Rashomon Capacity: Measuring Predictive Multiplicity in Classification

Hsiang Hsu and Flavio P. Calmon

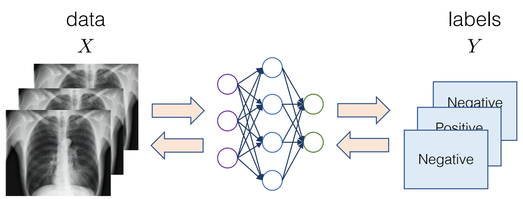
School of Engineering and Applied Science, Harvard University

Take-Away

1. Rashomon Capacity is a **computationally tractable metric** for predictive multiplicity
2. **Disclosing predictive multiplicity** is critical for **individual-level** decision-making, and current practice does not report it!

Rashomon Effect and Predictive Multiplicity

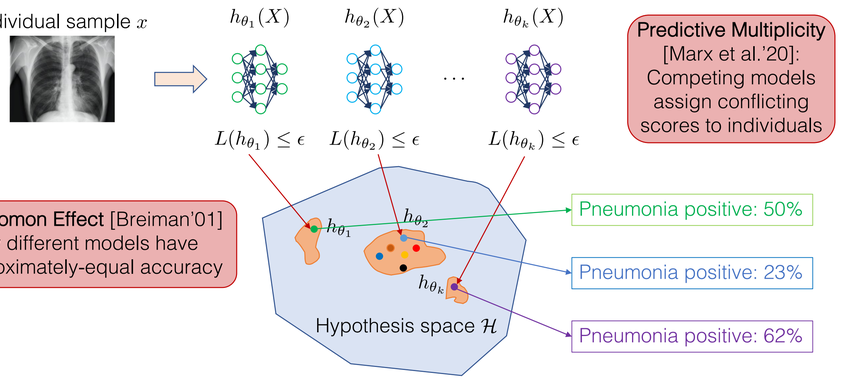
Training Time



Random Initializations
- Set random seeds
- Initialize weights
- Dropout rates
- Shuffling for batches
- …

$h_{\theta_1}(X), h_{\theta_2}(X), \cdots, h_{\theta_k}(X), \cdots, \theta_k \in \Theta$: model parameters (i.e., weights)

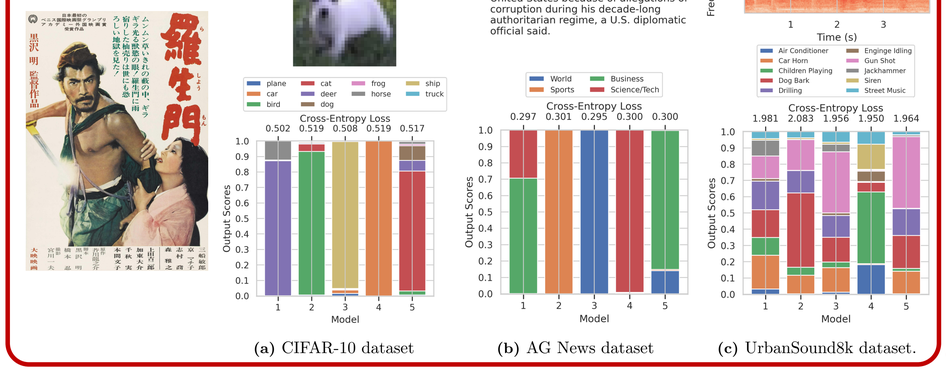
Test Time test loss: $L(h_\theta) = \mathbb{E}[\ell(h_\theta(X), Y)]$



Predictive Multiplicity [Marx et al.'20]: Competing models assign conflicting scores to individuals

Rashomon Effect [Breiman'01] Many different models have approximately-equal accuracy

Rashomon set: $\mathcal{R}(\mathcal{H}, \epsilon) \triangleq \{h_\theta \in \mathcal{H}; L(h_\theta) \le \epsilon\}$



(a) CIFAR-10 dataset (b) AG News dataset (c) UrbanSound8k dataset.

Rashomon Capacity

Desirable Properties
1. $1 \le m(x) \le c$
2. $m(x) = 1 \Rightarrow$ predictions from all models match
3. $m(x) = c \Rightarrow$ there are models in the Rashomon Set; that assign each of the c classes
4. Monotonic in $|\mathcal{R}(\mathcal{H}, \epsilon)|$

1. $h_{\theta_1}(x) = [0.05, 0.90, 0.05] \in \Delta_c$, $h_{\theta_2}(x) = [0.62, 0.34, 0.04] \in \Delta_c$, $h_{\theta_k}(x) = [0.25, 0.30, 0.45] \in \Delta_c$
2. $h_{\theta_1}(x) = [0.05, 0.90, 0.05] \in \Delta_c$, $h_{\theta_2}(x) = [0.05, 0.90, 0.05] \in \Delta_c$, $h_{\theta_k}(x) = [0.05, 0.90, 0.05] \in \Delta_c$
3. $h_{\theta_1}(x) = [\underline{1.00}, 0.00, 0.00] \in \Delta_c$, $h_{\theta_2}(x) = [0.00, \underline{1.00}, 0.00] \in \Delta_c$, $h_{\theta_k}(x) = [0.00, 0.00, \underline{1.00}] \in \Delta_c$
4. $h_{\theta_1}(x) = [0.05, 0.90, 0.05] \in \Delta_c$, $h_{\theta_2}(x) = [0.62, 0.34, 0.04] \in \Delta_c$, $h_{\theta_k}(x) = [0.25, 0.30, 0.45] \in \Delta_c$, $h_{\theta_{k+1}}(x) = [0.83, 0.08, 0.09] \in \Delta_c$

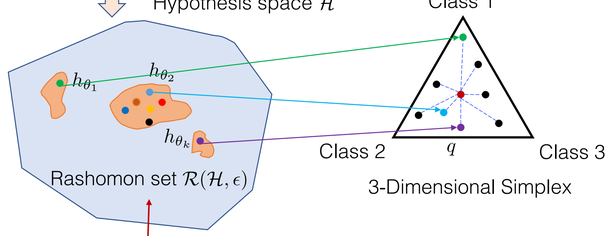
$m(\{h_{\theta_1}, \cdots, h_{\theta_k}\}) \le m(\{h_{\theta_1}, \cdots, h_{\theta_{k+1}}\})$

How do we measure the score variations?

$m : \mathcal{M}_\epsilon(x) \to \mathbb{R}^+$

Individual sample x

$\mathcal{M}_\epsilon(x) \triangleq \{h_\theta(x); h_\theta \in \mathcal{R}(\mathcal{H}, \epsilon)\}$

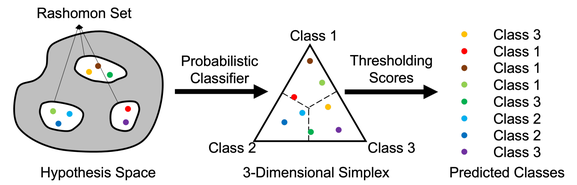


Distribution over models in the Rashomon set

$\sup_{P_{\mathcal{M}_{\epsilon'}(x)}} \inf_{\boldsymbol{q} \in \Delta_c} \mathbb{E}_{h \sim P_{\mathcal{M}_\epsilon(x)}} [D_{KL}(h(x) \| \boldsymbol{q})] \triangleq C_\epsilon(x)$

Rashomon Capacity: $m_\epsilon(x) \triangleq 2^{C_\epsilon(x)}$

Other Metrics



Rashomon ratio
Pattern Rashomon ratio

Rashomon Capacity

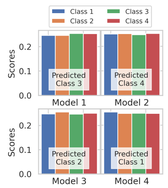
Ambiguity
Discrepancy

Rashomon Ratio

$\hat{\mathcal{R}}(\mathcal{H}, \epsilon) \triangleq \frac{\text{vol}(\mathcal{R}(\mathcal{H}, \epsilon))}{\text{vol}(\mathcal{H})}$

Pattern Rashomon Ratio

$\tilde{\mathcal{R}}(\mathcal{H}, \epsilon) \triangleq \frac{\sum_{j=0}^{2^n-1} \mathbb{1}[\exists h \in \mathcal{R}(\mathcal{H}, \epsilon): \zeta(h, \mathcal{D}) = \text{binary}(j)]}{\sum_{j=0}^{2^n-1} \mathbb{1}[\exists h \in \mathcal{H}: \zeta(h, \mathcal{D}) = \text{binary}(j)]}$



Ambiguity

$\alpha_\epsilon(\hat{h}) = \frac{1}{n} \sum_{i=1}^n \max_{h \in \mathcal{R}(\mathcal{H}, \epsilon)} \mathbb{1}\left[\arg\max h(\mathbf{x}_i) \neq \arg\max \hat{h}(\mathbf{x}_i)\right]$

Discrepancy

$\delta_\epsilon(\hat{h}) = \max_{h \in \mathcal{R}(\mathcal{H}, \epsilon)} \frac{1}{n} \sum_{i=1}^n \mathbb{1}\left[\arg\max h(\mathbf{x}_i) \neq \arg\max \hat{h}(\mathbf{x}_i)\right]$

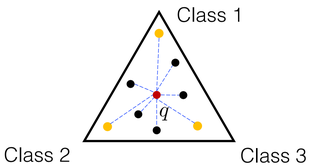
Computational Challenges

1. How to choose ϵ?
2. Approximating the Rashomon set without exhaustively searching?

Using a reference model and approximating the true Rashomon set by a Rashomon subset

$$\widetilde{\mathcal{R}}(\mathcal{H}, \epsilon') \triangleq \{h_{\theta_i} \in \mathcal{H}; L(h_{\theta_i}) \le \hat{L}(h_{\theta^*}) + \epsilon'\}_{i=1}^K \subseteq \mathcal{R}(\mathcal{H}, \hat{L}(h_{\theta^*}) + \epsilon')$$

For each sample x, here are at most c models in a Rashomon subset $\widetilde{\mathcal{R}}(\mathcal{H}, \epsilon)$ whose output scores yield the same Rashomon Capacity for x as the entire Rashomon set.

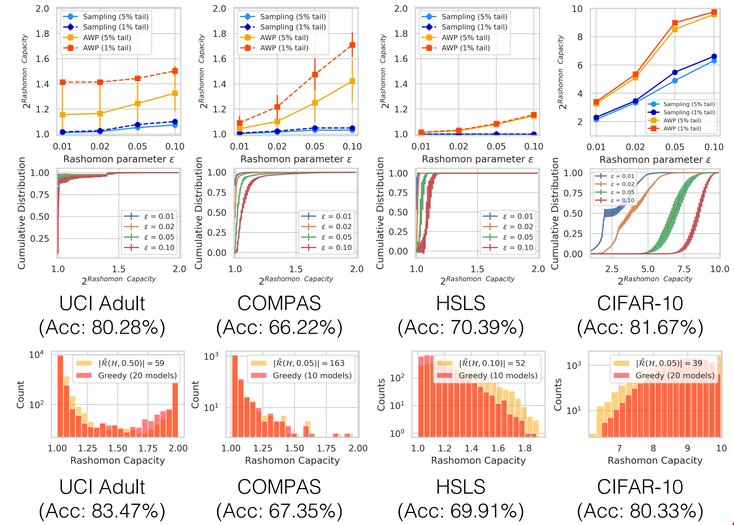


Adversarial Weight Perturbation (AWP)

$\mathbf{p}_k = h_{\hat{\theta}}(\mathbf{x}_i)$, where $\hat{\theta} = \underset{\theta \in \Theta,\ h_\theta \in \mathcal{R}(\mathcal{H}, \epsilon)}{\arg\max} [h_\theta(\mathbf{x}_i)]_k$, for all classes $k \in [c]$

Rashomon Capacity can be computed by the Blahut—Arimoto (BA) algorithm

Empirical Results



UCI Adult (Acc: 80.28%) | COMPAS (Acc: 66.22%) | HSLS (Acc: 70.39%) | CIFAR-10 (Acc: 81.67%)

UCI Adult (Acc: 83.47%) | COMPAS (Acc: 67.35%) | HSLS (Acc: 69.91%) | CIFAR-10 (Acc: 80.33%)