Bayesian-Optimal Multi-Classification implies Abstract Representations

Aman Bhargava

March 20, 2024

Consider an intelligent agent making decisions in some environment. The agent receives noisy observations conditioned on the environment state, and must produce optimal decisions (i.e., learn a multi-classification objective). We show that the agent must represent an estimate of the de-noised environment state if it optimally estimates decision output using noisy observations.

1 Problem Statement

Noisy Multi-Classifier: Formalize the "environment state" as $X \sim P(X)$ with sample space $\mathcal{X}$ and a corresponding ground truth decision set $P(Y_i|X)$ for $i \in [N]$ (e.g., multi-classification on the environment state). Denote the i.i.d. noise process $X_i \sim P(\tilde{X}_i|X)$ from which observations $\tilde{X}_i$ are sampled. We consider optimal estimators of the ground truth readout Y given noisy measurements $\tilde{X}$ denoted $P(\hat{Y}|\tilde{X}_1, \ldots, \tilde{X}_T)$.

$$\begin{tikzcd}
X \arrow[r, "\text{noise}"] \arrow[dr] & \{\tilde{X}_t\}_{t\in[T]} \arrow[r, "\text{agent}"] & \{\hat{Y}_i\}_{i\in[N]} \\
 & \{Y_i\}_{i\in[N]} &
\end{tikzcd} \tag{1}$$

Geometry: Let X reside in a metric space $\mathcal{X}$. Let each Y_i be defined in terms of a binary discriminator $\phi_i : \mathcal{X} \to \{0, 1\}$. Let the equivalence classes of $\mathcal{X}$ under each discriminator ϕ_i be connected (i.e., $\{x|\phi_i(x) = 1, x \in \mathcal{X}\}$ is connected for each ϕ_i).

Claim: Under fairly general conditions,

$$I(Z(t);X) = I(\tilde{X};X) \tag{2}$$

$$I(Z(t);X) = I(\tilde{X}_1 \dots \tilde{X}_T;X) \tag{3}$$

Proof Sketch:

- Derive $\hat{Y}_i \sim P(Y_i|\tilde{X}_t)$ using Bayes theorem. Due to independence, $P(Y_i|\tilde{X}_1,\dots,\tilde{X}_T)$ follows.
- Show that $P(Y_i|\tilde{X}_1,\dots,\tilde{X}_T)$ represents a distance between an implied $\hat{X} \sim P(X|\tilde{X}_1,\dots,\tilde{X}_T)$ and the boundary of the set $\{x|\phi(x) = 1, x \in \mathcal{X}\}$.
- Show that the set of N distances along with knowledge of the boundaries narrows $\hat{X}$ down to a point.

Results:

- Linear decision boundaries + Gaussian noise: proven in "Disentangling Representations in RNNs through Multi-task Learning".
- If X is continuously deformable to some $U = g(X)$ such that decision boundaries in U are linear, then the model must represent an optimal estimate of coordinates in U (new result).