**

3. **To get hired as a data scientist, dont follow the herd**
4. **9 Must-have skills you need to become a Data Scientist, updated**
5. **10 Free Must-See Courses for Machine Learning and Data Science**
6. **What does a data scientist REALLY look like?**
7. **10 Best Mobile Apps for Data Scientist / Data Analysts**

3. **To get hired as a data scientist, dont follow the herd**
4. **Introduction to Deep Learning with Keras**
5. **How Machines Understand Our Language: An Introduction to Natural Language Processing**
6. **Intro to Data Science for Managers**
7. **Generative Adversarial Networks – Paper Reading Road Map**

Latest News

- Deep Learning Cheat Sheets
- KDnuggets 18:n45, Nov 28: Your Favorite Python IDE/editor? ...
- SQL, Python, and R in One Platform
- What Python editors or IDEs you used the most in 2018?
- Humana: Principal Data Scientist/Informatics Principal ...
- Making Machine Learning Accessible [Webinar Replay]

Top Stories
Last Week

Most Popular

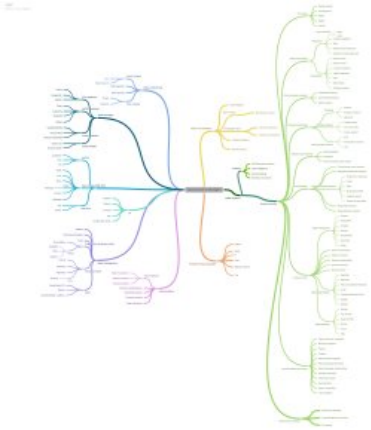
1. NEW **What is the Best Python IDE for Data Science?**

2. NEW **10 Free Must-See Courses for Machine Learning and Data Science**
3. **9 Must-have skills you need to become a Data Scientist, updated**
4. NEW **The Most in Demand Skills for Data Scientists**
5. NEW **To get hired as a data scientist, dont follow the herd**

6. NEW **Intro to Data Science for Managers**
7. NEW **The Big Data Game Board**

Most Shared

1. **Intro to Data Science for Managers**



2. **An Introduction to AI**
3. **The Big Data Game Board**
4. **6 Goals Every Wannabe Data Scientist Should Make for 2019**
5. **Top KDnuggets tweets, Nov 14-20: 10 Free Must-See Courses for Machine Learning and Data Science; Great list of #MachineLearning Resources**
6. **Using a Keras Long Short-Term Memory (LSTM) Model to Predict Stock Prices**
7. **How Important is that Machine Learning Model be Understandable? We analyze poll results**

More Recent Stories

- Making Machine Learning Accessible [Webinar Replay]
- Drexel University: 2 Teaching Faculty Positions in Data Scienc...
- How to Engineer Your Way Out of Slow Models
- Bringing Machine Learning Research to Product Commercialization
- Data Pro Cyber Monday – Choose Your Savings
- 3 Challenges for Companies Tackling Data Science
- My secret sauce to be in top 2% of a Kaggle competition
- Global Legal Entity Identifier Foundation (GLEIF): Data Analys...
- Top Stories, Nov 19-25: What is the Best Python IDE for Data S...
- Data Science Strategy Safari: Aligning Data Science Strategy t...
- Top 5 domains Big Data analytics helps to transform
- Intro to Data Science for Managers
- Monash University: Lecturer/Sr Lecturer – Digital Health...
- Monash University: Research Fellow (Digital Civics) [Melbourne...
- 6 Goals Every Wannabe Data Scientist Should Make for 2019
- Cartoon: Thanksgiving, Big Data, and Turkey Data Science.
- Top tweets, Nov 14-20: 10 Free Must-See Courses for Machine...
- Join the World's Biggest Deep Learning Summit – KDnugg...
- An Introduction to AI
- WPI: Post-Doctoral Fellow [Worcester, MA]

X