Learning with Class-Conditional Random Label Noise

References

Foundations of Machine Learning

Learning with Noisy Labels. NIPS 2013

Convexity, Classification, and Risk Bounds

Statistical behavior and consistency of classification methods based on convex risk minimization

Rademacher and Gaussian Complexities: Risk Bounds and Structural Results

Large-scale Multi-label Learning with Missing Labels. ICML 2014

Outline

- Preliminary
  - Empirical Risk Minimization (ERM)
  - Rademacher Complexity
  - Generalization Bound
- Learning with Noisy Label
  - Class-Conditional random label Noise (CCN)
- CCN in Multi-Label Learning

Empirical Risk Minimization (ERM) Framework

In machine learning, our goal is to find a predictor minimize the expected loss of f given below:

$$L(f(\cdot)) = \mathbf{E}_{X,Y}\ell\big(f(X), Y\big) \longrightarrow L(f(\cdot)) = \mathbf{E}_{X,Y} I\big(f(X), Y\big)$$

Risk

The expectation with respective to the true underlying distribution D

?

Binary classification: $y \in \{\pm 1\}$

0-1 loss $$I(f(x), y) = \begin{cases} 1, & \text{if } f(x)y < 0, \\ 1, & \text{if } f(x) = 0 \text{ and } y = -1 \\ 0, & \text{otherwise.} \end{cases}$$

Empirical Risk Minimization (ERM) Framework

Given a set of training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ independently drawn from D, the minimization of the empirical risk given below can be regarded as stochastic approximation of risk:

$$\frac{1}{n}\sum_{i=1}^{n} I\big(f(X_i), Y_i\big)$$

Nonconvexity → NP-hard

Idea: minimize a convex upper bound of the 0-1 loss function I

$$\frac{1}{n}\sum_{i=1}^{n} \phi\big(f(X_i)Y_i\big)$$

- Least squares: $\phi(v) = (1-v)^2$.
- Modified least squares: $\phi(v) = \max(1-v, 0)^2$.
- SVM: $\phi(v) = \max(1-v, 0)$.
- Exponential: $\phi(v) = \exp(-v)$.
- Logistic regression: $\phi(v) = \ln(1+\exp(-v))$.

Empirical Risk Minimization (ERM) Framework

Risk Bound

$$L(f(\cdot)) \leq \hat{L}(f(\cdot)) + \epsilon$$

A penalty term that is large for more complex function class

Example

Theorem 1 *Let F be a class of $\{\pm 1\}$-valued functions defined on a set $\mathcal{X}$. Let P be a probability distribution on $\mathcal{X} \times \{\pm 1\}$, and suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ and (X, Y) are chosen independently according to P. Then, there is an absolute constant c such that for any integer n, with probability at least $1 - \delta$ over samples of length n, every f in F satisfies*

$$P(Y \neq f(X)) \leq \hat{P}_n(Y \neq f(X)) + c\sqrt{\frac{\mathrm{VCdim}(F)}{n}},$$

data-independent

where $\mathrm{VCdim}(F)$ denotes the Vapnik-Chervonenkis dimension of F,

$$\hat{P}_n(S) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_S(X_i, Y_i),$$

Rademacher Complexity

Definition 3.1 (Empirical Rademacher complexity) *Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[a, b]$ and $S = (z_1, \ldots, z_m)$ a fixed sample of size m with elements in $\mathcal{Z}$. Then, the* empirical Rademacher complexity *of $\mathcal{G}$ with respect to the sample S is defined as:*

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i)\right], \tag{3.1}$$

where $\boldsymbol{\sigma} = (\sigma_i)_{i \in [m]}$, called Rademacher variables are independent uniform random variables taking values in $\{-1, +1\}$; $\forall i \in [m], \mathbb{P}(\sigma_i = -1) = \mathbb{P}(\sigma_i = +1) = 1/2$.

Definition 3.2 (Rademacher complexity) *Let $\mathcal{D}$ denote the distribution according to which samples are drawn. For any integer $m \geq 1$, the* Rademacher complexity *of $\mathcal{G}$ is the expectation of the empirical Rademacher complexity over all samples of size m drawn according to $\mathcal{D}$:*

$$\mathfrak{R}_m(\mathcal{G}) = \mathop{\mathbb{E}}_{S \sim \mathcal{D}^m}[\widehat{\mathfrak{R}}_S(\mathcal{G})]. \tag{3.2}$$

Generalization Bound

Theorem 3.3 *Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[0,1]$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size m, each of the following holds for all $g \in \mathcal{G}$:*

data-dependent

$g(z) = \ell(f(z), y)$

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathfrak{R}_m(\mathcal{G}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{3.3}$$

$$\textit{and} \quad \mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}\,. \tag{3.4}$$

Proof sketch.

$\widehat{\mathbb{E}}_S[g] = \frac{1}{m}\sum_{i=1}^{m} g(z_i)$

$\Phi(S) = \sup_{g\in\mathcal{G}}\left(\mathbb{E}[g] - \widehat{\mathbb{E}}_S[g]\right)$ McDiarmid's inequality

Foundations of Machine Learning. C3

Learning with Noisy Label

Symmetric label noise

$$\tilde{Y} = \begin{cases} Y & with\ probability\ \rho \\ -Y & with\ probability\ (1-\rho) \end{cases}$$

Class-conditional label noise (CCN)

$$P(\tilde{Y} = -1|Y = +1) = \rho_{+1}, P(\tilde{Y} = +1|Y = -1) = \rho_{-1}, \text{ and } \rho_{+1} + \rho_{-1} < 1 \text{ (known)}$$

Positive-Unlabeled (PU) Learning

$$\mathcal{S} = \mathcal{S}_{+1} \cup \mathcal{S}_{\text{unl}} \longrightarrow \underline{\mathcal{S}} = \mathcal{S}_{+1} \cup \mathcal{S}_{-1}$$

$$P(\tilde{Y} = -1|Y = +1) = \rho$$

$$P(\tilde{Y} = +1|Y = -1) = 0$$

$$\mathcal{S}_{+1} = \{(\mathbf{x}_i, 1)\}_{i=1}^{m} \qquad \mathcal{S}_{\text{unl}} = \{\mathbf{x}_i\}_{i=m+1}^{n} \qquad \mathcal{S}_{-1} = \{(\mathbf{x}_i, -1)\}_{i=m+1}^{n}$$

Learning with Noisy Label

Problem setup

Risk: $R_D(f) = \mathbb{E}_{(X,Y)\sim D}\left[1_{\{\text{sign}(f(X))\neq Y\}}\right]$

Bayes optimal function: $f^\star(x) = \text{sign}(\eta(x) - 1/2)$ where $\eta(x) = P(Y = 1|x)$

Bayes risk: $R^* = R_D(f^*)$

Loss function: $\ell(t, y)$

Modified loss function: $\tilde{\ell}(t, \tilde{y})$ for noisy label

1. Empirical $\tilde{\ell}$-risk on the observed sample: $\widehat{R}_{\tilde{\ell}}(f) := \frac{1}{n}\sum_{i=1}^{n} \tilde{\ell}(f(X_i), \tilde{Y}_i)$.
2. As n grows, we expect $\widehat{R}_{\tilde{\ell}}(f)$ to be close to the $\tilde{\ell}$-risk under the noisy distribution D_ρ:

$$R_{\tilde{\ell},D_\rho}(f) := \mathbb{E}_{(X,\tilde{Y})\sim D_\rho}\left[\tilde{\ell}(f(X), \tilde{Y})\right] .$$

3. ℓ-risk under the "clean" distribution D: $R_{\ell,D}(f) := \mathbb{E}_{(X,Y)\sim D}[\ell(f(X), Y)]$.

Learning with Noisy Label

Unbiased Estimator: $\tilde{\ell}(t,y) := \dfrac{(1-\rho_{-y})\,\ell(t,y) - \rho_y\,\ell(t,-y)}{1-\rho_{+1}-\rho_{-1}}$ → $\mathbb{E}_{\tilde{y}}\left[\tilde{\ell}(t,\tilde{y})\right] = \ell(t,y)$

Proof. Considering that y = +1 and y = -1

$$(1-\rho_{+1})\tilde{\ell}(t,+1) + \rho_{+1}\tilde{\ell}(t,-1) = \ell(t,+1)$$
$$(1-\rho_{-1})\tilde{\ell}(t,-1) + \rho_{-1}\tilde{\ell}(t,+1) = \ell(t,-1)$$

→

$$\tilde{\ell}(t,+1) = \frac{(1-\rho_{-1})\ell(t,+1) - \rho_{+1}\ell(t,-1)}{1-\rho_{+1}-\rho_{-1}}$$
$$\tilde{\ell}(t,-1) = \frac{(1-\rho_{+1})\ell(t,-1) - \rho_{-1}\ell(t,+1)}{1-\rho_{+1}-\rho_{-1}}$$

Our goal is to learn a good predictor in the presence of label noise by minimizing

$$\hat{f} \leftarrow \underset{f\in\mathcal{F}}{\operatorname{argmin}}\ \widehat{R}_{\tilde{\ell}}(f)$$

Unbiasedness of $\tilde{\ell}$ ⟶ the above term converge to $R_{\ell,D}(f)$

Use Rademacher complexity to give a performance guarantee for this procedure

Learning with Noisy Label

Lemma 2. *Let $\ell(t, y)$ be L-Lipschitz in t (for every y). Then, with probability at least $1 - \delta$,*

$$\max_{f \in \mathcal{F}} |\widehat{R}_{\tilde{\ell}}(f) - R_{\tilde{\ell}, D_\rho}(f)| \leq 2L_\rho \mathfrak{R}(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

where $\mathfrak{R}(\mathcal{F}) := \mathbb{E}_{X_i, \epsilon_i} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \epsilon_i f(X_i)\right]$

Proof.

$$\max_{f \in \mathcal{F}} |\widehat{R}_{\tilde{\ell}}(f) - R_{\tilde{\ell}, D_\rho}(f)| \leq 2 \cdot \mathfrak{R}(\tilde{\ell} \circ \mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

where $\mathfrak{R}(\tilde{\ell} \circ \mathcal{F}) := \mathbb{E}_{X_i, \tilde{Y}_i, \epsilon_i} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \tilde{\ell}(f(X_i), \tilde{Y}_i)\right]$

Assume ℓ is L-Lipschitz then $\tilde{\ell}$ is L_ρ-Lipschitz

Theorem 3.3 *Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[0, 1]$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample S of size m, each of the following holds for all $g \in \mathcal{G}$:*

$$\mathbb{E}[g(z)] \leq \frac{1}{m} \sum_{i=1}^{m} g(z_i) + 2\mathfrak{R}_m(\mathcal{G}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \qquad (3.3)$$

$$\mathfrak{R}(\tilde{\ell} \circ \mathcal{F}) \leq L_\rho \cdot \mathfrak{R}(\mathcal{F})$$

Learning with Noisy Label

Theorem 3 (Main Result 1). *With probability at least* $1-\delta$,

$$R_{\ell,D}(\hat{f}) \leq \min_{f\in\mathcal{F}} R_{\ell,D}(f) + 4L_\rho \mathfrak{R}(\mathcal{F}) + 2\sqrt{\frac{\log(1/\delta)}{2n}}$$

Proof.

Let f^* be the minimizer of $R_{\ell,D}(\cdot)$

$$\hat{f} \leftarrow \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{R}_{\tilde{\ell}}(f)$$

$$\begin{aligned}
&R_{\ell,D}(\hat{f}) - R_{\ell,D}(f^\star) \\
&= R_{\tilde{\ell},D_\rho}(\hat{f}) - R_{\tilde{\ell},D_\rho}(f^\star)
\end{aligned}$$

Due to unbiasedness of $\tilde{\ell}$

$$\begin{aligned}
&= \widehat{R}_{\tilde{\ell}}(\hat{f}) - \widehat{R}_{\tilde{\ell}}(f^\star) + (R_{\tilde{\ell},D_\rho}(\hat{f}) - \widehat{R}_{\tilde{\ell}}(\hat{f})) \\
&\quad + (\widehat{R}_{\tilde{\ell}}(f^\star) - R_{\tilde{\ell},D_\rho}(f^\star)) \\
&\leq 0 + 2\max_{f\in\mathcal{F}} |\widehat{R}_{\tilde{\ell}}(f) - R_{\tilde{\ell},D_\rho}(f)|
\end{aligned}$$

$$\widehat{R}_{\tilde{\ell}}(\hat{f}) - \widehat{R}_{\tilde{\ell}}(f^\star) \leq 0$$

$$(R_{\tilde{\ell},D_\rho}(\hat{f}) - \widehat{R}_{\tilde{\ell}}(\hat{f})) \leq \max_{f\in\mathcal{F}} |\widehat{R}_{\tilde{\ell}}(f) - R_{\tilde{\ell},D_\rho}(f)|$$

$$\max_{f\in\mathcal{F}} |\widehat{R}_{\tilde{\ell}}(f) - R_{\tilde{\ell},D_\rho}(f)| \leq 2L_\rho \mathfrak{R}(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

CCN in Multi-Label Learning

ERM Framework

$$\ell(\boldsymbol{y}, f(\boldsymbol{x}; Z)) = \sum_{j=1}^{L} \ell(\boldsymbol{y}^j, f^j(\boldsymbol{x}; Z)) \quad \text{where } \mathcal{Z} = \{Z \in \mathbb{R}^{d\times q} : rank(Z) \leq \varepsilon\}$$

Class-Conditional random label Noise

Uniform $P(y^j = -1 | y^j = +1) = \rho_{+1}$ $P(y^j = +1 | y^j = -1) = \rho_{-1}$ $\forall j \in [q]$

Non-uniform $P(y^j = -1 | y^j = +1) = \rho_{+1}^j$ $P(y^j = +1 | y^j = -1) = \rho_{-1}^j$ $\forall j \in [q]$

Partial Multi-Label Learning:

$$P(y^j = -1 | y^j = +1) = 0 \qquad P(y^j = +1 | y^j = -1) = \rho_{-1}$$

Weak Label Learning

$$P(y^j = -1 | y^j = +1) = \rho_{+1} \qquad P(y^j = +1 | y^j = -1) = 0$$

Thanks