Setting Up Your Public Data for Success

Rachael Tatman
Kaggle
rachael@kaggle.com

1. Introduction

If you have gone to all the trouble of preparing your data and releasing it to the public, you probably want people to actually use it! In this talk, I draw on lessons we have learned on Kaggle‘s public data platform to give you concrete recommendations for making your data more accessible, whether youre still collecting data, preparing it for release or have already released it.

While it does take time to prepare data for sharing and to follow the steps suggested here, it pays off in impact and citations; research papers that make their data open see a 9% increase in number of citations [3].

1.1 During data collection

Consider the potential audience for your data. The major consumers of public data – researchers, journalists, learners, educators and hobbyists – all have different needs. By considering who would be most interested in your data once it‘s released you can guide your data collection and formatting to better suit that community.

Collect data that is as rich as possible. Even if you aren’t planning on analyzing some factor, if you can collect information on it ethically and easily then considering including it in your data. It will let others ask more detailed questions and expand the potential audience for your data.

1.2. Before releasing the data

Prepare clear documentation and rich metadata. Describing what is in your data is a fairly simple step for you but extremely difficult for someone coming to your data for the first time. Make sure to include information on how and why the data was collected, the contents of each file and any information necessary during analysis, like the contents and units of columns in tabular data files. (See Gebru et al for one recommendation for formatting metadata [1].)

Provide sample code. Particularly if your data requires a particular type of visualization or preprocesing, providing sample code makes it more likely people will play around with your data. For example, if you’re sharing an acoustic dataset that is best visualized using a wideband spectrogram, then providing sample code for producing a wideband spectrogram would be a good a idea.

Consider releasing a version of your data that can be used as a drop-in replacement for another popular dataset. This may not be practical for all datasets, but datasets that can be used in place of another popular dataset can see wide adoption. A good example is the Fashion-MNIST dataset [4], which is made up of small labelled images of garments and can be used to replace the MNIST dataset [2], which contains small labelled images of hand written digits.

Pose questions or problems that could be answered using your data. Especially for learners, coming up with interesting questions in a new domain can be hard. By posing example questions, you give them an easy “in to start working with the data.

1.3. After you have released your data

Reach out to relevant communities. If you think theres a community that would be particularly interested in your data, reach out to them to let them know its available. Post about it on Twitter, e-mail your colleagues, write a blog post, whatever you’re comfortable doing to get the word out!

Consider hosting your data on multiple platforms. I’m a strong advocate in redundancy in data sharing (I’ve personally shared academic data on a service that was later shut down). In addition to serving as a fail-safe, hosting data on multiple platforms exposes it to users on those platforms, which can help improve adoption.

Update your dataset periodically. If you collect more data, make sure to update your dataset to keep it fresh. This is especially important if your data has any sort of time series component; users will quickly notice that a dataset is out of date and may choose not to use it for that reason.

References

[1] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumeé III, and K. Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

[2] Y. LeCun, C. Cortes, and C. Burges. Mnist handwritten digit database. *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2, 2010.

[3] H. A. Piwowar and T. J. Vision. Data reuse and the open data citation advantage. *PeerJ*, 1:e175, 2013.

[4] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.