Learning with Noisy Labels: Variable Selection and Misclassification Probability Modeling

Hui GUO

Department of Computer Science
Western University, Canada

Joint work with Grace Y. Yi

Problem Formulation

Classification problem with label noise:

- Input vector: $X \in \mathbb{R}^p$
  True label: $Y \in \{0, 1\}$
  Noisy label: $Y^* \in \{0, 1\}$
- Main study (noisy) data: $\mathcal{D}_{\mathsf{M}} \triangleq \{\{X_i, Y_i^*\} : i \in \mathcal{M}\}$ with size n
  Validation data: $\mathcal{D}_{\mathsf{V}} \triangleq \{\{X_i, Y_i, Y_i^*\} : i \in \mathcal{V}\}$ with size n_v
  - e.g., diagnoses using high-precision tools
- **Goal:** Learning a valid classifier using $\mathcal{D}_{\mathsf{M}}$ and $\mathcal{D}_{\mathsf{V}}$ to accurately predict the true label for future inputs

Observed Log-Likelihood Function

$$\ell^* = \log\left\{ \prod_{i\in\mathcal{V}} \underbrace{f(y_i|x_i)f(y_i^*|y_i,x_i)}_{\text{validation data}} \cdot \prod_{i\in\mathcal{M}} \underbrace{f(y_i^*|x_i)}_{\text{main study data}} \right\}$$

- $\mu(x) \triangleq P(Y=1|X=x)$ - conditional mean of the true label
  $\mu^*(x) \triangleq P(Y^*=1|X=x)$ - conditional mean of the noisy label

Misclassification probabilities:
$\gamma_{01}(x) = P(Y^*=1|Y=0,X=x)$
$\gamma_{10}(x) = P(Y^*=0|Y=1,X=x)$

✓ Parametric Modeling
✓ Semiparametric Modeling

$$\mu^*(x) = \gamma_{01}(x) + \{1-\gamma_{01}(x)-\gamma_{10}(x)\}\,\mu(x)$$

✓ Variable Selection

Regression model:
$g(\mu(x)) = \beta_0^* + x^T\beta$

- $g(\cdot)$: link function
- $\bar{\beta} = (\beta_0^*, \beta^T)^T$: parameters with true value $\bar{\beta}_0 = (\beta_{00}^*, \beta_0^T)^T$
- $\beta_0 = (\beta_{0I}^T, \beta_{0II}^T)^T$: Non-zero and zero elements of β_0

$\Rightarrow$ Express ℓ^* in terms of μ, γ_{01}, and γ_{10}:

$$\ell^* = \log\left\{ \prod_{i\in\mathcal{V}} \left[\mu(x_i)\{1-\gamma_{10}(x_i)\}^{y_i^*}\gamma_{10}(x_i)^{1-y_i^*}\right]^{y_i} \left[\{1-\mu(x_i)\}\gamma_{01}(x_i)^{y_i^*}\{1-\gamma_{01}(x_i)\}^{1-y_i^*}\right]^{1-y_i} \cdot \prod_{i\in\mathcal{M}} \mu^*(x_i)^{y_i^*}\{1-\mu^*(x_i)\}^{1-y_i^*} \right\}$$

Misclassification Probability Modeling

- **Parametric method:**

$$\text{logit } \gamma_{01}(x) = g_0(x;\nu);\ \text{logit } \gamma_{10}(x) = g_1(x;\nu)$$

  - $\theta = (\bar{\beta}^{\mathsf{T}}, \nu^{\mathsf{T}})^{\mathsf{T}}$: all involved parameters, with true value $\theta_0 = (\bar{\beta}_0^{\mathsf{T}}, \nu_0^{\mathsf{T}})^{\mathsf{T}}$
  - Log-likelihood function: $\ell^* \Rightarrow \ell^*(\theta)$

- **Semiparametric method:**

$$\widehat{\gamma}_{10}(x) = \frac{\sum_{i\in\mathcal{V}} y_i(1-y_i^*)\widetilde{K}_{i,x}}{\sum_{i\in\mathcal{V}} y_i \widetilde{K}_{i,x}};\quad \widehat{\gamma}_{01}(x) = \frac{\sum_{i\in\mathcal{V}}(1-y_i)y_i^* \widetilde{K}_{i,x}}{\sum_{i\in\mathcal{V}}(1-y_i)\widetilde{K}_{i,x}}.$$

  - $\widetilde{K}_{i,x}$: kernel estimator

$$\widetilde{K}_{i,x} = \underbrace{h^{-p_1} K\{(x_i^{\mathrm{C}} - x^{\mathrm{C}})/h\}}_{\text{continuous components}} \cdot \underbrace{\omega^{\sum_{t=1}^{p_2} \mathbb{I}(x_{it}^{\mathrm{D}} \neq x_t^{\mathrm{D}})}}_{\text{discrete components}}$$

  - Log-likelihood function: $\ell^* \Rightarrow \widehat{\ell}^*(\bar{\beta})$

Penalized Log-Likelihood Function Maximization

Penalized Log-Likelihood Function:

$$Q^* = \ell^* - n\sum_{j=1}^{p} \rho_{\lambda_n}(|\beta_j|)$$

ℓ^*: $\ell^*(\theta)$ or $\widehat{\ell^*}(\bar{\beta})$

ρ_{λ_n}: penalty function for variable selection

λ_n: nonnegative tuning parameter

e.g., SCAD, MCP

Theorem (Consistency).

Under regularity conditions, there exists a local maximizer of Q^* such that

$$\|\widehat{\bar{\beta}} - \bar{\beta}_0\|_2 = O_p(n^{-1/2} + a_n),$$

where $a_n = \max\{|\rho'_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0,\ j = 1, ..., p\}$.

Penalized Log-Likelihood Function Maximization

Theorem (Oracle Property).

Under regularity conditions, any $\sqrt{n}$-consistent estimator obtained from maximizing Q^* have the following properties: as $n \to \infty$,

(i) (Sparsity) with probability tending to 1, $\widehat{\beta}_{\mathrm{II}} = 0$,

(ii) (Asymptotic normality)

- For the parametric method,

$$\sqrt{n}(I_{11}^* + \Sigma)\left\{(\widehat{\bar{\beta}}_{\mathrm{I}} - \bar{\beta}_{\mathrm{I0}}) + (I_{11}^* + \Sigma)^{-1} b\right\} \xrightarrow{d} \mathcal{N}(0, I_{11}^*).$$

- For the semiparametric method,

$$\sqrt{n}\left\{I_{11} + \Sigma\right\}\left[(\widehat{\bar{\beta}}_{\mathrm{I}} - \bar{\beta}_{\mathrm{I0}}) + \left\{I_{11} + \Sigma\right\}^{-1} b\right] \xrightarrow{d} \mathcal{N}(0, I_{11} + \Sigma_{\mathrm{sp}}).$$

Thank You!