Learning from Complementary Labels

Takashi Ishida[1,2,3] **Gang Niu**[2,3] **Weihua Hu**[2,3] **Masashi Sugiyama**[3,2]

[1] Sumitomo Mitsui Asset Management, Tokyo, Japan

[2] The University of Tokyo, Tokyo, Japan

[3] RIKEN, Tokyo, Japan

{ishida@ms., gang@ms., hu@ms., sugi@}k.u-tokyo.ac.jp

ICML 2019

Ordinary Multi-Class Classification

The goal of ordinary multi-class classification is to learn a classifier $f(\boldsymbol{x}) : \mathbb{R}^d \to \{1, \ldots, K\}$ that minimizes the risk with $\mathcal{L}(f(\boldsymbol{x}), y)$

$$R(f) = \mathbb{E}_{p(\boldsymbol{x},y)}\left[\mathcal{L}\left(f(\boldsymbol{x}), y\right)\right]$$

$$f(\boldsymbol{x}) = \underset{y \in \{1,\ldots,K\}}{\arg\max}\ g_y(\boldsymbol{x})$$

Binary classifier

$$g_y(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$$

One vs All (OVA) and One vs One (Pairwise Comparison, PC)

$\ell(z) : \mathbb{R} \to \mathbb{R}$ a large loss for a small z

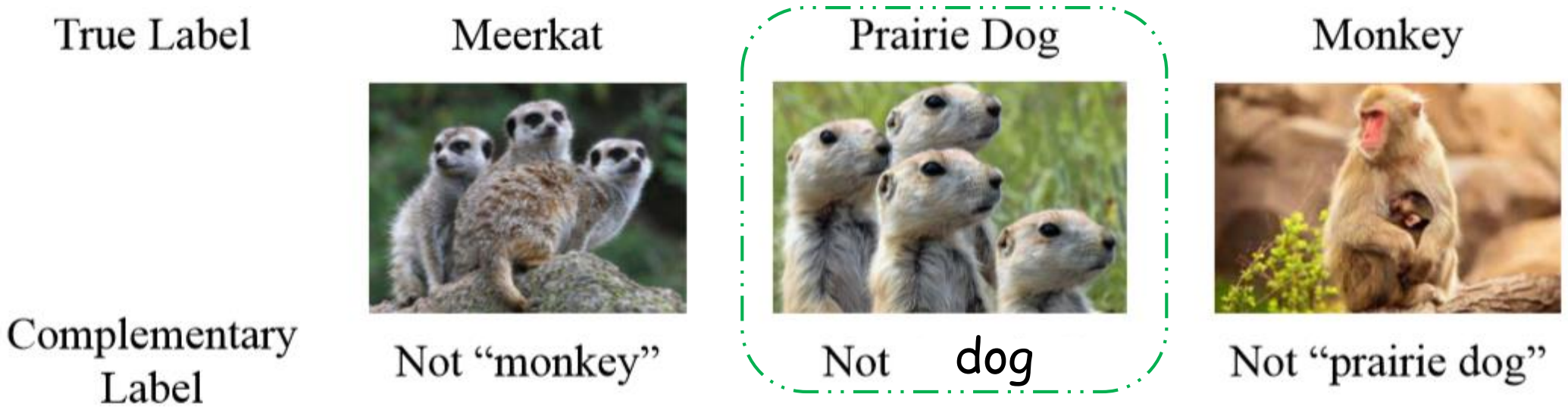
$$\mathcal{L}_{\mathrm{OVA}}(f(\boldsymbol{x}), y) = \ell\left(g_y(\boldsymbol{x})\right) + \frac{1}{K-1}\sum_{y' \neq y} \ell\left(-g_{y'}(\boldsymbol{x})\right)$$

$$\mathcal{L}_{\mathrm{PC}}\left(f(\boldsymbol{x}), y\right) = \sum_{y' \neq y} \ell\left(g_y(\boldsymbol{x}) - g_{y'}(\boldsymbol{x})\right).$$

$\frac{K(K-1)}{2}$ ~~K~~ classifiers

Ordinary Multi-Class Classification

- Critical Issue: training requires a large set of labeled data → labeling cost is expensive
- Existing Solutions
  - ✓ Active learning
  - ✓ Weakly supervised learning, e.g., partial label learning…
- Solution in the paper: complementary label learning



Complementary Label Learning

- Problem Formulation

$x \longleftrightarrow \cancel{y}\ \bar{y}$ $\quad \{(\boldsymbol{x}_i, \bar{y}_i)\}_{i=1}^{n} \xrightarrow{?}$ learn a multi-class classifier

- A basic assumption

$$\bar{p}(\boldsymbol{x}, \bar{y}) = \frac{1}{K-1} \sum_{y \neq \bar{y}} p(\boldsymbol{x}, y)$$

All $p(x, y)$ for $y \neq \bar{y}$ equally contribute to $\bar{p}(x, \bar{y})$

- The complementary loss $\bar{\mathcal{L}}(f(\boldsymbol{x}), \bar{y})$

Complementary Label Learning

- Unbiased Estimation of the Risk

$$R(f) = \mathbb{E}_{p(\boldsymbol{x},y)}\big[\mathcal{L}\big(f(\boldsymbol{x}), y\big)\big]$$

Theorem 1. *The classification risk (1) can be expressed as*

$$R(f) = (K-1)\mathbb{E}_{\overline{p}(\boldsymbol{x},\overline{y})}\big[\overline{\mathcal{L}}\big(f(\boldsymbol{x}), \overline{y}\big)\big] - M_1 + M_2,$$

if there exist constants $M_1, M_2 \geq 0$ *such that for all* $\boldsymbol{x}$ *and* y*, the complementary loss satisfies*

$$\sum_{\overline{y}=1}^{K} \overline{\mathcal{L}}\big(f(\boldsymbol{x}), \overline{y}\big) = M_1 \quad \textit{and} \quad \overline{\mathcal{L}}\big(f(\boldsymbol{x}), y\big) + \mathcal{L}\big(f(\boldsymbol{x}), y\big) = M_2.$$

Proof.

$$(K-1)\mathbb{E}_{\overline{p}(\boldsymbol{x},\overline{y})}[\overline{\mathcal{L}}(f(\boldsymbol{x}), \overline{y})] = (K-1)\int \sum_{\overline{y}=1}^{K} \overline{\mathcal{L}}(f(\boldsymbol{x}), \overline{y})\,\overline{p}(\boldsymbol{x}, \overline{y})\,\mathrm{d}\boldsymbol{x}$$

Assumption:

$$\overline{p}(\boldsymbol{x}, \overline{y}) = \frac{1}{K-1}\sum_{y \neq \overline{y}} p(\boldsymbol{x}, y)$$

$$= (K-1)\int \sum_{\overline{y}=1}^{K} \overline{\mathcal{L}}(f(\boldsymbol{x}), \overline{y}) \left(\frac{1}{K-1}\sum_{y \neq \overline{y}} p(\boldsymbol{x}, y)\right) \mathrm{d}\boldsymbol{x} = \int \sum_{y=1}^{K}\sum_{\overline{y} \neq y} \overline{\mathcal{L}}(f(\boldsymbol{x}), \overline{y})\,p(\boldsymbol{x}, y)\,\mathrm{d}\boldsymbol{x}$$

$$= \mathbb{E}_{p(\boldsymbol{x},y)}\left[\sum_{\overline{y} \neq y} \overline{\mathcal{L}}(f(\boldsymbol{x}), \overline{y})\right] = \mathbb{E}_{p(\boldsymbol{x},y)}[M_1 - \overline{\mathcal{L}}(f(\boldsymbol{x}), y)] = M_1 - \mathbb{E}_{p(\boldsymbol{x},y)}[\overline{\mathcal{L}}(f(\boldsymbol{x}), y)]$$

Complementary Label Learning

$$(K-1)\mathbb{E}_{\overline{p}(\boldsymbol{x},\overline{y})}[\overline{\mathcal{L}}(f(\boldsymbol{x}),\overline{y})] = M_1 - \mathbb{E}_{p(\boldsymbol{x},y)}[\overline{\mathcal{L}}(f(\boldsymbol{x}),y)] \qquad \overline{\mathcal{L}}\big(f(\boldsymbol{x}),y\big) + \mathcal{L}\big(f(\boldsymbol{x}),y\big) = M_2$$

$$\begin{aligned}(K-1)\mathbb{E}_{\overline{p}(\boldsymbol{x},\overline{y})}[\overline{\mathcal{L}}(f(\boldsymbol{x}),\overline{y})] - \mathbb{E}_{p(\boldsymbol{x},y)}[\mathcal{L}(f(\boldsymbol{x}),y)] &= M_1 - \mathbb{E}_{p(\boldsymbol{x},y)}[\overline{\mathcal{L}}(f(\boldsymbol{x}),y) + \mathcal{L}(f(\boldsymbol{x}),y)] \\ &= M_1 - \mathbb{E}_{p(\boldsymbol{x},y)}[M_2] \\ &= M_1 - M_2,\end{aligned}$$

$$\Rightarrow \quad R(f) = (K-1)\mathbb{E}_{\overline{p}(\boldsymbol{x},\overline{y})}\big[\overline{\mathcal{L}}\big(f(\boldsymbol{x}),\overline{y}\big)\big] - M_1 + M_2$$

- Two Conditions

$$\sum_{\overline{y}=1}^{K} \overline{\mathcal{L}}\big(f(\boldsymbol{x}),\overline{y}\big) = M_1$$

A multi-class version of a symmetric constraint

$$\ell(z) + \ell(-z) = C$$

$$\overline{\mathcal{L}}\big(f(\boldsymbol{x}),y\big) + \mathcal{L}\big(f(\boldsymbol{x}),y\big) = M_2$$

pattern x belongs to class y — correct

⇓

pattern x does not belongs to class y — incorrect

Complementary Label Learning

- Empirical risk: $\widehat{R}(f) = \frac{K-1}{n}\sum_{i=1}^{n}\overline{\mathcal{L}}\big(f(\boldsymbol{x}_i), \overline{y}_i\big) - M_1 + M_2$

$\ell(z) : \mathbb{R} \to \mathbb{R}$ a large loss for a small z

- Complementary loss function:

$$\overline{\mathcal{L}}_{\mathrm{OVA}}(f(\boldsymbol{x}), \overline{y}) = \frac{1}{K-1}\sum_{y \neq \overline{y}} \ell\big(g_y(\boldsymbol{x})\big) + \ell\big(-g_{\overline{y}}(\boldsymbol{x})\big)$$

$$\overline{\mathcal{L}}_{\mathrm{PC}}\big(f(\boldsymbol{x}), \overline{y}\big) = \sum_{y \neq \overline{y}} \ell\big(g_y(\boldsymbol{x}) - g_{\overline{y}}(\boldsymbol{x})\big).$$

Theorem 2. *If binary loss $\ell(z)$ satisfies*

$$\ell(z) + \ell(-z) = 1, \tag{11}$$

then $\overline{\mathcal{L}}_{\mathrm{OVA}}$ satisfies conditions (7) with $M_1 = K$ and $M_2 = 2$, and $\overline{\mathcal{L}}_{\mathrm{PC}}$ satisfies conditions (7) with $M_1 = K(K-1)/2$ and $M_2 = K-1$.

$$\text{Zero-one loss: } \ell_{0\text{-}1}(z) = \begin{cases} 0 & \text{if } z > 0, \\ 1 & \text{if } z \leq 0, \end{cases}$$

$$\text{Sigmoid loss: } \ell_{\mathrm{S}}(z) = \frac{1}{1+e^{z}},$$

$$\text{Ramp loss: } \ell_{\mathrm{R}}(z) = \frac{1}{2}\max\Big(0, \min\big(2, 1-z\big)\Big)$$

Complementary Label Learning

Estimation Error Bounds

Theorem 6. *For any $\delta > 0$, with probability at least $1 - \delta$,*

$$R(\widehat{f}) - R(f^*) \leq 4K(K-1)L_\ell \mathfrak{R}_n(\mathcal{G}) + (K-1)\sqrt{\frac{8\ln(2/\delta)}{n}},$$

if $(\widehat{g}_1, \ldots, \widehat{g}_K)$ is trained by minimizing $\widehat{R}(f)$ is w.r.t. $\overline{\mathcal{L}}_{\mathrm{OVA}}$, and

$$R(\widehat{f}) - R(f^*) \leq 8K(K-1)^2 L_\ell \mathfrak{R}_n(\mathcal{G}) + (K-1)^2\sqrt{\frac{2\ln(2/\delta)}{n}},$$

if $(\widehat{g}_1, \ldots, \widehat{g}_K)$ is trained by minimizing $\widehat{R}(f)$ is w.r.t. $\overline{\mathcal{L}}_{\mathrm{PC}}$.

Experiments

Table 1: Means and standard deviations of classification accuracy over five trials in percentage, when the number of classes ("cls") is changed for the MNIST dataset. "PC" is (10), "OVA" is (9), "Sigmoid" is (13), and "Ramp" is (14). Best and equivalent methods (with 5% t-test) are highlighted in boldface.

Method	3 cls	4 cls	5 cls	6 cls	7 cls	8 cls	9 cls	10 cls
OVA Sigmoid	**95.2 (0.9)**	**91.4 (0.5)**	**87.5 (2.2)**	**82.0 (1.3)**	**74.5 (2.9)**	**73.9 (1.2)**	**63.6 (4.0)**	**57.2 (1.6)**
OVA Ramp	**95.1 (0.9)**	**90.8 (1.0)**	**86.5 (1.8)**	**79.4 (2.6)**	**73.9 (3.9)**	**71.4 (4.0)**	**66.1 (2.1)**	**56.1 (3.6)**
PC Sigmoid	**94.9 (0.5)**	**90.9 (0.8)**	**88.1 (1.8)**	**80.3 (2.5)**	**75.8 (2.5)**	**72.9 (3.0)**	**65.0 (3.5)**	**58.9 (3.9)**
PC Ramp	**94.5 (0.7)**	**90.8 (0.5)**	**88.0 (2.2)**	**81.0 (2.2)**	**74.0 (2.3)**	**71.4 (2.4)**	**69.0 (2.8)**	**57.3 (2.0)**

Experiments

Dataset	Class	Dim	# train	# test	PC/S	PL	ML
WAVEFORM1	$1 \sim 3$	21	1226	398	**85.8(0.5)**	**85.7(0.9)**	79.3(4.8)
WAVEFORM2	$1 \sim 3$	40	1227	408	**84.7(1.3)**	**84.6(0.8)**	74.9(5.2)
SATIMAGE	$1 \sim 7$	36	415	211	**68.7(5.4)**	60.7(3.7)	33.6(6.2)
PENDIGITS	$1 \sim 5$	16	719	336	**87.0(2.9)**	76.2(3.3)	44.7(9.6)
	$6 \sim 10$		719	335	**78.4(4.6)**	71.1(3.3)	38.4(9.6)
	even #		719	336	**90.8(2.4)**	76.8(1.6)	43.8(5.1)
	odd #		719	335	**76.0(5.4)**	67.4(2.6)	40.2(8.0)
	$1 \sim 10$		719	335	**38.0(4.3)**	33.2(3.8)	16.1(4.6)
DRIVE	$1 \sim 5$	48	3955	1326	**89.1(4.0)**	77.7(1.5)	31.1(3.5)
	$6 \sim 10$		3923	1313	**88.8(1.8)**	78.5(2.6)	30.4(7.2)
	even #		3925	1283	**81.8(3.4)**	63.9(1.8)	29.7(6.3)
	odd #		3939	1278	**85.4(4.2)**	74.9(3.2)	27.6(5.8)
	$1 \sim 10$		3925	1269	**40.8(4.3)**	32.0(4.1)	12.7(3.1)
LETTER	$1 \sim 5$	16	565	171	**79.7(5.3)**	**75.1(4.4)**	28.3(10.4)
	$6 \sim 10$		550	178	**76.2(6.2)**	66.8(2.5)	34.0(6.9)
	$11 \sim 15$		556	177	**78.3(4.1)**	67.4(3.3)	28.6(5.0)
	$16 \sim 20$		550	184	**77.2(3.2)**	68.4(2.1)	32.7(6.4)
	$21 \sim 25$		585	167	**80.4(4.2)**	75.1(1.9)	32.0(5.7)
	$1 \sim 25$		550	167	**5.1(2.1)**	**5.0(1.0)**	**5.2(1.1)**
USPS	$1 \sim 5$	256	652	166	**79.1(3.1)**	70.3(3.2)	44.4(8.9)
	$6 \sim 10$		542	147	**69.5(6.5)**	**66.1(2.4)**	37.3(8.8)
	even #		556	147	**67.4(5.4)**	**66.2(2.3)**	35.7(6.6)
	odd #		542	147	**77.5(4.5)**	69.3(3.1)	36.6(7.5)
	$1 \sim 10$		542	127	**30.7(4.4)**	26.0(3.5)	13.3(5.4)

Thanks