Machine Learning Training to Support Diversity of Opinion

Johanne Christensen, Benjamin Watson
North Carolina State University
jtchrist,bwatson@ncsu.edu

Abstract

Drawing from a small qualitative study of users during a training task for a machine learning system, we explore the implications of restricting the collection of training data to categorical labels alone in domains where subjectivity may be a necessity for serving the needs of a diverse user base. We advocate for new methods of producing labeled training data for machine learning that can discover and support diverse opinions. Such methods might include collecting additional data, using proxy measures of data accuracy, and applying survey methodology.

Introduction

The common practice in data collection of converging on a simple “ground truth” has implications for the functionality of systems intended to interact with a diverse set of users. While some low level tasks may have datasets with objectively correct answers, many tasks have an inherent amount of subjectivity. This is especially true in domains where the data is complex and high-dimensional or unstructured.

When a task can be framed as a question with an unconditional answer, as in ”does this photo have a cat in it?”, we don’t consider that data domain to be subjective. But when the task’s question depends even in part on an opinion or other context not explicitly captured in the observed data, the answer may vary dependent on these contextual factors. For example, in medicine, doctors often do not agree on recommended actions (Ross and Swetlitz). Appropriate recommendations may vary depending on treatment guidelines in specific locales, availability of treatments, and/or patient preference. Making recommendations based on US standards of care can produce results that are not useful to patients in other countries. And ultimately, these systems do not support a broadly diverse group of users. Training a system to support doctors in making treatment recommendations to patients requires a dataset that adequately models the necessary flexibility to address each individual situation.

Choosing to converge on a simple ground truth can erase minority opinions that nevertheless have validity to a subset of users in a diverse user base. In order to build systems that support a diverse set of users, the systems must be trained to support a diverse set of opinions that represent and respect these users’ needs. Therefore, labeling training data in a way that supports this diversity is essential. In this position paper, we explore how the design of the training data collection process relative to diversity may impact utility and adoption.

A Think Aloud Study of Relevancy Labeling

From a small pilot study of six participants asked to label training data for a question answer system, we discovered that even this small set of people produced a variety of schemes for determining relevancy. We asked participants to think aloud as they rated the relevancy of answers to subjective questions, such as ”How can I avoid or minimize jet lag?”. Our study raised questions about how we can capture the nuance inherent in this type of dataset.

While labeling, participants expressed preferences for different answer types when choosing how to accept an answer. P2 said, ”I’m weeding out personal stories,” while P6 commented during rating answers for the same question, ”The personal experience ones seem to answer the question better”. Objectively, neither participant’s choice is more or less correct than the other’s; both schemes – factual answers vs personal experiences – could provide helpful information to different users, depending on the user’s needs. In particular, the subjectivity within this domain lends itself to to multiple and sometimes contradictory correct answers.

Supporting Diversity of Opinion

Where diverse opinions exist, they must be supported by a dataset that reflects that diversity. AI implementations currently struggle to provide the flexible solutions necessary to solve problems for different types of users, and not just a single approximation of all types. In this section, we make several claims advocating that better methods for collecting data can be utilized throughout the design, development, and deployment of these systems to meet these goals.

Claim 1: *to accurately reflect subjectivity and diversity, we must improve our data collection methods.*

Lack of support for diverse opinions such as the ones described in our think aloud study can have potentially detrimental effects for users. If a user is being presented with factual answers when anecdotal ones would be more helpful to them, they will not find the same value in the system than a different user whose needs are more closely aligned with the majority (or plurality) opinion. Furthermore, users from already marginalized groups can experience this type of interaction as invalidating or discriminatory. At a minimum, system design should address potential areas of discrimination against groups of users. Bringing these concerns to data labeling and modeling training is a way to proactively design systems to not only avoid discrimination but also to effectively serve the different needs of the entire user base.

Claim 2: *to accurately capture subjectivity and diversity in a dataset, we should apply survey methodology.*

One solution for accurate capture of subjectivity is to utilize survey sampling methods. If the demographics of a user base are known or can be reasonably estimated, proportional sampling of experts from these demographics may be a way to collect a diverse set of labeled data. In other instances, SMEs may have sufficient knowledge of possible user preferences to construct diversity of opinions themselves. The participants of our study were able to recognize and detail possible alternative opinions that they themselves did not favor. A doctor with sufficient clinical experience would likewise be able to determine a range of preferences that patients may exhibit and a system should be trained to accommodate.

Claim 3: *richer data collection and data clustering may alternatively capture diversity of opinion.*

In other domains, possible variances of opinion may not be obvious until SMEs begin labeling data. Processes for eliciting these differing labeling schemes should include tools to assist in this task. Rather than simply requiring SMEs to provide labels, we believe that asking them to elaborate why they chose the label they did will surface additional information that is useful, including in cases where multiple labels may be correct depending on circumstance. Additionally, finding areas where the support of diverse opinions is necessary may be easier once the SME is engaged with the data, which may itself reveal distinct clusters.

Claim 4: *to support modeling of human subjectivity and diversity, data collection should be enriched.*

In our study, participants engaged in a process of mental modeling while they each derived their own understanding of how to determine relevancy for each question. Equally apparent from the think aloud results and the researcher's observation of how participants approached the task is that participants are generally not explicitly aware of the cognitive process they are engaging to build these models. Yet the fact that participants are able to clearly articulate what kind of answers they prefer suggests that there is more information that can be collected from subject matter experts (SMEs) than just labels.

We hypothesize that users engaging in the training task are subconsciously building a rich set of data beyond the training labels themselves that could potentially be utilized. More specifically, that there is an underlying structure of the domain that adequately captures the diversity a system should be supporting. The process of labeling data can be utilized to address or discover that structure.

Claim 5: *when workers are scarce, measure attention to the task, and consistency in their task.*

In domains where subject matter experts are scarce, crowdsourcing large amounts of training data and averaging out genuine inaccuracies is not feasible to distinguish from consistent vs careless outliers in SME opinions. Therefore, collecting training data with some measure of label correctness is essential to ensure the quality of the training data used to build a model.

”Correctness” as a proxy for data quality can be considered along two vectors: *attention* and *consistency*. Attention can be measured as task load (via a scale like NASA-TLX (Hart 2006)) or the SME's ability to elaborate on their own thoughts. Consistency means that SMEs have some rubric that is being applied evenly across multiple sessions of rating. As it is difficult for users to maintain consistency in a mental model when a returning to an interrupted task (LaToza, Venolia, and DeLine 2006), having tooling (Christensen et al. 2018) to serve as an external memory aid for SMEs should improve their consistency and reduce error and improve the overall data quality. In addition, if SMEs articulate their rubric, data managers can bring apparent inconsistencies to the attention of SMEs for possible correction.

Claim 6: *incorporate end user feedback to identify gaps in model subjectivity and diversity.*

Corrective feedback is often discussed in terms of allowing end users to provide corrections to a model which is making incorrect predictions. However, in subjective domains, it may not always be clear when the feedback is correcting a genuine error or suggesting an alternative result. More research is needed to answer questions on how to implement and understand user feedback in these domains.

Conclusions

In domains where user needs are a factor, training to support diverse opinions may increase the utility of the system across a more diverse set of users. Instead of converging on a single popular answer to train a model, we can better serve our user base if we consider expanding the process of training a machine learning model to inherently support the flexibility needed to assist on complex, nuanced, and context dependent problems.

References

Christensen, J.; Watson, B.; Rindos, A.; and Joines, S. 2018. Building bridges: A case study in structuring human-ml training interactions via ux.

Hart, S. G. 2006. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, 904–908. Sage publications Sage CA: Los Angeles, CA.

LaToza, T. D.; Venolia, G.; and DeLine, R. 2006. Maintaining mental models: a study of developer work habits. In *Proceedings of the 28th international conference on Software engineering*, 492–501. ACM.

Ross, C., and Swetlitz, I. Ibm pitched its watson supercomputer as a revolution in cancer care. it's nowhere close. https://www.statnews.com/2017/09/05/watson-ibm-cancer/. Accessed: 2019-11-01.