Convergence Behavior of an Adversarial Weak Supervision Method

Steven An [1] Sanjoy Dasgupta [1]

[1]Department of Computer Science and Engineering
University of California, San Diego

Abstract

Suppose one has rules-of-thumb (e.g. "document contains 'birdie' $\Rightarrow$ 'sports' is its topic") that predict on unlabeled datapoints. How can one combine them to form a single predictor when rules have different accuracies and are sometimes contradictory? There are two overarching approaches to this problem: probabilistic and adversarial. The former is popular in practice, but we show that the latter has strong theoretical results and good experimental performance.

Formal Model

- n unlabeled datapoints $X = \{x_1, x_2, \ldots, x_n\} \subset \mathcal{X}$
- k labels denoted $\mathcal{Y} = \{1, 2, \ldots, k\}$
- v labeled datapoints $X^L = \{(x_{n+1}, y_{n+1}), \ldots, (x_{n+v}, y_{n+v})\}$
- p rules-of-thumb that can each abstain:

$$h^{(1)}, h^{(2)}, \ldots, h^{(p)} \colon X \to \mathcal{Y} \cup \{?\}$$

Goal: For each datapoint x_i and labels $1 \leq \ell \leq k$, infer the underlying conditional label probability

$$\eta_{i\ell} = \Pr(y_i = \ell \mid x_i).$$

Representing the Probabilistic Approach: One-Coin Dawid-Skene (OCDS)

Labels y_i are drawn i.i.d. and rule predictions $h^{(j)}(x_i)$, $h^{(j')}(x_i)$ are independent given label y_i. Important (underlying) quantities are the class frequencies $\tau_\ell = \Pr(y = \ell)$ and rule accuracies $b_j = \Pr(h^{(j)}(x) = y)$. $h^{(j)}$ predicts correct label y_i w.p. b_j. Otherwise, a random label is selected from $\mathcal{Y} \setminus \{y_i\}$ w.p. $1 - b_j$. I.e. whether $h^{(j)}$ is right on x_i is a coin flip.

- OCDS prediction for $\eta_{i\ell}$ is the posterior label probability with respect to generative process, using τ_ℓ, b_j.
- In practice, we estimate τ, b_j. EM (which maximizes likelihood) is one way, but other methods are possible.

Representing the Adversarial Approach: Balsubramani-Freund (BF)

If we don't specify a generative process and instead suppose an adversary labels the data, how should we predict?

- By using X^L we can approximate the rule accuracies, allowing us to construct a polytope P of feasible labelings. W.h.p. this contains the underlying labelings.
- A lower (upper) bound on $h^{(j)}$'s accuracy on X is a halfspace. P is the intersection of many such halfspaces.
- To pick a single labeling g, we play a game where the adversary picks labeling $z \in P$ maximizing our prediction g's loss. We pick g minimizing the worst case log-loss the adversary can inflict on us. Formally, the minimax game is

$$\min_{g} \max_{z \in P} -z^\top \log g.$$

Questions of interest: What is the optimal g for the above game? What are its theoretical properties? How does it compare to other methods of combining rules-of-thumb?

Our Contributions

Functional Form of OCDS/BF Predictions

OCDS and BF generate predictions from the same exponential family, which we denote by $\mathcal{G}$.

BF Prediction is Max Entropy Distribution

The optimal choice of g from above is the maximum entropy distribution with respect to P.

BF's Consistency/Rates of Convergence

As the polytope shrinks, BF's prediction converges to the unique best approximator in $\mathcal{G}$ to the underlying conditional label probabilities η. We provide rates of convergence in terms of the polytope P's "size".

Our Contributions (Cont.)

OCDS + EM's Inconsistency

We exhibit a problem where OCDS + its EM algorithm never converges to the best approximator of η in $\mathcal{G}$.

Model and Approximation Uncertainty

- Model Uncertainty is the irreducible error from having to approximate η via picking predictions from $\mathcal{G}$.
- Approximation Uncertainty is the reducible error from not picking the best approximator of η in $\mathcal{G}$.

I.e. the bigger $\mathcal{G}$ is, the lower the model uncertainty. Since BF and OCDS both predict from $\mathcal{G}$, their model uncertainties are the same and it suffices to study their approximation uncertainty to fairly compare them.

BF/OCDS Error Comparison

For every OCDS prediction, we provide a sufficient condition for BF to generate a prediction that is no worse than said OCDS prediction. I.e. we give a concrete condition on the size of P required for the BF prediction to be no worse than the OCDS one with respect to log loss.

Select Experimental Results Showing BF's Viability

Table 1. Avg. Log Loss of BF ($v = 100$) vs Other WS Methods ($v = 0$)

Method	Basketball	Domain	IMDB	SMS	Yelp	Youtube
MV	2.40	5.48	6.39	0.79	5.90	1.27
OCDS	3.75	22.32	2.91	0.78	1.73	17.63
DP	1.31	9.21	0.68	0.53	2.61	0.72
EBCC	0.45	1.80	0.73	0.43	0.81	0.69
HyperLM	1.31	1.29	0.62	0.68	**0.60**	**0.42**
BF	**0.39**	**1.12**	**0.59**	**0.42**	0.64	0.50
$\frac{1}{n} d(\eta, g^*)$	0.32	1.01	0.57	0.25	0.54	0.21