On the benefits of output sparsity for multi-label classification

Evgenii Chzhen
http://echzhen.com
Université Paris-Est, Télécom Paristech

Joint work with: Christoph Denis, Mohamed Hebiri, Joseph Salmon

Outline

Outline

Framework and notation

We have N observations and each observation belongs to a set of labels.

- Observations: $X_i \in \mathbb{R}^D$,
- Label vectors = binary vectors: $Y_i = (Y_i^1, \ldots, Y_i^L)^\top \in \{0,1\}^L$,
- N, L, D - huge and probably $N \approx L$,
- Y_i consists of at most K ones (active labels) and $K \ll L$.

Outline

Motivation

0-type error vs 1-type error

$\hat{Y}^l = 1$ when $Y^l = 0$

$\hat{Y}^l = 0$ when $Y^l = 1$

Motivation

0-type error vs 1-type error

$\hat{Y}^l = 1$ when $Y^l = 0$

$\hat{Y}^l = 0$ when $Y^l = 1$

Example

$$Y = (\underbrace{1, \dots, 1}_{10}, \underbrace{0, \dots, 0}_{90})^\top \ ,$$

$$\hat{Y}_0 = (\underbrace{1, \dots, 1}_{10}, \underbrace{\color{red}{1, \dots, 1}}_{5}, \underbrace{0, \dots, 0}_{85})^\top \ ,$$

$$\hat{Y}_1 = (\underbrace{1, \dots, 1}_{5}, \underbrace{\color{red}{0, \dots, 0}}_{5}, \underbrace{0, \dots, 0}_{90})^\top \ .$$

- Same amount of mistakes but of different type
- Which one is better for a user?

Motivation

0-type error vs 1-type error

$\hat{Y}^l = 1$ when $Y^l = 0$ $\qquad$ $\hat{Y}^l = 0$ when $Y^l = 1$

Hamming loss

$$\mathcal{L}_H(Y, \hat{Y}) = \sum_{l=1}^{L} \mathbb{1}_{\{Y^l \neq \hat{Y}^l\}} = \sum_{Y^l=0} \mathbb{1}_{\{\hat{Y}^l=1\}} + \sum_{Y^l=1} \mathbb{1}_{\{\hat{Y}^l=0\}}$$

- For Hamming loss $\hat{Y}_0$ and $\hat{Y}_1$ are the same,
- Hamming loss does not know anything about sparsity K,
- But Hamming is separable, hence easy to optimize.

Outline

Our approach: add weights

Weighted Hamming loss

$$\mathcal{L}(Y,\hat{Y}) = p_0 \sum_{Y^l=0} \mathbb{1}_{\{\hat{Y}^l=1\}} + p_1 \sum_{Y^l=1} \mathbb{1}_{\{\hat{Y}^l=0\}} \ ,$$

such that $p_0 + p_1 = 1$.

Our approach: add weights

Weighted Hamming loss

$$\mathcal{L}(Y,\hat{Y}) = p_0 \sum_{Y^l=0} \mathbb{1}_{\{\hat{Y}^l=1\}} + p_1 \sum_{Y^l=1} \mathbb{1}_{\{\hat{Y}^l=0\}} \ ,$$

such that $p_0 + p_1 = 1$.

Examples

- Hamming loss: $p_0 = p_1 = 0.5$
- [Jain et al., 2016] : $p_0 = 0$ and $p_1 = 1$
- Our choice: $p_0 = \frac{2K}{L}$ and $p_1 = 1 - p_0$

Why our choice of weights?

Consider the following situation

- $Y = (\underbrace{1, \ldots, 1}_{K}, \underbrace{0, \ldots, 0}_{L-K})^\top$
- $\hat{Y}_0 = (0, \ldots, 0)^\top$: predicts all labels inactive,
- $\hat{Y}_1 = (1, \ldots, 1)^\top$: predicts all labels active,
- $\hat{Y}_{2K} = (\underbrace{1, \ldots, 1}_{2K}, \underbrace{0, \ldots, 0}_{L-2K})$: makes K mistakes of 0-type
- Do not forget that $K \ll L$

Why our choice of weights?

Consider the following situation

- $Y = (\underbrace{1, \ldots, 1}_{K}, \underbrace{0, \ldots, 0}_{L-K})^\top$
- $\hat{Y}_0 = (0, \ldots, 0)^\top$: predicts all labels inactive,
- $\hat{Y}_1 = (1, \ldots, 1)^\top$: predicts all labels active,
- $\hat{Y}_{2K} = (\underbrace{1, \ldots, 1}_{2K}, \underbrace{0, \ldots, 0}_{L-2K})$: makes K mistakes of 0-type
- Do not forget that $K \ll L$

Classical Hamming loss

- $\hat{Y}_1$ is almost the worst
- $\hat{Y}_0$ is the same as $\hat{Y}_{2K}$

Why our choice of weights?

Consider the following situation

- $Y = (\underbrace{1, \dots, 1}_{K}, \underbrace{0, \dots, 0}_{L-K})^\top$
- $\hat{Y}_0 = (0, \dots, 0)^\top$: predicts all labels inactive,
- $\hat{Y}_1 = (1, \dots, 1)^\top$: predicts all labels active,
- $\hat{Y}_{2K} = (\underbrace{1, \dots, 1}_{2K}, \underbrace{0, \dots, 0}_{L-2K})$: makes K mistakes of 0-type
- Do not forget that $K \ll L$

[Jain et al., 2016]

- $\hat{Y}_0$ is the worst
- $\hat{Y}_1$ is the same as $\hat{Y}_{2K}$

Why our choice of weights?

Consider the following situation

- $Y = (\underbrace{1, \dots, 1}_{K}, \underbrace{0, \dots, 0}_{L-K})^\top$
- $\hat{Y}_0 = (0, \dots, 0)^\top$: predicts all labels inactive,
- $\hat{Y}_1 = (1, \dots, 1)^\top$: predicts all labels active,
- $\hat{Y}_{2K} = (\underbrace{1, \dots, 1}_{2K}, \underbrace{0, \dots, 0}_{L-2K})$: makes K mistakes of 0-type
- Do not forget that $K \ll L$

Our choice

- $\hat{Y}_0$, $\hat{Y}_1$ are almost the worst
- $\hat{Y}_{2K}$ is almost the best

Outline

Numerical results

Synthetic dataset with controlled sparsity: $N = 2D = 2L = 200$

Settings	Median output sparsity		Recall (micro)		Precision (micro)	
	Our	Std	Our	Std	Our	Std
$K = 2$	2.47	0.04	1.0	0.02	0.80	1.0
$K = 6$	6.83	0.43	1.0	0.07	0.88	1.0
$K = 10$	9.85	1.81	0.90	0.18	0.91	1.0
$K = 14$	10.90	4.11	0.72	0.29	0.93	0.99
$K = 18$	10.98	6.61	0.58	0.36	0.95	0.99

- When $K \ll L$ we output MORE active labels,
- Hence, better Recall and worse Precision,
- When $K > 10$ our setting are violated.

Conclusion

- For sparse datasets: errors of 0/1-type are not the same for a user;
- Use our framework if you agree with the previous idea;
- We do not introduce a new algorithm per se, but we construct a new loss;
- We provide a theoretical justification to our framework (generalization bounds and analysis of convex surrogates).

Thank you for your attention!