Learning setting

Binary classification: $S = \{(x_i, y_i), i \in [m]\}$ sampled from $D$ over $\mathbb{R}^d \times \{-1, 1\}$. Linear (or kernel) models $h \in \mathcal{H}$. Minimize the empirical risk associated with a surrogate loss $\ell(x)$

$$\arg\min_{h \in \mathcal{H}} \mathbb{E}_S [\ell(h(x))] = \arg\min_{h \in \mathcal{H}} R_{S, \ell}(h)$$

Loss factorization

Define: mean operator $\mu_S \equiv \mathbb{E}_S [y|x]$ Define: “double sample” $S_{2x} = \{(x_i, \sigma), i \in [m], \sigma \in \{\pm 1\}\}$ Assume: Linear model $h$ Assume: Linear-odd loss (LOL)

$$\ell_o(x) = (\ell(x) - \ell(-x))/2 = ax, \text{ for any } a \in \mathbb{R}$$

Factorization:

$$R_{S, \ell}(h) = \frac{1}{2} \mathbb{E}_{S_{2x}} [\ell(h(x)) + \ell(h(-x))] + \mathbb{E}_{S} [\ell_o(h(x))]$$

$$= \frac{1}{2} R_{S_{2x}, \ell}(h) + a h(\mu_S)$$

Theorem Assume $\ell$ is a-LOL and L-Lipschitz. Suppose $\mathbb{R}^d \supset \mathcal{X} = \{x : ||x||_2 \leq X \land \infty \land \mathcal{X} = \{\theta : \||\theta||_2 \leq B < \infty\}$. Let $(c(X, B) = \max_{x \in \mathcal{X}} \mathbb{E}(\ell(x) X) B)$ and $\theta = \arg\min_{\theta \in \mathcal{X}} R_{D, \ell}(\theta)$. Then for any $\delta > 0$, with probability at least $1 - \delta$:

$$c(X, B) \leq \frac{1}{k \log \frac{1}{\delta}} + 2|\mu| \max_{x \in \mathcal{X}} \mathbb{E}_D [||\mu_D - \mu_S||_2]$$

Weakly supervised learning

E.g., noisy labels, partially missing, or multi-instance. We solve it by the same meta-algorithm on 2 steps: (1) estimate $\mu$ and (2) call any standard empirical risk minimizer plugging $\mu$ into the loss computed on the “double sample” $S_{2x}$.

An example of (2): “$\mu$-SGD” is adapted as follow.

Input: $S_{2x}, \mu, \ell$ is a-LOL; $\lambda > 0; T > 0$

$$m' \leftarrow |S_{2x}|$$

$\theta^{0} \leftarrow 0$

For any $t = 1, \ldots, T$:

Pick $i = i^t \in [m']$ uniformly at random

$$\eta^t \leftarrow (\lambda)^{-1}$$

Pick any $v \in \partial \ell(y_i(\theta^t, x_i))$

$$\theta^{t+1} \leftarrow (1 - \eta^t\lambda)\theta^t - \eta^t (v + a\mu/2)$$

$$\theta^{t+1} \leftarrow \min \{\theta^{t+1}, \theta^t + \sqrt{1/\lambda} \}$$

Output: $\theta^{t+1}$

Learning with noisy labels

In a noisy sample $S$ labels are flipped by class-dependent noise $p_+, p_- \in [0, 1/2]$, while feature vectors are the same. By the method of Natarajan et. al.‘13, we obtain an unbiased estimator of the sufficient statistic $\mu$ (instead of $\ell$) as

$$\hat{\mu}_S = \mathbb{E}_S \left[ \frac{y - (1 - p_+ - p_-)}{1 - p_+ - p_-} x \right]$$

This leads to the following bound

$$R_{D, \ell}(\theta) - \inf_{\theta \in \mathcal{X}} R_{D, \ell}(\theta) \leq \left( \frac{\sqrt{2} + 1}{4} \right) \cdot XBL + \frac{1}{m} \log \left( \frac{1}{\delta} \right)$$

Algorithm $\mu$SGD with noisy labels

Input: $S, \ell \in \text{LOL}; \lambda > 0; T > 0$

$S_{2x} = \{(x_i, \sigma), i \in [m], \sigma \in \{\pm 1\}\}$

$\mu_S \leftarrow$ Equation (1)

Output: $\theta$

bonus: LOL are approximately robust

Theorem Assume $\{\theta \in \mathcal{H}: \||\theta||_2 \leq B\}$. Then for every a-LOL

$$R_{S, \ell}(\theta^*) - R_{D, \ell}(\theta^*) \leq 4a|B| \max(p_+, p_-) ||\mu_D - \mu_S||_2$$

Moreover, if $\ell$ is also once differentiable and $\gamma$ strongly convex, then $\|\theta^* - \theta^0\|_2 \leq (2/\gamma) \cdot 4a|B| \max(p_+, p_-) ||\mu_D - \mu_S||_2$.

Experiments

We call our 2-step algorithm with the adaptation of SGD and compare it to standard SGD, with the same parameters. Datasets from UCI are artificially corrupted with asymmetric label noise. With our approach, we can still learn when one of the two noise rates is of about 50% (=one label is random).

Discussions (a sample of)

Proximal methods

$$\theta^{t+1} \leftarrow \arg\min_{\theta} \left( \ell R_{S_{2x}, \ell}(\theta^t + a\mu/2) \right)$$

Covariance operator

$$\pi = \mathbb{E}_{S_1}[y > 0]$$

$$\mu_S = \text{cov}_S [x, y] + (2p_+ - 1) \mathbb{E}_S [x]$$

Ask me about: kernels, learning reductions, regression.