Toward a Characterization of Loss Functions for Distribution Learning

Nika Haghtalab[1], Cameron Musco[2], and Bo Waggoner[3]

[1]Cornell University [2]UMass Amherst [3]U. Colorado

[1,2,3]Research conducted while at Microsoft Research

Summary

A common task in e.g. natural language processing is to learn a discrete distribution over a very large domain. But how do we **evaluate** a learned distribution $\mathbf{q}$ given samples from the truth $\mathbf{p}$? This paper proposes an axiomatic approach to selecting a loss function and finds that imposing the requirement of **calibration** allows many loss functions to satisfy the axioms.

Setting

True distribution: $\mathbf{p} \in \Delta_N$ — N exponentially large
Learned distribution: $\mathbf{q} \in \Delta_N$ — given to us by some algorithm
Loss functions: $\ell(\mathbf{q}, x)$ — loss of $\mathbf{q}$ on sample $x \in [N]$
Expected loss: $\ell(\mathbf{q}; \mathbf{p})$ — of $\mathbf{q}$ on a sample drawn from $\mathbf{p}$
Empirical distribution: $\hat{\mathbf{p}}$ — of some set of m samples
Empirical loss: $\ell(\mathbf{q}; \hat{\mathbf{p}})$ — average loss of $\mathbf{q}$ on the m samples

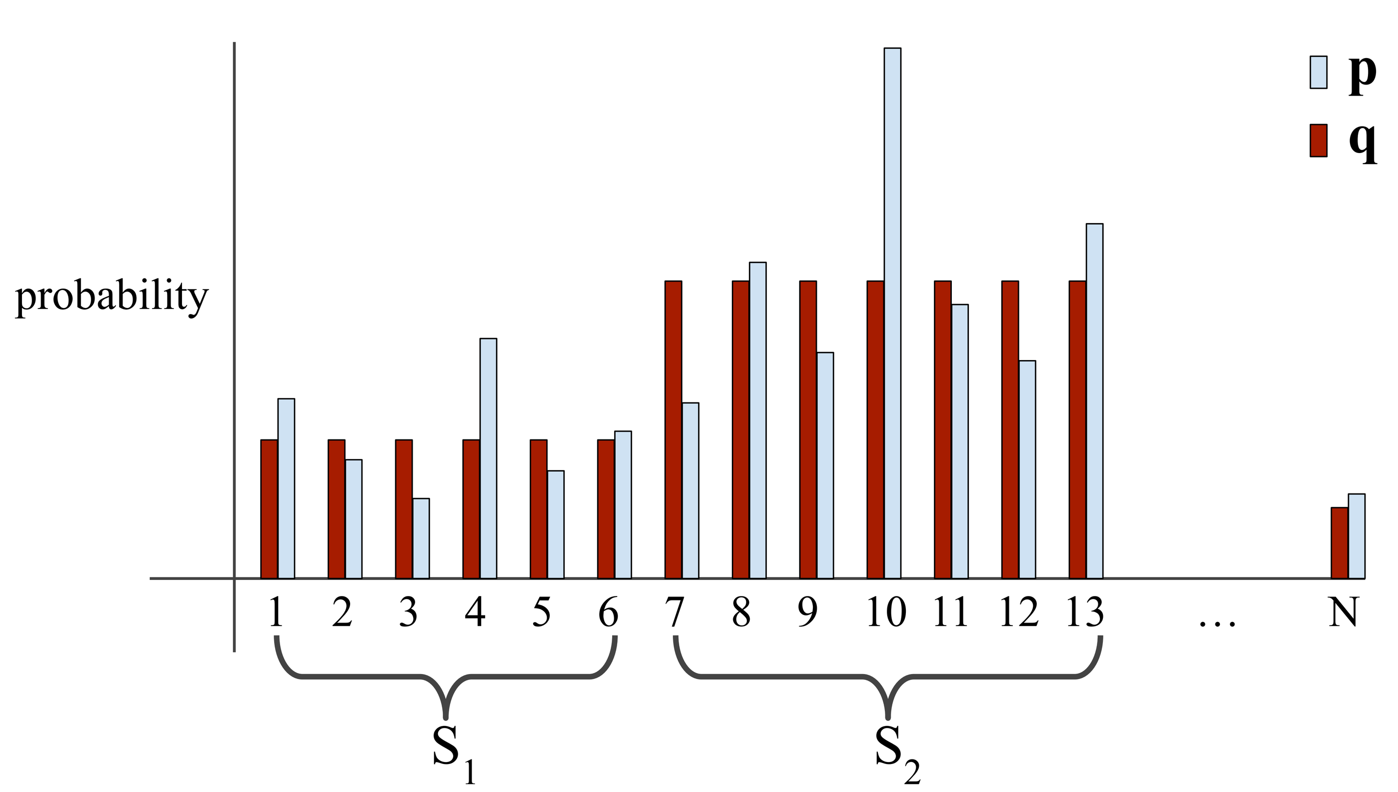
Log loss: $\ell(\mathbf{q}, x) = \ln\left(\frac{1}{x}\right)$

Calibration

$\mathbf{q}$ is **calibrated**[1, 2] with respect to $\mathbf{p}$ if the domain is partitioned by $S_1, \ldots, S_k$ where, for each S_i:

(1) $\mathbf{q}$ is uniform on S_i — i.e. $\mathbf{q}(x) = \mathbf{q}(y)$ for $x, y \in S_i$

(2) $\mathbf{q}(S_i) = \mathbf{p}(S_i)$ — $\implies$ average probabilities are equal



(need not be contiguous in general)

The Axioms

(1) **local:** $\ell(\mathbf{q}, x)$ depends only on q_x. — not $q_{x'}$ for any $x' \neq x$

(2) **strictly proper:** $\ell(\mathbf{q}; \mathbf{p}) > \ell(\mathbf{p}; \mathbf{p})$ for all $\mathbf{q} \neq \mathbf{p}$.
i.e. true distribution minimizes expected loss

(3) **β-strongly proper:** If $\|\mathbf{p} - \mathbf{q}\|_1 \geq \epsilon$, then $\ell(\mathbf{q}; \mathbf{p}) - \ell(\mathbf{p}; \mathbf{p}) \geq \frac{\beta}{2}\epsilon^2$.
log loss is 1-strongly proper $\iff$ Pinsker's inequality

(4) **sample proper:** If $\|\mathbf{p} - \mathbf{q}\|_1 \geq \epsilon$, then when drawing $m = \text{poly}(\frac{1}{\epsilon}, \log(N))$ samples, $\ell(\mathbf{q}; \hat{\mathbf{p}}) > \ell(\mathbf{p}; \hat{\mathbf{p}})$ w.high prob.
log loss is sample proper (folklore).

(5) **concentrating:** For any $\gamma > 0$, when drawing $m = \text{poly}(\frac{1}{\gamma}, \log(N))$ samples, $|\ell(\mathbf{q}; \hat{\mathbf{p}}) - \ell(\mathbf{q}; \mathbf{p})| \leq \gamma$ w.high prob.
log loss **does not** concentrate!

Key Points

(A) No loss function can satisfy all 5 axioms.
(B) But if we restrict to **calibrated** distributions $\mathbf{q}$, many losses satisfy all 5!
(C) We believe restricting to calibrated $\mathbf{q}$ is natural and may be feasible for learning algorithms.

Capturing Properties of Calibration

Lemma 1: If $\mathbf{q}$ is calibrated with respect to $\mathbf{p}$, then on any partition element S_i,

$$\mathbb{E}\left[\frac{1}{\mathbf{p}(X)} \,\middle|\, X \in S_i\right] = \mathbb{E}\left[\frac{1}{\mathbf{q}(X)} \,\middle|\, X \in S_i\right] = \frac{|S_i|}{\mathbf{p}(S_i)}.$$

Implication: If $\ell(\mathbf{q}, x) = f\left(\frac{1}{q_x}\right)$ for (left-strongly) concave f, then ℓ is (strongly) proper over calibrated $\mathbf{q}$.

Lemma 2: If $\mathbf{q}$ is calibrated with respect to $\mathbf{p}$, then for all x, $q_x \geq \left(\frac{1}{N}\right) p_x$.

Implication: If $\ell(\mathbf{q}, x) = f\left(\log(\frac{1}{q_x})\right)$ for left-strongly-concave, polynomial f, then ℓ is sample proper and concentrates over calibrated $\mathbf{q}$.

Results and Applications

Summary: Prove general conditions under which a loss of the form $\ell(\mathbf{q}, x) = f\left(\frac{1}{q_x}\right)$, for some f, satisfies axioms (1)-(5). (see "Capturing Properties of Calibration") Examples: Loss functions such as $\ell(\mathbf{q}, x) = \log(\log(\frac{e}{q_x}))$, $\sqrt{\log(\frac{1}{q_x})}$, $\left(\log(\frac{1}{q_x})\right)^2$, etc. satisfy (1)-(5).

Why satisfy the axioms? (Note: The space N may be **exponentially large**, e.g. all sentences of ≤ 50 words.)

(1) Can efficiently compute the loss from implicit representations of $\mathbf{q}$.
(2) Classical forecasting axiom: ground truth minimizes expected loss.
(3) Worse predictions have significantly larger expected loss.
(4) Few samples suffice to distinguish correct/incorrect distributions.
(5) Few samples suffice to accurately estimate actual expected loss.

Extensions and appendices:

- Results all extend to *approximate* calibration.
- One can efficiently post-process a learning algorithm to approximately calibrate it.

Implications for Practice

- ML currently struggles to rigorously evaluate distributions over large sample spaces (GANs, NLP applications, ...).
- This paper suggests imposing **calibration** on learning algorithms and evaluating with **loss functions** satisfying the axioms.
- The axioms may explain why log loss is so popular in practice...
- ...and open up alternatives such as $\text{poly}(\log(\frac{1}{q_x}))$ and more.

References

[1] A. Philip Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

[2] Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

Neural Information Processing Systems 2019
Vancouver, Canada