Data Reliability Scoring

Yiling Chen[1] Shi Feng[1] Paul Kattuman[2] Fang-Yi Yu[3]

Harvard University[1] University of Cambridge[2] George Mason University[3]

Key Contributions

Can we assess the reliability of a dataset without access to the ground truth?

- Formalize the problem of reliability scoring based on observations from unknown experiments
- Propose dataset reliability orderings to benchmark reliability
- Design the Gram Determinant Score which can preserve dataset reliability orderings
- Conduct experiment on synthetic noise models and CIFAR-10 embeddings

Problem Formulation

Basic Setup

Reliability score $S(\hat{\boldsymbol{x}}, \boldsymbol{y})$ should assign a higher expected score to datasets that are closer to the true data.

Reports $\hat{\boldsymbol{x}}$	$\hat{x}_1$	$\hat{x}_2$	...	$\hat{x}_N$
			⇑ Misreport matrix Q	
True data $\boldsymbol{x}$	x_1	x_2	...	x_N
	↓	↓	Experiment P	↓
Observations $\boldsymbol{y}$	y_1	y_2	...	y_N

- Three datasets of length N
  - (Hidden) True data $\boldsymbol{x} = (x_1, \ldots, x_N) \in [d]^N$
  - Reported data $\hat{\boldsymbol{x}} = (\hat{x}_1, \ldots, \hat{x}_N) \in [d]^N$
  - Observable data $\boldsymbol{y} = (y_1, \ldots, y_N) \in \mathcal{Y}^N$
- Experiment $\boldsymbol{P} \in \mathbb{R}^{\mathcal{Y} \times d}$: $y_n \sim P_{x_n}$ independently for all n.
- Misreport matrix $\boldsymbol{Q} \in \mathbb{R}^{d \times d}$: $Q(i,j) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}[x_n = i, \hat{x}_n = j]$. Conditional frequency matrices $\boldsymbol{Q}_{\hat{\boldsymbol{x}}|\boldsymbol{x}}$ and $\boldsymbol{Q}_{\boldsymbol{x}|\hat{\boldsymbol{x}}}$ are column stochastic.

Examples

1. Insurance Reimbursement: Let $\boldsymbol{x}$ represent the true disease states of patients, $\hat{\boldsymbol{x}}$ the diagnoses reported by the hospital, and $\boldsymbol{y}$ auxiliary signals such as blood test results or imaging biomarkers.
2. Image Labeling: If the dataset includes image labels, let $\boldsymbol{x}$ be the true labels, $\hat{\boldsymbol{x}}$ the reported labels, and $\boldsymbol{y}$ encoder-based representations or other observations derived from the images.

Reliability Orderings of Datasets

To compare the reliability of different reported datasets, $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{x}}'$, relative to the true data $\boldsymbol{x}$, we propose strict partial orderings of reported datasets.

1. Exact Match Ordering: $\hat{\boldsymbol{x}} \succ^{\boldsymbol{x}}_{EXACT} \hat{\boldsymbol{x}}'$ if $\hat{\boldsymbol{x}} = \boldsymbol{x}$ but $\hat{\boldsymbol{x}}' \neq \boldsymbol{x}$.
2. Blackwell dominant ordering: $\hat{\boldsymbol{x}} \succ^{\boldsymbol{x}}_{Blackwell} \hat{\boldsymbol{x}}'$ if $\hat{\boldsymbol{x}}'$ is a garbling of $\hat{\boldsymbol{x}}$.
3. Hamming ordering: $\hat{\boldsymbol{x}} \succ^{\boldsymbol{x}}_{Hamming} \hat{\boldsymbol{x}}'$ if $\hat{\boldsymbol{x}}$ is closer to $\boldsymbol{x}$ than $\hat{\boldsymbol{x}}'$ in Hamming distance.

Reliability Scoring Preserving Reliability Orderings

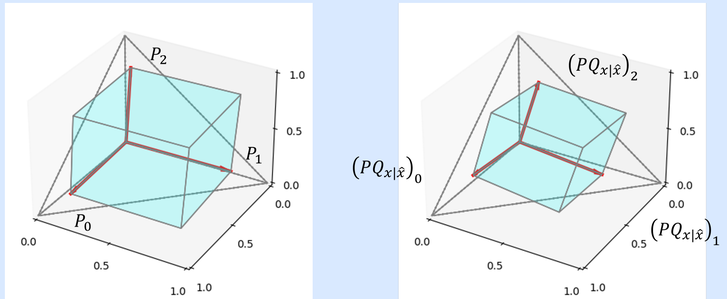
Given a reliability ordering $\succ^{\cdot}$, a reliability score S *preserves* the partial ordering $\succ^{\cdot}$ under experiment $\boldsymbol{P}$, if

$$\mathbb{E}_{\boldsymbol{y}\sim P(\boldsymbol{x})}[S(\hat{\boldsymbol{x}}, \boldsymbol{y})] > \mathbb{E}_{\boldsymbol{y}\sim P(\boldsymbol{x})}[S(\hat{\boldsymbol{x}}', \boldsymbol{y})]. \tag{1}$$

for all $\boldsymbol{x}, \hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}' \in \mathcal{X}^N$ with $\hat{\boldsymbol{x}} \succ^{\boldsymbol{x}} \hat{\boldsymbol{x}}'$.

Additionally, S *asymptotically* preserves $\succ^{\cdot}$ under a set of experiments $\mathcal{P}$ and a set of misreport matrices $\mathcal{Q}$, if for all $\boldsymbol{P} \in \mathcal{P}$ and $\boldsymbol{Q}, \boldsymbol{Q}' \in \mathcal{Q}$ and large enough N, S preserves $\succ^{\cdot}$ under $\boldsymbol{P}$ for all $\boldsymbol{x}, \hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}'$ with $\hat{\boldsymbol{x}} \succ^{\boldsymbol{x}} \hat{\boldsymbol{x}}'$ and misreport matrices $\boldsymbol{Q}, \boldsymbol{Q}'$.

Geometric Intuition of Gram Determinant



(a) True data $\boldsymbol{x}$ (b) Reported data $\hat{\boldsymbol{x}}$

Figure 1. Parallelepipeds spanned by (normalized) observations with $d = 3$ and $|\mathcal{Y}| = 3$

Given $\boldsymbol{P}$, the average distribution of observation under each true data label $x \in [d]$ is P_x, and the volume of the spanned parallelepiped is the *Gram determinant* of $\boldsymbol{P}$

$$vol(P_0, \ldots, P_{d-1}) = \sqrt{\det(\boldsymbol{P}^\top \boldsymbol{P})}.$$

Given $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$, the distribution under each reported label x becomes $(\boldsymbol{P}\boldsymbol{Q}_{\boldsymbol{x}|\hat{\boldsymbol{x}}})_x$, with volume $\sqrt{\det((\boldsymbol{P}\boldsymbol{Q}_{\boldsymbol{x}|\hat{\boldsymbol{x}}})^\top(\boldsymbol{P}\boldsymbol{Q}_{\boldsymbol{x}|\hat{\boldsymbol{x}}}))} \leq \sqrt{\det(\boldsymbol{P}^\top\boldsymbol{P})}$

Gram Determinant Reliability Scores

Gram determinant scores measure the volume spanned by the vectors of observed experiment outcomes.

Let $\boldsymbol{G} = \boldsymbol{P}^\top\boldsymbol{P}$ where $\boldsymbol{G}(x,x') = \langle P_x, P_{x'}\rangle$, the Gram matrix of reports $\hat{\boldsymbol{x}}$ be $\hat{\boldsymbol{G}} = \boldsymbol{Q}^\top\boldsymbol{G}\boldsymbol{Q}$ where $\hat{\boldsymbol{G}}(x,x') := \frac{1}{N^2}\sum_{n,n':\hat{x}_n=x,\hat{x}_{n'}=x'}\langle P_{x_n}, P_{x_{n'}}\rangle$, and the *Gram determinant score* is

$$\Gamma := \det\left(\hat{\boldsymbol{G}}\right). \tag{2}$$

The Gram determinant score in preserves

- exact match ordering under $\mathcal{P}_{indep}$ and $\mathcal{Q}_{nonperm}$,
- Blackwell ordering under $\mathcal{P}_{indep}$ and $\mathcal{Q}_{reg}$, and
- $\frac{1}{4L}$-Hamming ordering under $\mathcal{P}_{indep}$ and $\mathcal{Q}_{L,1/64L^2d^2}$.

Properties of Gram determinant reliability scores

Estimating Gram Determinant Scores

1. $\hat{\boldsymbol{G}}$ admits estimator $\bar{\boldsymbol{G}}(x,x') = \frac{1}{N^2}\sum_{n,n'\in[N]:\hat{x}_n=x,\hat{x}_{n'}=x'}\mathbf{1}[y_n = y_{n'}]$ which is consistent as $N \to \infty$.
2. We can decrease the bias of the above plug-in estimator $\bar{\boldsymbol{G}}$ through a carefully designed sampling scheme.

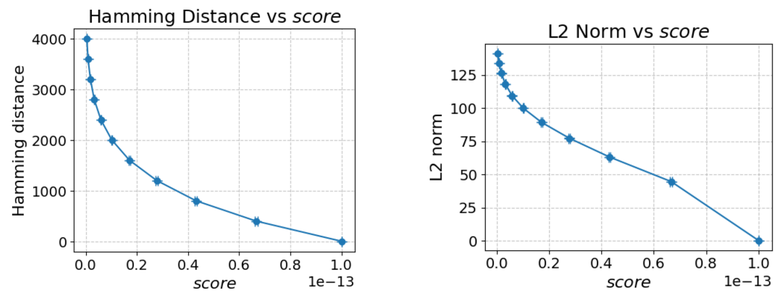
These estimators for the Gram determinant scores asymptotically preserve the above reliability orderings.

Kernelized Gram Determinant Scores

If $\mathcal{Y}$ has a kernel K, we can define $\boldsymbol{G}_K$ so that $\boldsymbol{G}_K(x,x') = \langle P_x, P_{x'}\rangle_K := \mathbb{E}_{y\sim P_x, y'\sim P_{x'}}[K(y,y')]$, $\hat{\boldsymbol{G}}_K = \boldsymbol{Q}^\top\boldsymbol{G}_K\boldsymbol{Q}$ and *Gram determinant score* with kernel K as $\Gamma_K := \det\left(\hat{\boldsymbol{G}}_K\right)$.

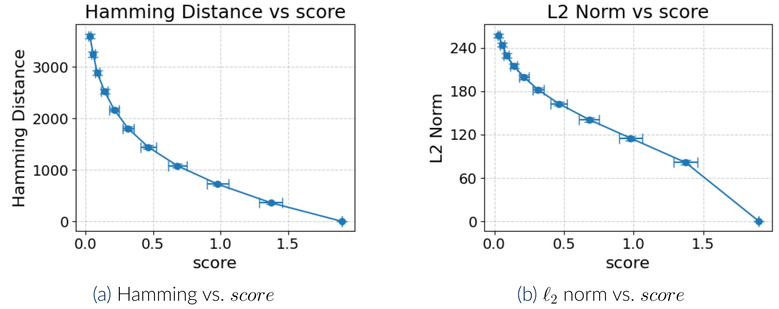
- Linear feature: $K(y,y') = \langle\phi(y), \phi(y')\rangle$, $\phi : \mathcal{Y} \to \mathbb{R}^k$ e.g. one-hot encoder.
- Gaussian RBF: $K(y,y') = \exp\left(-\frac{\|y-y'\|_2^2}{2\sigma^2}\right)$.
- Dot product: $K(y,y') = y^\top y'$ e.g., observation are forecast vectors.

Empirical Results



(a) Hamming vs. *score* (b) ℓ_2 norm vs. *score*

Figure 2. Gram Determinant Score on Synthetic Data



(a) Hamming vs. *score* (b) ℓ_2 norm vs. *score*

Figure 3. Gram Determinant Score with Kernels on Image Data

The Twenty-Sixth ACM Conference on Economics and Computation