Statistical inference on black-box generative models in the data kernel perspective space

Hayden Helm, Aranyak Acharyya, Younger Park, Brandon Duderstadt, Carey Priebe

Nomic AI x Helivan x Johns Hopkins University

Motivation

Fine-tuning, prompt augmentation have increased the number of effective "models" being deployed

⇒ **it is imperative to develop statistical tools to study *populations of models*.**

This work extends recent theoretical results on low-dimensional representations of black-box generative models to *inference* on the models.

The low-dimensional representations – or perspectives – of the models can be used to predict properties of fine-tuning mixtures // other model-level covariates such as sensitivity, toxicity, etc.

Inference in the low-dimensional space can offer greatly improved evaluation efficiency.

Methods

Consider models $\{f_i\}$, $i \in [n]$.
Consider queries $\{q_j\}$, $j \in [m]$.
Consider replicates $\{f_i(q_j)_k\}$, $k \in [r]$.
Let g be an embedding function that maps model responses $f_i(q_j)_k$ to $\mathbb{R}^p$
Let $\bar{X}_i \in \mathbb{R}^{m \times p}$ denote the matrix whose j-th row is $\bar{x}_{ij} = \frac{1}{r}\sum_k g(f_i(q_j)_k)$
Let $y_i \in \mathbb{R}$ be a model-level covariate ($y_i = s(f_i) \in \mathbb{R}^p$)

With $D_{ii'} = \frac{1}{m}||\bar{X}_i - \bar{X}_{i'}||_F$
and $\hat{\psi} := \text{MDS}(D) \in \mathbb{R}^{n \times d}$
and decision function $h : \mathbb{R}^d \to \mathbb{R}^p$
then ..

Theoretical Results:

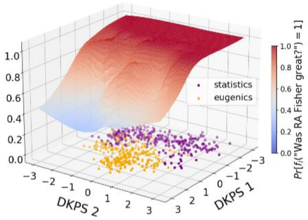
1. **Inference using the estimates of the true low-dimensional representations converges to inference using the true low-dimensional representations.**
2. **If h_{n} is consistent using true low-dimensional representations then h_{n} is consistent when using estimated low-dimensional representations**

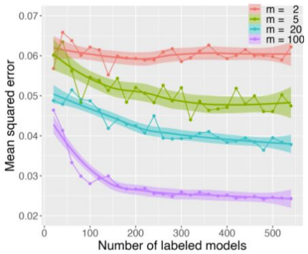
Theorem 1. *Under technical assumptions described in Appendix A,*

$$\mathcal{R}_\ell(P_{\psi Y}, h(\cdot\,;\hat{\mathcal{T}}_n)) \to \mathcal{R}_\ell(P_{\psi Y}, h(\cdot\,;\mathcal{T}_n))$$

as $m, r \to \infty$, for every n.

Theorem 2. *Under technical assumptions described in Appendix A, if $(h(\cdot\,;\mathcal{T}_1), \ldots, h(\cdot\,;\mathcal{T}_n))$ is consistent for $P_{\psi Y}$ with respect to $\mathcal{H}$, then $(h(\cdot\,;\hat{\mathcal{T}}_1), \ldots, h(\cdot\,;\hat{\mathcal{T}}_n))$ is consistent for $P_{\psi Y}$ with respect to $\mathcal{H}$ as $n, m, r \to \infty$.*



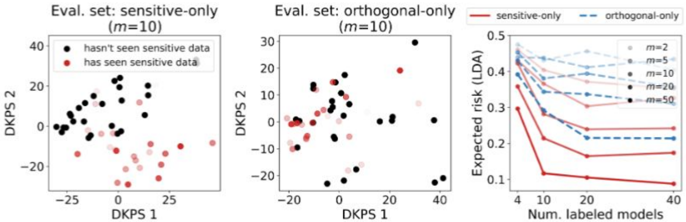


Predicting presence of data type in training mixture

y_{i} = "Was data type X in fine-tuning mixture" \in {0,1}

n=50 (25 where y_{i} = 0, 25 where y_{i} = 1); queries either relevant to X or orthogonal

h = 1-NN



Predicting sensitivity / toxicity

y_{i} = toxicity // bias \in R

n=59 (model graph around AlphaMonarch-7B)

queries from toxicity // bias benchmarks

h = 1-NN