Weak Supervision,

noisy labels, and error propagation

Marat Freytsis

hep-ai journal club — December 11, 2018

based on Yu *et al.* [arXiv:1402.5902], Cohen, MF, Osdiek [arXiv:1706.09451] + bits of others

Why Weak supervision?

Fully supervised learning on real data often difficult/impossible

- Individual labels are prohibitively expensive to assign
- Personalized information legally protected (*e.g.*, medical, demographic data)
- For quantum systems, unique labels may be unphysical

Several classes of learning tasks on partially labels well developed

- semi-supervised: augmenting labeled with unlabeled data
- multiple instance: presence of signal in bag is marked but not identified

One which nicely maps onto many scientific data measurements is Learning from Label Proportions

Plan

- Learning from Label Proportions
- Viability and generalization error
- Proportion uncertainties, stability, and error propagation

Learning from Label Proportions

general setting

Domain of instance features denoted by $\mathcal{X}$ and (discrete) labels by $\mathcal{Y}$. Data consists of bags of events with features $\tilde{\mathbf{x}} = (\mathbf{x}_1, \ldots \mathbf{x}_r)$ and labels $\tilde{y} = (y_1, \ldots, y_r)$, drawn iid from a distribution over $(\mathcal{X} \times \mathcal{Y})^r$.

Learner has no access to labels, but instead receives label proportions $(\tilde{\mathbf{x}}, f_i(\tilde{y}))$, with $f_i(\tilde{y}) = \sum_{n=1}^{r} \mathbb{I}_{y_n=i}/r$. From a set of m bags, the task is to find a classifier from individual events to labels.

For experimental measurements, $f_i(\tilde{y})$ can be naturally interpreted as, *e.g.*, a rate/cross-section measurement/calculation even if individual events cannot be perfects separated by their features

Is this even possible?

heuristic argument

Consider binary classification ($y_i = \{0, 1\}$). Discretize data into bins $b_{m,j}$. If 2 bags are present, in each bin

$$\begin{aligned} b_{A,j} &= f_{A,1} b_{1,j} + (1 - f_{A,1}) b_{0,j} \\ b_{B,j} &= f_{B,1} b_{1,j} + (1 - f_{B,1}) b_{0,j} \end{aligned} \implies \begin{aligned} b_{0,j} &= \frac{f_{A,1} b_{B,j} - f_{B,1} b_{A,j}}{f_{A,1} - f_{B,1}} \\ b_{1,j} &= \frac{(1-f_{B,1}) b_{A,j} - (1-f_{A,1}) b_{B,j}}{f_{A,1} - f_{B,1}} \end{aligned}$$

and the distributions can be inverted algebraically.

Requirements:

- Number of bags $\geq$ number of classes to be distinguished, with label proportions unique for each bag.
- The bags need to be drawing from the same underlying distribution for each class, *i.e.*, however the label proportions were made different should be uncorrelated from the distribution over $(\mathcal{X} \times \mathcal{Y})^r$.

Classification in practice

Don't want to discretize, no guarantee events sample feature space densely enough it even makes sense. How to classify events? Modify loss function!

1. direct attack:

$$\ell_{\text{LLP}} = \arg\min_{h \in \mathcal{H}} \ell(\langle h(\mathbf{x}_i)\rangle_{\text{batch}}, \langle f(\tilde{y})\rangle_{\text{batch}})$$

   typically need re-optimization of hyperparameters
2. clever trick (classification without labels):

$$\ell_{\text{CWoLa}} = \arg\min_{h \in \mathcal{H}} \ell(h(\mathbf{x}_i), f(\tilde{y}))$$

Metodiev *et al.* [arXiv:1708.02949]

   with your fully-supervised loss function of choice

Classification without labels

why does the second version work at all?

Theorem

Given mixed samples M_1 and M_2 defined in terms of pure samples S and B with signal fractions $f_1 > f_2$, an optimal classifier trained to distinguish M_1 from M_2 is also optimal for distinguishing S from B.

Proof.

The optimal classifier to distinguish examples drawn from p_{M_1} and p_{M_2} is the likelihood ratio $L_{M_1/M_2}(\mathbf{x}) = p_{M_1}(\mathbf{x})/p_{M_2}(\mathbf{x})$. Similarly, the optimal classifier to distinguish examples drawn from p_S and p_B is the likelihood ratio $L_{S/B}(\mathbf{x}) = p_S(\mathbf{x})/p_B(\mathbf{x})$. Where p_B has support, we can relate these two likelihood ratios algebraically:

$$L_{M_1/M_2} = \frac{p_{M_1}}{p_{M_2}} = \frac{f_1 p_S + (1-f_1)p_B}{f_2 p_S + (1-f_2)p_B} = \frac{f_1 L_{S/B} + (1-f_1)}{f_2 L_{S/B} + (1-f_2)},$$

which is a monotonically increasing rescaling of the likelihood $L_{S/B}$ as long as $f_1 > f_2$, since $\partial_{L_{S/B}} L_{M_1/M_2} = (f_1 - f_2)/(f_2 L_{S/B} - f_2 + 1)^2 > 0$. If $f_1 < f_2$, then one obtains the reversed classifier. Therefore, $L_{S/B}$ and L_{M_1/M_2} define the same classifier. □

Only makes sense for binary classification!

Still need to know label proportions to calibrate classifier.

Plan

- Learning from Label Proportions
- **Viability and generalization error**
- Proportion uncertainties, stability, and error propagation

When is all of this viable?

All of this should clearly work in at least some cases, but can we know when will fails?

It turns out the classification without labels results are more general than they seem. Under mild assumptions (more later) a classifier which can accurately predict bag proportions can be guaranteed to achieve low error on event labels.

More precisely, for $\phi_r(h) : \mathcal{X}^r \to \mathbb{R}$, $\phi_r(h)(\tilde{\mathbf{x}}) = \sum_{n=1}^{r} h(\mathbf{x}_i)/r$, the classifier selected by

$$\arg\min_{h \in \mathcal{H}} \sum_{\text{bags}} \ell(\phi_r(h), f(\tilde{y}))$$

will also solve the original task with high accuracy.

Generalization errors for label proportions

For a given empirical bag label proportion error for loss function ℓ, $\mathrm{err}^{\ell}(h)$, it is possible to prove a bound on the expected error over the full distribution $\mathcal{X} \times \mathcal{Y}$,

$$\mathrm{err}^{\ell}_{G}(h) = \mathbb{E}_{(\tilde{\mathbf{x}},\tilde{y})}\ell(\phi_r(h), f(\tilde{y})).$$

As a function of the VC dimension of the hypothesis class, with probability $1-\delta$, $\mathrm{err}^{\ell}_{G}(h) \leq \mathrm{err}^{\ell}(h) + \epsilon$ if the number of bags m is

$$m \geq \frac{64}{\epsilon^2}\left(2VC(\mathcal{H})\log\frac{12r}{\epsilon} + \log\frac{4}{\delta}\right).$$

The mild dependence on bag size r means that destabilizing the method by adding more data is not a large concern.

for this proof and following, see arXiv:1402.5902

Event errors from proportion errors

With some mild assumptions, the above founds can be extended to individual events.

If $\mathrm{err}_G^\ell(h) \leq \epsilon$ with probability $1-\delta$, and each bag is at least $(1-\eta)$-pure $1-\rho$ of the time, then $h(\mathbf{x})$ correctly classifies a fraction $(1-\tau)(1-\delta-\rho)(1-2\eta-\epsilon)$ of N events with probability

$$1 - e^{-\frac{N\tau^2}{2}(1-\delta-\rho)(1-2\eta-\epsilon)}.$$

Unfortunately, these bounds are somewhat weak. Guaranteed high performance generically requires extremely pure samples.

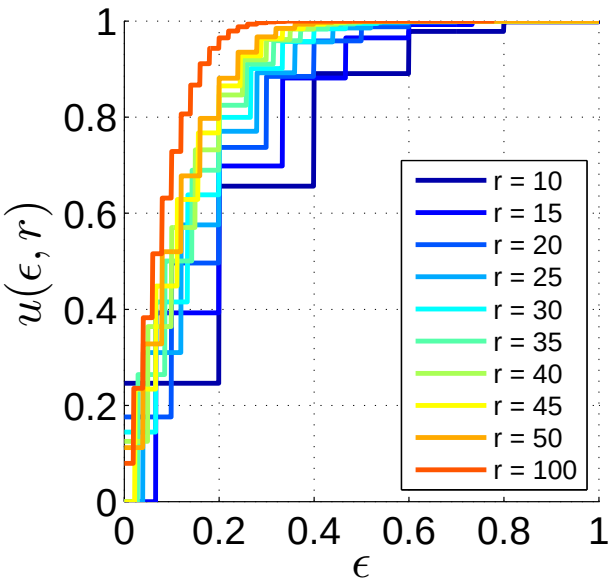
Class distribution independence

The preceding was so weak because no conditional independence of the underlying distributions from the bags was assumed, *i.e.*, the assumption that allowed us to invert the class distributions earlier.

If all bags are drawn from mixtures of underlying class distributions with different fractions, the probability of event error can be written as a generative model.

For binary classification, the probability of getting a classifier with error $\leq \epsilon$ is then bounded from below by $u(\epsilon, r)$.

The general answer becomes quite involved in this case, and I won't attempt to reproduce it.



Plan

- Learning from Label Proportions
- Viability and generalization error
- Proportion uncertainties, stability, and error propagation

Label uncertainties

The supervised aspect comes from the provided label proportions. What if these are wrong?

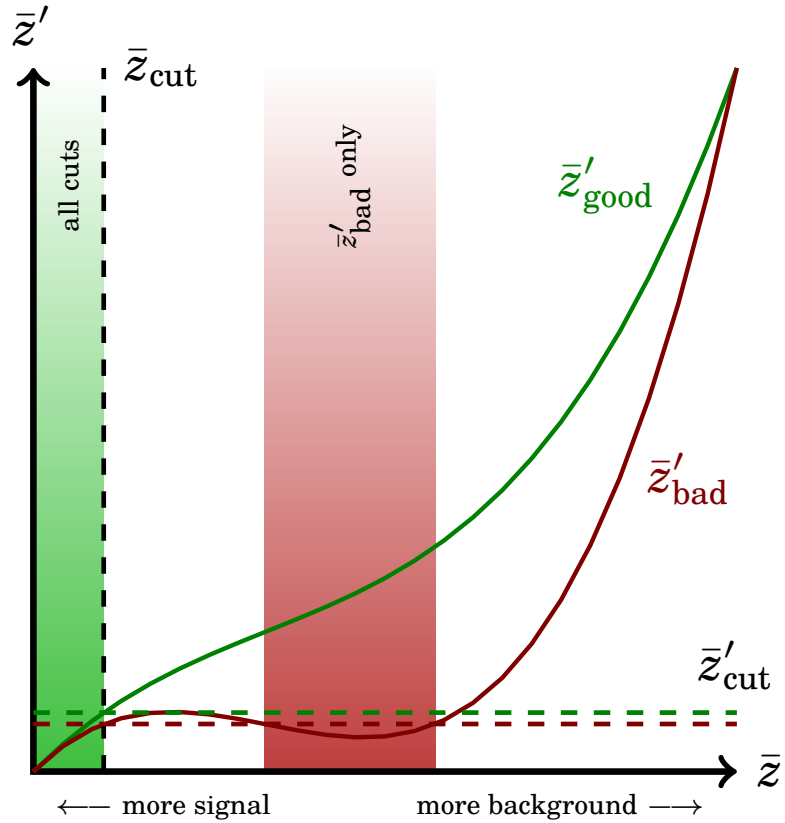
Return to the heuristic argument

$$\begin{aligned} b_{A,j} &= f_{A,1} b_{1,j} + (1 - f_{A,1}) b_{0,j} \\ b_{B,j} &= f_{B,1} b_{1,j} + (1 - f_{B,1}) b_{0,j} \end{aligned} \quad \Longrightarrow \quad \begin{aligned} b_{0,j} &= \frac{f_{A,1} b_{B,j} - f_{B,1} b_{A,j}}{f_{A,1} - f_{B,1}} \\ b_{1,j} &= \frac{(1-f_{B,1}) b_{A,j} - (1-f_{A,1}) b_{B,j}}{f_{A,1} - f_{B,1}} \end{aligned}$$

A Neyman–Pearson-optimal classifier is $z = b_0/(b_0 + b_1)$. The dependence on the error from a shift/uncertainty in any label proportion can be worked out analytically.

Label insensitivity

cartoon version



As long as the resulting distortion is monotonic, the classifiers are equivalent

Label insensitivity

concrete example

For a binary classifier and 2 bags with error $f_{A,1} \to f_{A,1} + \delta$,

$$\bar{z}' = \frac{1 - f_B}{1 - 2f_B} \frac{\frac{1-f_A-\delta}{1-f_B} - r(\mathbf{x})}{\frac{1-2f_A-2\delta}{1-2f_B} - r(\mathbf{x})} = \bar{z}_i + \delta \left(\frac{\frac{1-f_B}{1-2f_B} - \frac{\bar{z}_i}{1-2f_B} + 2(\bar{z}_i^2 - \bar{z}_i)}{\frac{f_A-f_B}{1-2f_B} + 2\delta(\frac{1-f_B}{1-2f_B} - \bar{z}_i)} \right),$$

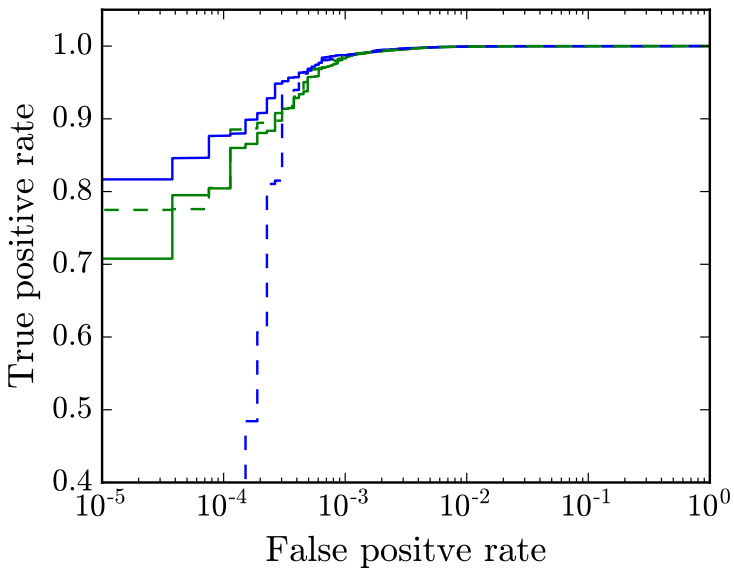
where $r(\mathbf{x}) = b_A(\mathbf{x})/b_B(\mathbf{x})$ is the ratio of inferred bag distributions.

The classifier remains equivalent to the optimal one if

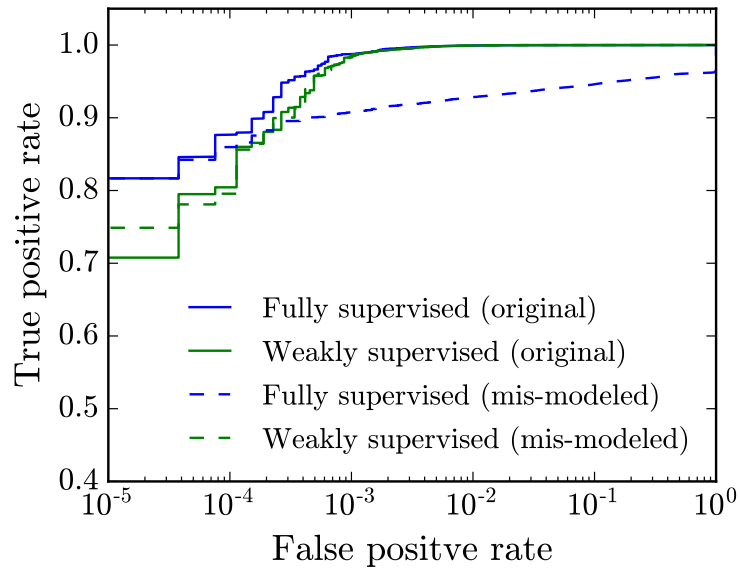
$$\delta \lesssim \frac{f_A - f_B}{3 - 2\min(f_B, 1 - f_B)}$$

A numerical study

impact of mismodelling



randomly swap 15% of each class



swap the 10% (15%) most signal-like (background-like)

Using random mutli-gaussian mixture models

Concluding thoughts

- Can bounds on generalization errors be made stronger without assuming distribution independence? (Or assuming something weaker)
- Understand how optimality arguments change with finite statistics/correlations?
- Can we propagate input uncertainties through the network?
  - Where would this be useful?
- Thank you!