Learning from Corrupted Binary Labels via Class-Probability Estimation

Aditya Menon, Brendan van Rooyen, Cheng Soon Ong, Bob Williamson

Q: Can we learn a good classifier when labels have been corrupted (e.g. label noise, no negative labels)?
A: If corruption rates are unknown, we can do well on balanced error and AUC;
If corruption rates are known, we can do well on a range of other measures (e.g. F-score);
We can estimate corruption rates from outputs of class-probability estimation (e.g. kernel logistic regression).

Classification with Corrupted Binary Labels

Problem: Learning when labels are corrupted in some way.

Class-conditional label noise (CCN learning)

- ✔ Labels flipped with
- ✔ class-dependent
- ✔ probability.
- ✖

Positive and unlabelled data (PU learning)

In lieu of –'ve samples, pool of unlabelled samples.

Three questions:
(1) Don't know corruption parameters → can we still learn?
(2) Know corruption parameters → can we learn more?
(3) Can we estimate the corruption parameters?

Assumed Corruption Model

Mutually contaminated distributions framework (Scott et al, 2013): corrupted class-conditionals are mixtures of original

Clean distribution $D = (P, Q, \pi)$ **(Ideally observed)** ⟶ **Corrupted distribution** $D_{\mathrm{corr}} = (P_{\mathrm{corr}}, Q_{\mathrm{corr}}, \pi_{\mathrm{corr}})$ **(Actually observed)**

$$P_{\mathrm{corr}} = (1-\alpha)\cdot P + \alpha \cdot Q$$
$$Q_{\mathrm{corr}} = \beta \cdot P + (1-\beta)\cdot Q$$

α, β **Corruption rates**

CCN learning: If +'ve (-'ve) labels are flipped w.p. ρ_+ (ρ_-),

$$\alpha = \pi_{\mathrm{corr}}^{-1} \cdot (1-\pi)\cdot \rho_-$$
$$\beta = (1-\pi_{\mathrm{corr}})^{-1} \cdot \pi \cdot \rho_+$$
$$\pi_{\mathrm{corr}} = (1-\rho_+)\cdot\pi + \rho_- \cdot (1-\pi)$$

PU learning: Since all observed +'ves are actually +'ve,

$$\alpha = 0$$
$$\beta = \pi$$
π_{corr} = arbitrary

Balanced Error and AUC are "Corruption-Immune"

Balanced Error (BER) of a classifier *f*: $(\mathrm{FPR}(f) + \mathrm{FNR}(f))/2$.
- favoured over 0-1 error under class imbalance

Fact: Clean and corrupted BER satisfy:

$$\mathrm{BER}^{D_{\mathrm{corr}}}(f) = (1-\alpha-\beta)\cdot \mathrm{BER}^{D}(f) + \frac{\alpha+\beta}{2}.$$

⇒ can minimise BER as-is on corrupted data
⇒ does not require knowledge of corruption parameters!
⇒ can obtain regret bound for strongly proper composite loss minimisation

Similarly, for area under the ROC curve (AUC) of scorer *s*:

$$\mathrm{AUC}^{D_{\mathrm{corr}}}(s) = (1-\alpha-\beta)\cdot \mathrm{AUC}^{D}(s) + \frac{\alpha+\beta}{2}$$

⇒ similar regret bound as for BER

But what about other performance measures?

Structure of Corrupted Class-Probabilities

For many measures, optimal to threshold (clean) class-probabilities, η. In general, the corrupted class-probabilities η_{corr} satisfy:

$$\eta_{\mathrm{corr}}(x) = \phi_{\alpha,\beta,\pi}(\eta(x))$$

where $\phi_{\alpha,\beta,\pi}$ is monotone for fixed α, β, π .

Know α, β, π → can classify on clean distribution:
- find optimal threshold on corrupted distribution, or
- find equivalent corrupted risk

Bad news: Beyond BER, we need to know α, β, π
- Only (non-trivial) measure whose:
  - corrupted threshold independent of α, β, π
  - corrupted risk = affine transform of clean risk
    - Equal FPR/FNR → eigenvector of corruption transform

Good news: We can estimate α, β, π from η_{corr}!

Estimating Corruption Parameters

Suppose *D* satisfies: $\inf_{x \in \mathcal{X}} \eta(x) = 0$ and $\sup_{x \in \mathcal{X}} \eta(x) = 1$

i.e., exist "deterministically +'ve and –'ve instances.

Then, if $\eta_{\min} = \inf_{x \in \mathcal{X}} \eta_{\mathrm{corr}}(x)$ and $\eta_{\max} = \sup_{x \in \mathcal{X}} \eta_{\mathrm{corr}}(x)$,

$$\alpha = \frac{\eta_{\min}\cdot(\eta_{\max}-\pi_{\mathrm{corr}})}{\pi_{\mathrm{corr}}\cdot(\eta_{\max}-\eta_{\min})} \qquad \beta = \frac{(1-\eta_{\max})\cdot(\pi_{\mathrm{corr}}-\eta_{\min})}{(1-\pi_{\mathrm{corr}})\cdot(\eta_{\max}-\eta_{\min})}.$$

Estimate corruption rates from class-probabilities!

CCN learning:

$$\rho_+ = 1 - \eta_{\max}$$
$$\rho_- = \eta_{\min}$$
$$\pi = \frac{\pi_{\mathrm{corr}} - \eta_{\min}}{\eta_{\max} - \eta_{\min}}$$

PU learning:

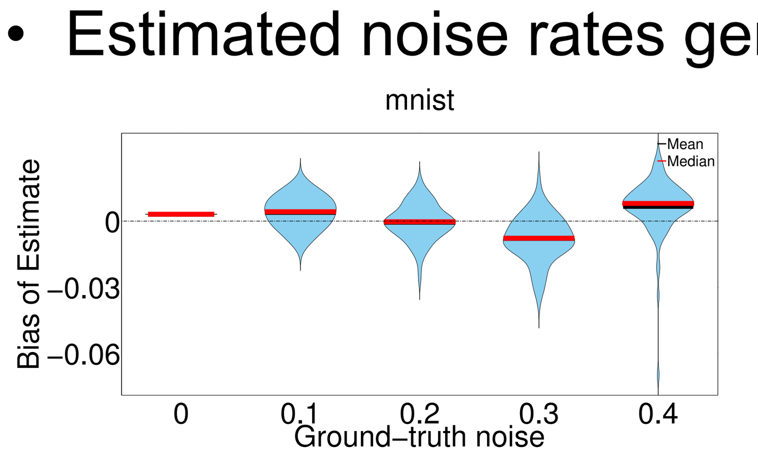
$$\pi = \frac{\pi_{\mathrm{corr}}}{1-\pi_{\mathrm{corr}}} \cdot \frac{1-\eta_{\max}}{\eta_{\max}}.$$

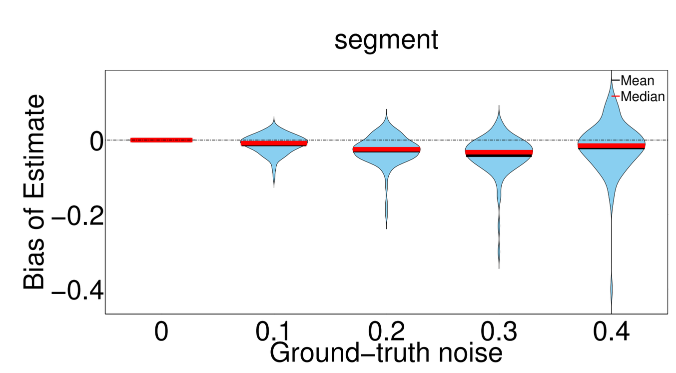
Experimental Validation

- Inject label noise of varying rates to UCI datasets
- Estimate noise rates via a neural network, since

$$\eta(x) = \sigma(\langle w, x\rangle) \implies \eta_{\mathrm{corr}}(x) = a \cdot \sigma(\langle w, x\rangle) + b$$

- Estimated noise rates generally reliable:





- Classification w/ noise estimates ~ w/ oracle noise
- Observe low degradation in both BER and AUC