6.864 HW2: Dataset Creation

Background

Data drives the process of machine learning, determining what a model learns and how well it will perform in the real world. When trained on poor-quality or unsuitable data, even the best-trained model can have unintended consequences.

Collecting and annotating a high-quality dataset is a nontrivial process, but one that often gets skimmed over. There are many steps involved: from deciding what data to even collect (and from where), to how to measure features and assign labels. Most of these steps involve people weighing different tradeoffs and making decisions. These choices can result in significantly different datasets and models. Even when you're just using an existing dataset, it's important to remember that data isn't pre-existing, objective ground truth. Rather, it reflects the judgments and values of the people who collected, measured, and annotated it.

This doesn't mean that data isn't useful -- it just means that when given a dataset, you should be able to critically question how and by whom it was created, and what its limitations might be. This will also give you insight about what the data should and should not be used for.

Annotating labels

In this assignment, we will explore one aspect of dataset creation: annotating data with labels. In supervised learning, labels determine what the model is actually optimizing.

The first decision involved in annotating labels is figuring out what the label should be. For example, let's say you want to build a system to predict the sentiment of a movie review. You could make your labeling scheme binary: either 'positive' or 'negative'. Perhaps you also want to include 'neutral.' Or, maybe you think a linear scale is less important, and want to label specific emotions the reviewer expresses like 'confusion' or 'boredom.'

An important concern at this stage is **construct validity**: does the label capture the concept that it's supposed to? Ultimately a label is **an operationalization** (a concrete *measurement*) to approximate some **construct** (aka an *idea* or *concept*) that's not directly encoded or observable. In the movie review example, human annotations of 'positive' and 'negative' are intended to approximate the higher-level construct of 'sentiment'. Perhaps you're convinced that these labels capture the concept of 'sentiment' effectively. But for other domains, it may not be so straightforward. For example, when someone applies for a loan or a credit card, banks typically want to know their 'creditworthiness,' or how much they trust the person to repay the loan. This more nebulous construct might be measured through a 'credit score,' which is a numerical score generated by a model that's supposed to represent 'creditworthiness.' Does that score actually capture 'creditworthiness'? Or does it miss important factors and/or include irrelevant ones?

After defining the concept you're interested in and the specific label you want to collect, you need to decide how to actually measure and assign these labels to specific examples. In the movie review example, you could use star ratings (if available) to map to the 'positive' and 'negative' labels. Or maybe you try to mine labels in some other way, like detecting specific emojis in the review and mapping them to emotion labels. More often, it's hard to find labels that naturally occur in the data; instead, human annotators manually label examples. Who these annotaters are may vary depending on things like the scale of the data, convenience, or domain speciality. In some cases, for small-scale data, it might just be the researchers collecting the data who then annotate it. There are also several online platforms where crowdsourced users are given instructions and paid to annotate examples. Or for more specialized domains, you might need to recruit people with more in-depth knowledge (e.g., doctors to label medical data).

Because labels guide the entire optimization process, it's important to think about who is doing the labeling, how many people are labeling each example, how labeler disagreements are dealt with, and how subjectivity and annotator differences affect the end result.

There are other sub-aspects of construct validity to think about here: how do the labels you obtain compare with other, existing ways of measuring that label (**convergent validity**)? Are they able to predict the things you think they should (**predictive validity**)? We can imagine developing different ways to measure these things that can complement other qualitative assessments.

While construct validity asks whether the label is a good operationalization of the target construct, **construct reliability** is a complementary concern focused on the consistency of measuring the label. For example, how consistent are the labels from different annotators for a given example (**interrater reliability**)? How consistent is a single annotator at different points in time (**test-retest reliability**)?

For more on different types of reliability & validity, and their implications for machine learning, check out this paper (especially sections 3.1 and 3.2): *Measurement and Fairness. Jacobs, A. and Wallach, H.*

Ethical Issues with Digital Crowdsourcing Platforms

When we talk about human annotation, it's important to be aware that large online platforms where workers are paid to label examples (e.g., Amazon Mechanical Turk) come with their own set of issues. Workers are often underpaid — a 2018 study by the United Nations' International Labor Organization (ILO) found that average hourly rates were $4.43, and went down to $3.31 when time spent doing unpaid work was included. Labeling tasks range from benign to quite disturbing, even leading to PTSD in some cases. In order to get paid, there's often pressure to not deviate from the majority answer. And, there's a separate problem of workers being heavily concentrated in a few countries — leading to labels that are disproportionality worse for examples that rely on cultural or contextual knowledge that they may not have.

Assignment Overview

Note: There is an intermediate due date. Parts 0 and 1 are due on Thursday, 3/18. Parts 2 and 3 are due the following Thursday, 3/25. If you have logistical issues (e.g., you're not on the group assignment list, you didn't receive instructions for part 2, etc), email 6864-hw2@mit.edu.

1. **Part 0: (due Thursday 3/18 @ 12pm)**: Short-answer reflections to a few hypothetical scenarios.
    a. Submit short-answer responses here: Form for Part 0 answers
2. **Part 1 (due Thursday 3/18 @ 12pm):** This part is done in groups. You will have the role of a researcher in charge of creating a dataset. You'll receive a hypothetical task, make decisions about what labels you want to collect, and write instructions to a group of annotators.
    a. Find your group members here: Part 1 Groups
    b. Your task is in the leftmost column of the spreadsheet. For Task A, find the info here (Task A info). For Task B, find the info here (Task B info).
    c. Submit short-answer responses here: Form for Part 1 answers
    d. Email your instructions doc to the Kerberos IDs indicated in the group assignment spreadsheet.
3. **Part 2 (due Thursday 3/25 @ 12pm):** This part is done individually. You'll now be an annotator. First, take the instructions you wrote and annotate a new set of examples according to them. Then, you will receive instructions from a different group, for a different task, and will be asked to annotate a set of examples by following their instructions.
    a. Fill out *both* of the following:
        i. Annotation form for Task A
        ii. Annotations form for Task B
4. **Part 3: (due Thursday 3/25 @ 12pm):** Pick one of the following readings to read & respond to on Canvas.
    a. Data and its (dis)contents: A survey of dataset development and use in machine learning research. **(DISCUSSION LINK)**
    b. Data Feminism Chapter 4 (What Gets Counted Counts). **(DISCUSSION LINK)**
    c. Datasheets for datasets. **(DISCUSSION LINK)**
    d. Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria. **(DISCUSSION LINK)**
    e. A.I. Is Learning From Humans. Many Humans + optionally These companies claim to provide "fair-trade" data work. Do they? **(DISCUSSION LINK)**