A Benchmark For Benchmarks*

Gaurav Sood

December, 2023

Benchmark datasets like MNIST, ImageNet, etc., abound in machine learning. Such datasets stimulate work on a problem by providing an agreed-upon mark to beat. Many of the benchmark datasets, however, are constructed in an ad hoc manner.[1] As a result, it is hard to understand why the best-performing models vary across different benchmark datasets (see here), to compare models, and to confidently prognosticate about performance on a new dataset. To address such issues, in the following paragraphs, we provide a framework for building a good benchmark dataset.

To build a good benchmark dataset, start by defining the problem precisely. Let me illustrate the point with an example. Let's say that we want to build a benchmark dataset for testing the performance of a cat-dog photo classifier. Famously, if in the benchmark dataset, all the photos of a dog are clicked outside the house and all the photos of a cat are clicked inside the house, the classifier that may perform the best is an indoor-outdoor classifier. And that may be a reasonable outcome if we expect to do well on such images. Or it may not be. If we want to measure the capacity to identify cats and dogs, we may need to build datasets that test the model's capability to do that without using irrelevant information (see Cronbach and Meehl 1955).

To highlight a different aspect of imprecision in problem definition, consider a different example. Say that we want to test the performance of a model that infers race from the last name of a person. We could build a benchmark dataset by combining various datasets with no particular criteria except the availability of data (see, e.g., Krstovski, Lu and Xu 2023). However, performance on such a benchmark would be hard to interpret because performance on such a dataset doesn't directly map to a well-defined problem in the world. To specify the problem we need to specify the population. If the population is adult Americans, and if say "you picked a person at random [whose] last name [was] 'Smith' in the US in 2010 and [were] asked ... to guess this person's race (as measured by the census), the best guess would be based on what is available from the aggregated Census [last name] file. It is the Bayes Optimal Solution." (Sood and Laohaprapanon, 2017)

*I am looking for feedback. Please email me at gsood07@gmail.com

[1]For instance, a survey of the text classification benchmarks suggests no organizing principle except that the tasks are text classification.

Once you have defined the problem, the attributes of good benchmark datasets are:

- **Authorized Content.** The data ought not to violate privacy and copyright.
- **High-quality Labels.**
  - **Reliability.** If you can't replicate a label, it is likely no good. One way to assess noise in labels is to measure agreement across labelers. If the agreement is low, it may be because the directions are bad or because the task is hard. If the task is hard, you could take pool judgments across multiple people and use the pooled judgment as the final label (still check if means across multiple people are reliable or not).
  - **Validity.** Even if a label can be consistently produced, it needn't map to the underlying concept. For instance, a broken clock is consistent but not a valid indicator of time. The bar is that the labels precisely map to the underlying construct. Here are some reasons why labels may not be valid:
    - **Confounders.** Labelers may use factors irrelevant to label. The result could be incorrect (but consistent) labels.
    - **Bias.** If coders are biased, they may produce problematic labels. For instance, see The Art Newspaper 2019 about issues with ImageNet labels.
- **Documentation.** Minus the documentation, users are left to their imagination or to their nails to dig out details about the population, how the labels were coded, etc. For instance, is news about a sale Business news—"A bargain hunter's paradise Massachusetts bargain hunters showed up in droves and shopped hard on yesterday's sales tax holiday..." (Zhang, Zhao and LeCun, 2015)? It could be. Or it could be an error. It depends on how the category was defined and the direction to labelers. All in all, how labels were created is important for comprehensibility and reproducibility.
- **Size.** The benchmark datasets need to be large enough so that the sampling variation is small enough that we can make credible judgments across algorithms.
- **Random Sample.** We would ideally like to randomly sample from the domain of the problem to build a dataset. Absent that, we need to think carefully about the sub-varieties of a problem to drive clarity about the perimeter of what we learn. For instance for text classification, some relevant dimensions may be:

  - **Topic.** E.g., news categories, sentiment, diseases from doctors' notes, movie categories, website classification, hate speech or not, spam or not, etc.
  - **Length of text.** Sentence, tweet, paragraph, paper, book, etc.
  - **Dependent Variable.** Multi-class and multi-label.
  - **Source of Text.** STS—closed-caption, ML transcription, manual transcript, LLM-assisted output, or typed.
  - **Dialect.** There are well-known differences in spelling but also words that exist in one dialect, e.g., Brinjal, that don't exist in another, and different words for the same object, e.g., Brinjal vs. Eggplant. Some dialects may also use words from the native language. Different dialects may also have different phrases to refer to the same thing, e.g., Do the needful (do what is required, or take care of it), Revert (for reply), etc. All of these issues can cause models to have different performance in different dialects.
  - **Age.** Age is its own zip code. Text from different eras may use language that is very different from today's.
  - **Out of Distribution.** Can a classifier accurately identify out-of-distribution data and pro-rate its predictions appropriately?

- **Diversity.** To measure performance on fairness-related issues, data should have an adequate representation of various races, genders, and such.
- **No Duplicates.** Having duplicates provides a misleading understanding of the performance. Doing well on one example now gets you credit for more.

Future

- Assemble a large set of text classification datasets and tally how often the best-performing model on one dataset is also the best-performing on another. Measure the standard deviation of the best-performing model and the variability of model performance across datasets.
- **Selective Benchmarking.** Do papers use benchmarks they perform best on?

References

Cronbach, Lee J and Paul E Meehl. 1955. "Construct validity in psychological tests." *Psychological bulletin* 52(4):281.

Krstovski, Kriste, Yao Lu and Ye Xu. 2023. "Inferring Gender From Name: A Large Scale Performance Evaluation Study.". Available at SSRN: https://ssrn.com/abstract=4572589 or http://dx.doi.org/10.2139/ssrn.4572589.

Sood, Gaurav and Suriyan Laohaprapanon. 2017. "ethnicolr." https://github.com/appeler/ethnicolr.

The Art Newspaper. 2019. "Leading online database to remove 600,000 images after art project reveals its racist bias." *The Art Newspaper* .
URL: *https://www.theartnewspaper.com/2019/09/23/leading-online-database-to-remove-600000-images-after-art-project-reveals-its-racist-bia*

Zhang, Xiang, Junbo Jake Zhao and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *NIPS*.