Designing Data: Proactive Data Collection and Iteration for Machine Learning Using Reflexive Planning, Monitoring & Density Estimation



Aspen Hopkins*, Fred Hohman, Luca Zappella, Xavier Suau Cuadros, Dominik Moritz *dataspen@mit.edu

Motivation

Contribution

- Lack of diversity in data collection causes failures when deploying ML
- Post-collection interventions are time and resource intens

r = race_representation(

ative American or Alaskan Nati liddle Eastern or Arab America

ast Asian or Asian America

outh Asian or Indian Ame

a = age_composition

Hispanic or Latinx

he dashboard is split into four parts: (1) Describing data expectations, (2

ompted with a series of questions to define the semantics/classes and representation you a tion. Many questions are intended to promp reflexivity prior to data collection as a tool to

ation,

a collection, (3) What you data looks like right now, and (4) How familiar your data is to your model. Make sur

Who is collecting or directing data collection

- Focusing only (pipeline is not (
- New methods i

Designing data is an iterative approach to data collection. It includes (1) **Pre-Collection Planning**, (2) **Collection Monitoring**, & (3) **Data Familiarity** (an application of density estimation). Each intervention complements the others, ensuring the final dataset provides as comprehensive coverage as possible.



ML Collect Dashboard

Welcome to the ML Collect dashboard. The dashboard is spi data collection, (3) What you data looks like right now, and ((1) **before** continuing to (2), (3), and (4).

(1) Describing data expectations

In this section, you will be prompted with a series of question aiming for in your data collection. Many questions are intendminimize biases embedded in data. Complete this section be

who is collecting or directing data cc By minimize biases embedded in data. Complete this section be (4): g = gender_composition() g.gender female trans female trans female trans male intersex non-binary or gender variant not listed

Building representative datasets is an architector tains is torically difference below to let the dimensions you want to consider in your data collection. For each dimension, set the architector of Alaskan Native architector of Alaskan Native south Asian or Indian American relies on the efficacy data requirements.

With reflexive planning & by documenting expected distributions, collectors ensure these specifications are as comprehensive as possible before collecting.

ML Collect Dashboard



Own arm Pre-Collection Planning Collect Collection Monitoring Train Data Familiarity (DE) Deploy



2. Collection Monitoring

Despite best efforts, data collected might not match expectations.

By comparing expected distributions to collected data, we capture a dataset's evolution, allowing users to make targeted adjustments.

By understanding how a model perceives data, we can focus data collection efforts on the most useful subsets, reweighting or replacing data accordingly.

We use density estimation (DE) to uncover samples that are unfamiliar to the model those that either are not represented appropriately, are challenging to learn, or were erroneously collected.

Here, we learn a Gaussian Mixture model (GMM) on a network's layer activations:

$$p(x \mid \lambda) = \sum_{i=1}^{M} w_i g(x \mid \mu_i, \Sigma_i)$$

Where x is the matrix of layer activations, $w_i, i = 1,...,M$ are the mixture weights, and $g(x | \mu_i, \Sigma_i), i = 1,...,M$ are the densities. PCA is used to reduce the dimensionality. The resulting log-likelihood values are the **familiarity scores.**



3. Familiarity

Despite increased rigor in collection, expected and actual data distributions may not match learning needs of a model. These are used to debug a dataset early in data collection. Later, it informs data iteration, improving diversity and coverage. While DE for OOD detection is well studied, our use of DE to direct data work is unique.

Does auditing to increase data diversity improve model generalizability?



Is data familiarity useful for auditing ML models and datasets?

Multiple

Left -

Right -

large -

small-

Female Male

26-30

31-35

41-45

46-50

21-25

short

medium -

Diverse Data Intersectional Group Accuracy



Asia Whit Asia Blac Blac Blac Blac Blac Blac Blac Smal Female Ato 26-30 31-35 26-30 31-35 26-30 31-35 26-30 26-30 26-30 26-30 26-30 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-40 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 87-70 B Post-Familiarity Intervention Intersectional Change in Accuracy



© 2023 Apple Inc. All rights reserved.