Deep Gamblers: Learning to Abstain with Portfolio Theory

Liu Ziyin (UTokyo), Zhikang T. Wang (UTokyo), Paul Liang (CMU), Ruslan Salakhutdinov (CMU), Louis-Philippe Morency (CMU), Masahito Ueda (UTokyo),

Classification and the Inadequacy of \( nll \) loss

Want to find: \( \theta = \arg \max_{\theta} \Pr(Y|\theta) \)

In practice, minimize negative log loss (\( nll \) loss): \( \min_{\theta} - \log(p(Y|\theta)) \)

Intuition: Prediction as Horse Race

Horse Race with Reservation

- \( m \) horses
- Betting strategy: \( \sum_{i=1}^{m} b_i \rightarrow \sum_{i=0}^{m} b_i \)
- Chance of winning: \( p_i \)
- Payoff if we bet on the winning horse: \( a_i \)
- Return after winning: \( S = a_i b_i \rightarrow a_i b_i + b_0 \)

Objective: maximize doubling rate:

\[
\max W = \max \mathbb{E} \log(S) = \max \sum_{i=1}^{m} p_i \log(a_i b_i + b_0)
\]

Classification Problem = Betting problem with Reservation

with \( a = 1, b_0 = 0 \)

Classification Problem \( \leq \) Betting problem with Reservation

Surprising Benefit:
- Training with gambler’s loss reduces overfit
- Improved performance when noisy label is present

The proposed method: the gambler’s loss

\[
\max \mathbb{E} \log(S) = \max \sum_{i=1}^{m} p_i \log(a_i b_i + b_0)
\]

The Learned Representation is Better Separable:

Figure 1. Top-5 rejected images in the MNIST testing set found by two methods. The number above image is the predicted uncertainty score (ours) or the entropy of the prediction baseline. For the top 2 images, our method chooses images that are hard to recognize, while that of the baseline can be identified unambiguously by human.

SOTA Performance...

Table 3: SVHN. The number is error percentage on the covered dataset; the lower the better. We see that our method achieves competitive results across all coverage. It is the SOTA method at coverage (0.85, 1.00).

Table 4: Cifar. The number is error percentage on the covered dataset; the lower the better. This dataset is binary classification, and the input images have large resolution.