An Introduction to Introductions

Praveen Nair

University of California San Diego

October 24, 2022

Quarter 1 Checkpoint

Your checkpoint for your Quarter 1 project is due on **Sunday, October 30th.**

This includes:

- Title & Abstract
- Introduction to your report
  - Introduction paragraph
  - Literature review
  - Description of data
- Any code you've produced so far

Title

A title should let your audience know what the paper will be about, and ideally, what contributions you've made in your report.

Password policies of most top websites fail to follow best practices

Kevin Lee   Sten Sjöberg   Arvind Narayanan

Department of Computer Science and Center for Information Technology Policy
Princeton University

Robust Meta-learning with Sampling Noise and Label Noise via Eigen-Reptile

Dong Chen, Lingfei Wu, Siliang Tang, Xiao Yun, Bo Long, Yueting Zhuang *Proceedings of the 39th International Conference on Machine Learning*, PMLR 162:3662-3678, 2022.

Titles Are Really Important!

Focus on a single message; *papers that simultaneously focus on multiple contributions tend to be less convincing about each and are therefore less memorable.*

The most important element of a paper is the title—think of the ratio of the number of titles you read to the number of papers you read. *The title is typically the first element a reader encounters, so its quality determines whether the reader will invest time in reading the abstract.*

The title not only transmits the paper's central contribution but can also serve as a constant reminder (to you) to focus the text on transmitting that idea.[1]

[1] https://doi.org/10.1371/journal.pcbi.1005619

Abstracts

In about a paragraph, you should describe the problem you intend to solve, how you approached it, and what your most important contributions were. **Just like the title, the abstract going to decide whether someone is going to continue reading your report!**

> *The abstract is, for most readers, the only part of the paper that will be read. This means that the abstract must convey the entire message of the paper effectively.*

Structuring an Abstract

Context: Describe what the previous research in the field is, and what results you're building on.

Recent work introduced the model of learning from discriminative feature feedback, in which a human annotator not only provides labels of instances, but also identifies discriminative features that highlight important differences between pairs of instances. It was shown that such feedback can be conducive to learning, and makes it possible to efficiently learn some concept classes that would otherwise be intractable. *However, these results all relied upon perfect annotator feedback. In this paper, we introduce a more...*[2]

[2] https://arxiv.org/abs/2003.03946

Context

Context: Identify a gap in the existing literature that motivates the work you've done.

...and makes it possible to efficiently learn some concept classes that would otherwise be intractable. **However, these results all relied upon perfect annotator feedback.** *In this paper, we introduce a more realistic, robust version of the framework, in which the annotator is allowed to make mistakes. We show how such errors can be handled algorithmically, in both an adversarial and a stochastic setting...*

Content

Content: Describe the main contribution of your paper, whether this is a discovery about your dataset, or a new methodology.

. . . However, these results all relied upon perfect annotator feedback. **In this paper, we introduce a more realistic, robust version of the framework, in which the annotator is allowed to make mistakes. We show how such errors can be handled algorithmically, in both an adversarial and a stochastic setting.** *In particular, we derive regret bounds in both settings that, as in the case of a perfect annotator, are independent of the number of features. We show that this result cannot be obtained by a naive reduction from the robust setting to the non-robust setting.*

Content

Content: Describe what makes this insight important to readers, or what attributes make your methology useful.

...In this paper, we introduce a more realistic, robust version of the framework, in which the annotator is allowed to make mistakes. We show how such errors can be handled algorithmically, in both an adversarial and a stochastic setting. **In particular, we derive regret bounds in both settings that, as in the case of a perfect annotator, are independent of the number of features. We show that this result cannot be obtained by a naive reduction from the robust setting to the non-robust setting.**

Introductions

Like an abstract, an introduction should situate your paper in the research space: a broad overview of why the field is important, what work you plan to do in it, and what data you're going to use to do that.

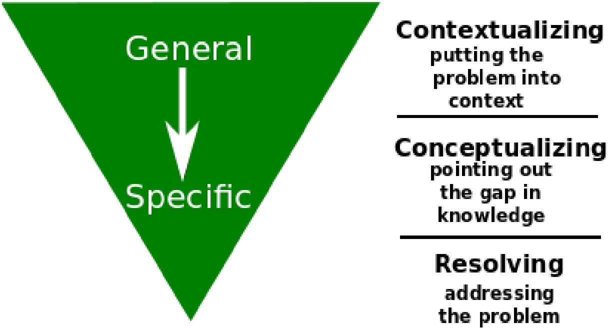

Fig. 1. The inverted pyramid approach (© P. Regoniel)

Example: Obermeyer (2019)

There is growing concern that algorithms may reproduce racial and gender disparities via the people building them or through the data used to train them (1–3). Empirical work is increasingly lending support to these concerns. For example, job search ads for highly paid positions are less likely to be presented to women (4), searches for distinctively Black-sounding names are more likely to trigger ads for arrest records (5), and image searches for professions such as CEO produce fewer images of women (6). Facial recognition systems increasingly used in law enforcement perform worse on recognizing faces of women and Black individuals (7, 8), and natural language processing algorithms encode language in gendered ways (9).[3]

[3]https://www.science.org/doi/full/10.1126/science.aax2342

Empirical investigations of algorithmic bias, though, have been hindered by a key constraint: Algorithms deployed on large scales are typically proprietary, making it difficult for independent researchers to dissect them. Instead, researchers must work "from the outside," often with great ingenuity, and resort to clever work-arounds such as audit studies. Such efforts can document disparities, but understanding how and why they arise—much less figuring out what to do about them—is difficult without greater access to the algorithms themselves. Our understanding of a mechanism therefore typically relies on theory or exercises with researcher-created algorithms (10–13). Without an algorithm's training data, objective function, and prediction methodology, we can only guess as to the actual mechanisms for the important algorithmic disparities that arise.

In this study, we exploit a rich dataset that provides insight into a live, scaled algorithm deployed nationwide today. It is one of the largest and most typical examples of a class of commercial risk-prediction tools that, by industry estimates, are applied to roughly 200 million people in the United States each year. Large health systems and payers rely on this algorithm to target patients for "high-risk care management" programs. These programs seek to improve the care of patients with complex health needs by providing additional resources, including greater attention from trained providers, to help ensure that care is well coordinated. Most health systems use these programs as the cornerstone of population health management efforts...

Literature Review

You don't need to have read every paper in the field, but you present the basic foundation of what has been done before.

If you're applying a model to data:

- What work has been done before on similar data?
- What tasks have similar models been used on before?

If your paper introduces a new method:

- What are previous frameworks you're building on?
- What do previous methods do wrong that you want to fix?

Data Description

Describing your data in detail isn't just important for readers of your paper; it's also vital for future researchers (like yourselves) who want to replicate it!

Things to consider:

- Source of the data (if it's public, then include a citation!)
- Steps you took to clean, process, or alter data
- Basic information about data: number of rows, what attributes are included, breakdown by some important variable
- Why the data is well-suited for the problem you're solving
- Limitations of the dataset for your analysis (e.g. selection bias, unrepresentativeness, missing values, etc.)

Example: Obermeyer (2019)

Working with a large academic hospital, we identified all primary care patients enrolled in risk-based contracts from 2013 to 2015. . . . Any patient who identified as Black was considered to be Black for the purpose of this analysis. Of the remaining patients, those who self-identified as races other than White (e.g., Hispanic) were so considered (data on these patients are presented in table S1 and fig. S1 in the supplementary materials). We considered all remaining patients to be White. **This approach allowed us to study one particular racial difference of social and historical interest between patients who self-identified as Black and patients who self-identified as White without another race or ethnicity; it has the disadvantage of not allowing for the study of intersectional racial and ethnic identities.**

Example: Obermeyer (2019)

Our main sample thus consisted of (i) 6079 patients who self-identified as Black and (ii) 43,539 patients who self-identified as White without another race or ethnicity, whom we observed over 11,929 and 88,080 patient-years, respectively (1 patient-year represents data collected for an individual patient in a calendar year). The sample was 71.2% enrolled in commercial insurance and 28.8% in Medicare; on average, 50.9 years old; and 63% female (Table 1).

Remember, visualization and tables are your friends for describing data!

Quarter 1 Checkpoint

Your checkpoint for your Quarter 1 project is due on **Sunday, October 30th.**

This includes:

- Title & Abstract
- Introduction to your report
  - Introduction paragraph
  - Literature review
  - Description of data
- Any code you've produced so far