Weak Supervision with Incremental Source Accuracy Estimation

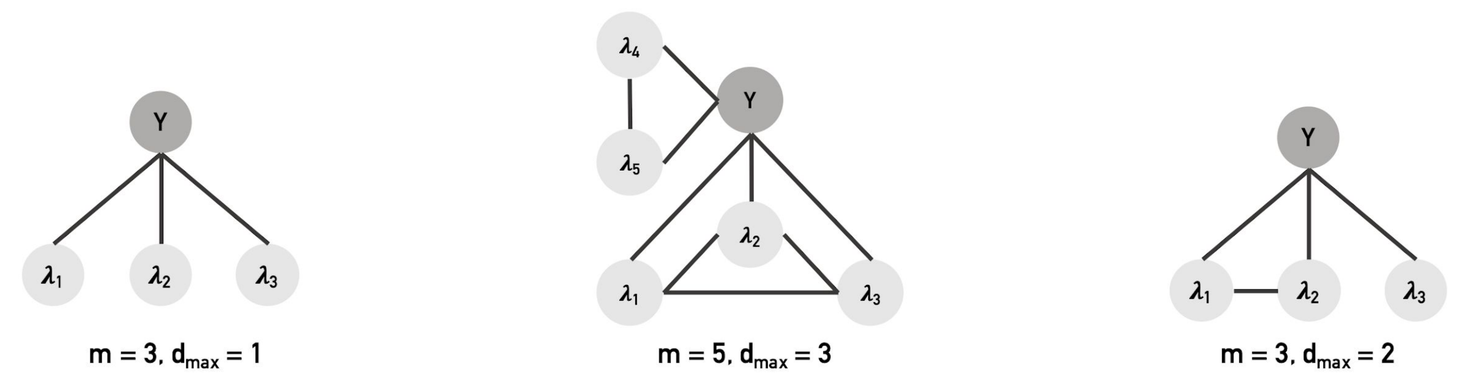
Richard Correro
rcorrero@stanford.edu

Weak Supervision Overview

Weak supervision approaches obtain labels for training data using noisier or higher-level sources than traditional supervision [1].

Recently proposed methods use generative models to combine labels from multiple noisy sources to generate probabilistic labels when true labels are unknown [2].

Varma et. al., [2] model the joint distribution of the weak supervision sources which produce noisy labels λ_i for $1 \leq i \leq m$ and the latent true label Y as a Markov random field associated with a graph $G = (V, E)$ where $V = \{\lambda_1, \ldots, \lambda_m\} \cup \{Y\}$. If λ_i is not independent of λ_j conditioned on Y and the other sources then $(\lambda_i, \lambda_j) \in E$.



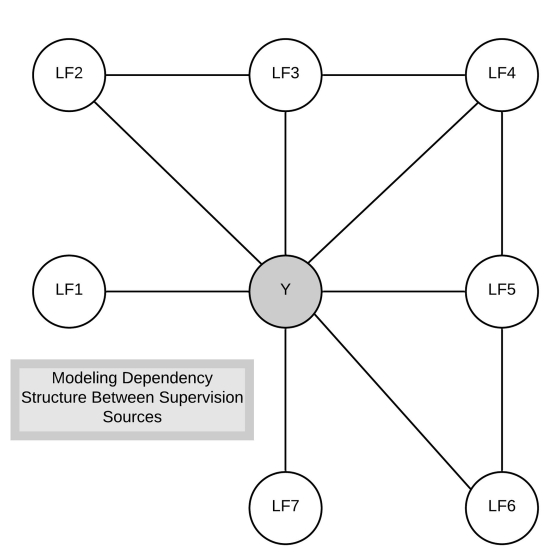
Model:

$$f_G(\lambda_1, \ldots, \lambda_m, y) = \frac{1}{Z} \exp\left(\sum_{\lambda_i \in V} \theta_i \lambda_i + \sum_{(\lambda_i, \lambda_j) \in E} \theta_{i,j} \lambda_i \lambda_j + \theta_Y y + \sum_{\lambda_i \in V} \theta_{Y,y} y \lambda_i \right)$$

where $\theta_{i,j}$ $1 \leq i, j \leq m+1$ denote the canonical parameters associated with the supervision sources and the Y and Z is a partition function. Using these parameters we may compute $f_G(Y|\lambda_1, \ldots, \lambda_m)$ for each of the training examples.

Previous Work

(Algorithm 1 comes from [2] and algorithm 2 from [3])



Let Σ denote the covariance matrix of the supervision sources and Y. Writing $O = \{\lambda_1, \ldots, \lambda_M\}$ and $S = \{Y\}$, we have

$$\Sigma = \begin{bmatrix} \Sigma_O & \Sigma_{OS} \\ \Sigma_{OS}^T & \Sigma_S \end{bmatrix}$$

Inverting Σ, we write

$$\Sigma^{-1} = \begin{bmatrix} K_O & K_{OS} \\ K_{OS}^T & K_S \end{bmatrix}$$

To learn G we assume G is sparse and thus Σ^{-1} is graph-structured. We may estimate Σ_O^{-1} empirically.

$$\Sigma_O^{-1} = K_O - zz^T$$

where K_O is sparse and zz^T is low-rank. We may learn the structure of G from K_O and we may learn the accuracies of the sources from z.

Algorithm 1 yields z and K_O and thus the source accuracies and the graph structure. Ratner, et. al. [3] propose a simpler algorithm for estimating z if the graph structure is already known. If E is alrady known we may construct a dependency mask $\Omega = \{(i, j) : (\lambda_i, \lambda_j) \notin E\}$ which we use in algorithm 2:

Algorithm 1: Weak Supervision Structure Learning and Source Estimation Using Robust PCA

Result: $\hat{G} = (V, \hat{E})$, $\hat{L}$
Inputs : Estimate of covariance matrix $\hat{\Sigma}_O$, parameter γ, threshold T
Solve : $(\hat{S}, \hat{L}) = \text{argmin}_{(S,L)} ||L||_* + \gamma ||S||_1$
s.t. $S - L = \hat{\Sigma}_O^{-1}$
$\hat{E} \leftarrow \{(i, j) : i < j, \hat{S}_{i,j} > T\}$

Algorithm 2: Source Estimation for Weak Supervision

Result: $\hat{\mu}$
Inputs : Observed labeling rates $\hat{\mathbb{E}}[O]$ and covariance $\hat{\Sigma}_O$; class balance $\hat{\mathbb{E}}[Y]$ and variance $\hat{\Sigma}_S$; dependency mask Ω
$\hat{z} \leftarrow \text{argmin}_Z ||\hat{\Sigma}_O^{-1} + zz^T||_\Omega$
$\hat{c} \leftarrow \Sigma_S^{-1}(1 + \hat{z}^T \hat{\Sigma}_O \hat{z})$
$\hat{\Sigma}_{OS} \leftarrow \hat{\Sigma}_O \hat{z} / \sqrt{\hat{c}}$
$\hat{\mu} \leftarrow \hat{\Sigma}_{OS} + \hat{\mathbb{E}}[Y]\hat{\mathbb{E}}[O]$

Problem

An incremental algorithm for learning source accuracies and graph structure would allow us to use weak-supervision labeling methods in out-of-core or online settings. For example, an incremental learning algorithm would allow for generating labels on streaming data in real-time. Although algorithm 1 estimates both the source accuracies μ and the dependency structure $\hat{E}$, it requires the entire dataset and cannot be implemented iteratively. Algorithm 2 is much more efficient but does not yield an estimate of the dependency structure of the labeling functions.

Solution

We develop a method which combines algorithms 1 and 2 to estimate both E and μ incrementally.

Method

Given an initial batch of samples we estimate $\hat{K}_O$ and $\hat{z}$ using algorithm 1. From $\hat{K}_O$ we may determine the graph structure and dependency mask $\hat{\Omega}$. Let $\hat{\mu}$ denote the the estimated accuracies of the supervision sources. Given $\hat{\Omega}$ we may estimate z' for a new batch X' of examples using

$$\hat{z}' \leftarrow \text{argmin}_{z'} ||\hat{\Sigma}_O^{-1} + z' z'^T||_\Omega$$

which may be solved by least-squares. From $\hat{z}'$ we estimate the source accuracies on X'

$$\hat{c} \leftarrow \Sigma_S^{-1}(1 + \hat{z}'^T \hat{\Sigma}_O \hat{z}')$$

$$\hat{\Sigma}_{OS} \leftarrow \hat{\Sigma}_O \hat{z}' / \sqrt{\hat{c}}$$

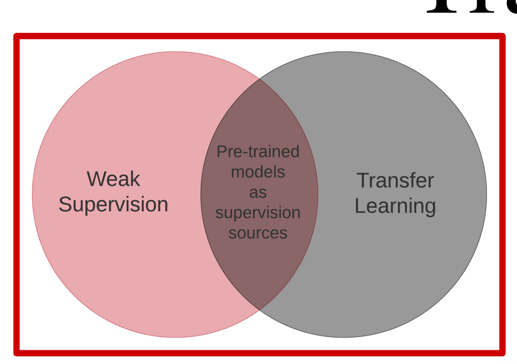
$$\hat{\mu}' \leftarrow \hat{\Sigma}_{OS} + \hat{\mathbb{E}}[Y]\hat{\mathbb{E}}[O]$$

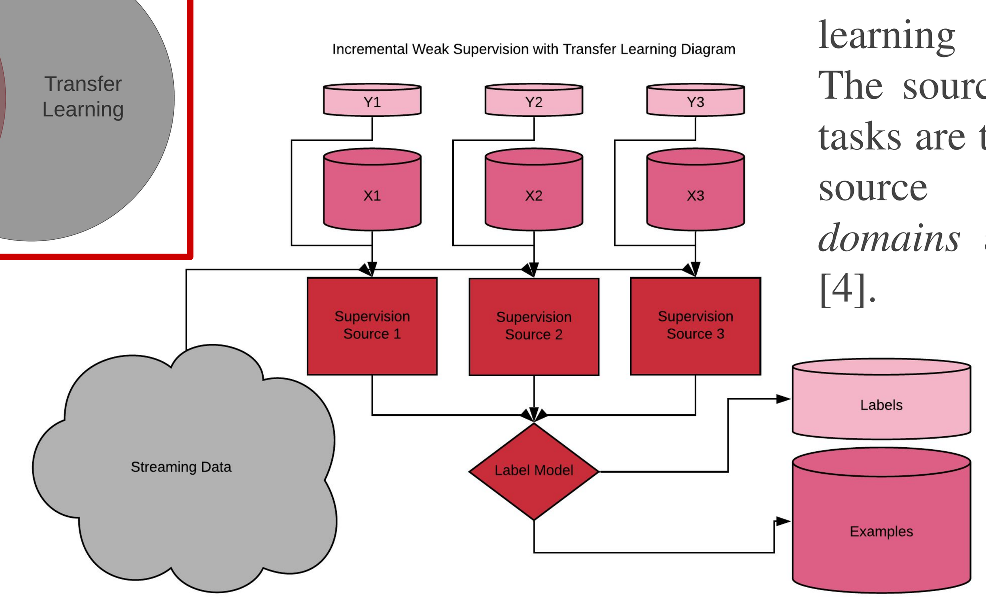
Using α we update $\hat{\mu}$

$$\hat{\mu} = (1 - \alpha)\hat{\mu} + \alpha \hat{\mu}'$$

We thus estimate source accuracies using an exponentially-weighted moving average of the batch-estimates $\hat{\mu}'$.

Transfer Learning





Transductive transfer learning problems: The source and target tasks are the same, but source and target *domains* are different [4].

Tests

We utilize our incremental algorithm in such a scenario. We attempt to label tweets by sentiment received in real-time, as they are broadcast, using the following weak supervision sources:

- Naïve Bayes model trained on a set of movie reviews labeled by sentiment ("positive" or "negative")
- Textblob Pattern Analyzer - A model which uses a lookup table to generate polarity and subjectivity estimates for a text
- Naïve Bayes model trained on a dataset consisting of tweets associated with US airlines and labeled by sentiment ("positive", "neutral", and "negative")

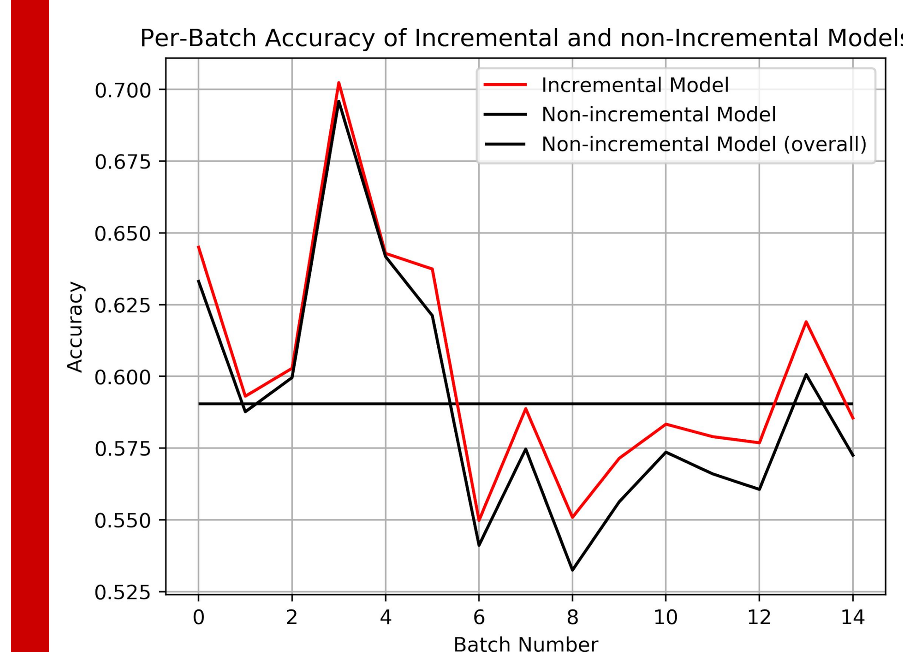
Data

We test our model using tweets concerning a presidential debate labeled by sentiment ("positive", "neutral", or "negative"). We sort the tweets by timestamp, partition them into batches, and stream them to our model to simulate the imagined use case in which the model is used to generate labels in real-time.

We compare the per-batch accuracy of the labels generated by our model with the accuracy of the non-incremental model.

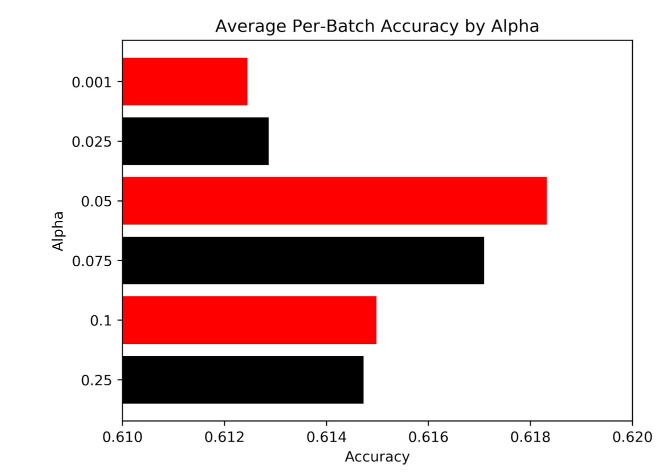
We test different values of α and compare accuracies.

Results



$$\texttt{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

Alpha	0.001	0.01	0.025	0.05	0.1	0.25
Accuracy	0.61245	0.61287	0.61832	0.61709	0.61498	0.61473



References

[1]. Alexander Ratner. Stephen Bach. Paroma Varma. Christopher Ré. (2017) "Weak Supervision: The New Programming Paradigm for Machine Learning". Available: https://hazyresearch.github.io/snorkel/blog/ws_blog_post.html

[2]. Paroma Varma. Frederic Sala. Ann He. Alexander Ratner. Christopher Ré. (2019) "Learning Dependency Structures for Weak Supervision Models". Preprint.

[3]. Alexander Ratner. Braden Hancock. Jared Dunnmon. Frederic Sala. Shreyash Pandey. Christopher Ré. (2018) "Training Complex Models with Multi-Task Weak Supervision". Preprint.

[4]. Sinno Jialin Pan. Qiang Yang (2009) "A Survey on Transfer Learning". *IEEE Transactions on Knowledge and Data Engineering*. Vol 22. Issue 10.