Sharp Analysis of Learning with Discrete Losses

Alex Nowak-Vila, Francis Bach and Alessandro Rudi

INRIA - Ecole Normale Supérieure - PSL Research University

Challenges in Structured Prediction

In structured prediction:

- the **number of possible labels is exponentially large** w.r.t to the natural dimension of the data.
- we generally have **more possible outputs than training data**.

Examples of typical structured prediction problems:

- *Multilabel prediction:* predict a subset of labels.
- *Sequence prediction:* predict a sequence over a fixed dictionary.
- *Ranking:* predict a permutation.

Q: When is learning statistically and computationally feasible in structured prediction?

Supervised Learning Setting

- **Spaces:** input space $\mathcal{X}$, **discrete** label space $\mathcal{Y}$ and **discrete** output space $\mathcal{Z}$.
- **Data:** n i.i.d observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ from a distribution P.
- **Structured loss:** a loss between outputs and labels

$$L : \mathcal{Z} \times \mathcal{Y} \longrightarrow \mathbb{R}.$$

- **Expected risk and Bayes risk:** minimize the expected risk $\mathcal{E}(f)$:

$$\mathcal{E}(f) = \mathbb{E}_{(X,Y)\sim P} L(f(X), Y) = \int_{\mathcal{X}} \ell(f(x), x) dP(x),$$

where $\ell(z, x) = \int_{\mathcal{Y}} L(z, y) dP(y|x)$ is the Bayes risk.

- **Bayes classifier:** the Bayes classifier $f^\star : \mathcal{X} \to \mathcal{Z}$ has the form:

$$f^\star(x) = \arg\min_{z\in\mathcal{Z}} \ \ell(z, x), \quad f^\star = \arg\min_{f:\mathcal{X}\to\mathcal{Z}} \ \mathcal{E}(f).$$

Quadratic Surrogate (QS) Estimator

Given a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined on the input space and $\lambda > 0$, the QS estimator introduced in [1] has the form

$$\widehat{f}_n(x) = \arg\min_{z\in\mathcal{Z}} \sum_{i=1}^{n} \alpha_i(x) L(z, y_i), \qquad (1)$$

where $\alpha(x) = (K + n\lambda I)^{-1} K_x \in \mathbb{R}^n$ with $K_x = (k(x, x_1), \ldots, k(x, x_n)) \in \mathbb{R}^n$ and $K \in \mathbb{R}^{n\times n}$ is defined by $K_{ij} = k(x_i, x_j)$.

General Analysis of the Estimator

It is known that the resulting estimator is consistent. Moreover, we have the following generalization bound ([1]):

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^\star) \ \leq \ C\, n^{-1/4},$$

where C is a constant. A priori, there is no control over the magnitude of the constant C ([2]).

- **If $C \sim |\mathcal{Y}|, |\mathcal{Z}|$, then the bound is generally non-informative.**

Goal: Characterize C for discrete losses.

Affine Decomposition of the Loss

We consider an **affine decomposition** of the loss matrix $L \in \mathbb{R}^{|\mathcal{Z}|\times|\mathcal{Y}|}$:

$$L = FU^\top + c\mathbf{1}. \qquad (2)$$

where $F \in \mathbb{R}^{|\mathcal{Z}|\times r}, U \in \mathbb{R}^{|\mathcal{Y}|\times r}$, $c \in \mathbb{R}$, $\mathbf{1} \in \mathbb{R}^{|\mathcal{Z}|\times|\mathcal{Y}|}$ is the matrix of ones and $r \in \mathbb{N}$.

- We will use it to characterize the statistical and computational complexity of learning with loss L.

Statistical Complexity Analysis

Let $n \in \mathbb{N}, \tau > 0$ and $\lambda_n = n^{-1/2}$. Assume that the loss L decomposes as Eq. (2). If $g^* \in \mathcal{G}$, we have that with probability $1 - 8e^{-\tau}$,

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^\star) \ \leq \ \mathsf{A} Q\, c\tau^2\, n^{-1/4},$$

where $\mathsf{A} = \sqrt{r}\|F\|_\infty U_{\max}$, $Q = \max_j \|g_j^*/U_{\max}\|$ and $U_{\max} = \max_{j,k} |U_{kj}|$.

The statistical complexity is characterized by the constant

$$\mathbf{A} = \sqrt{r}\|F\|_\infty U_{\max}.$$

- Provide **generalization of low-noise conditions for general losses:** improved rates with conditions of the form $P_{\mathcal{X}}(\gamma(X) \leq \varepsilon) = o(\varepsilon^p)$ for $p > 0$, where $\gamma(x) = \min_{z'\neq f^\star(x)} \ell(z', x) - \ell(f^\star(x), x)$.
- **Tightness:** Use calibration dimension introduced by [3] to analyze tightness of statistical bounds.

Computational Complexity Analysis

Computational complexity of solving Eq. (1) is the same as the one of solving the following minimization problem:

$$\min_{z\in\mathcal{Z}} \ F_z\, \theta, \qquad (3)$$

where $F_z \in \mathbb{R}^r$ is the z-th row of F and $\theta \in \mathbb{R}^r$.

The computational complexity is characterized by the cost of solving problem (3).

Analysis of Multilabel and Ranking Losses

- The label space of the following losses / scores is $\mathcal{Y} = \{0, 1\}^m$.

Multilabel and Ranking measures

Measure	$\mathcal{Z}$	Definition	r	A	$\mathrm{INF}_F(m)$
0-1 (↓)	$\mathcal{P}_m$	$1(z \neq y)$	2^m	$2^{m/2}$	$\mathcal{O}(n \wedge 2^m)$
Block 0-1 (↓)	$\mathcal{P}_m$	$1(z \in B_j, y \notin B_j, \ j \in [b])$	b	$\sqrt{b}$	$\mathcal{O}(b)$
Hamming (↓)	$\mathcal{P}_m$	$\frac{1}{m}\sum_{j=1}^m 1([z]_j \neq [y]_j)$	m	$\frac{1}{2}$	$\mathcal{O}(m)$
F-score (↑)	$\mathcal{P}_m$	$2\frac{\lvert z\cap y\rvert}{\lvert z\rvert+\lvert y\rvert}$	m^2+1	$\sqrt{2}m$	$\mathcal{O}(m^2)$
Prec@k (↑)	$\mathcal{P}_{m,k}$	$\frac{\lvert z\cap y\rvert}{k}$	m	$\sqrt{\frac{m}{k}}$	$\mathcal{O}(m\log k)$
NDCG (↑)	$\mathfrak{S}_m$	$\frac{1}{N(r)}\sum_{j=1}^m G([r]_j) D_{\sigma(j)}$	m	$\sqrt{m}\ (\sum_j D_j^2)^{\frac{1}{2}} G_{\max}$	$\mathcal{O}(m\log m)$.
PD (↓)	$\mathfrak{S}_m$	$\frac{1}{N(y)}\sum_{j,\ell=1}^m 1_{([y]_j<[y]_\ell)} 1_{(\sigma(j)>\sigma(\ell))}$	$\frac{m(m-1)}{2}$	$\frac{m}{4}$	$\mathrm{MWFAS}(m)$.
MAP (↑)	$\mathfrak{S}_m$	$\frac{1}{\lvert y\rvert}\sum_{j=1}^m \frac{[y]_j}{\sigma(j)}\sum_{\ell=1}^{\sigma(j)} y_{\sigma^{-1}(\ell)}$	$\frac{m(m+1)}{2}$	$\frac{1}{2}m\sqrt{\log(m+1)}$	$\mathrm{QAP}(m)$.

Experiments

- Perform experiments on multilabel and ranking datasets: compare with SSVM and threshold-based method.

Multilabel		bibtex	birds	CAL500	corel5k	enron	mediamill	medical	scene	yeast
	n	7395	645	502	5000	1702	43907	978	2407	2417
	d	1836	260	68	499	1001	120	1449	294	103
	m	159	19	174	374	53	101	45	6	14
0-1 (↓)	THBM	0.82	0.57	1.0	0.99	0.92	0.93	0.31	0.49	0.93
	SSVM	0.91	0.53	1.0	0.99	0.90	1.0	0.35	0.51	0.95
	QS	**0.78**	**0.52**	1.0	**0.95**	**0.86**	**0.86**	**0.29**	**0.34**	**0.76**
Ham (↓)	THBM	1.3e-2	7.9e-2	0.14	1.1e-2	**5.9e-2**	3.1e-2	**9.4e-3**	0.11	**0.26**
	SSVM	1.3e-2	6.4e-2	0.13	1.0e-2	7.1e-2	8.7e-2	1.07e-2	0.11	0.40
	QS	1.3e-2	**4.9e-2**	**0.14**	**9.4e-3**	8.6-2	3.1e-2	9.6e-3	0.11	0.42
F-score (↑)	THBM	0.44	0.25	0.46	0.25	0.51	**0.56**	0.80	0.63	**0.48**
	SSVM	0.19	0.16	0.33	0.11	0.49	0.40	0.74	0.57	**0.48**
	QS	**0.47**	**0.28**	**0.47**	**0.26**	**0.52**	**0.56**	**0.83**	**0.68**	0.47

Ranking		Ohsumed
	n	106
	d	25
	m	150
NDCG@3 (↑)	SSVM	0.47
	QS	**0.51**
NDCG@5 (↑)	SSVM	0.45
	QS	**0.48**
NDCG@10 (↑)	SSVM	0.43
	QS	**0.46**

- **Take-away message:** importance of being consistent and calibrated to the measure of interest.

Main References

- [1] Carlo Ciliberto, Lorenzo Rosasco and Alessandro Rudi. A consistent regularization approach for structured prediction. In *Advances in Neural Information Processing Systems*, 2016.
- [2] Anton Osokin, Francis Bach and Simon Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. In *Advances in Neural Information Processing Systems*, 2017.
- [3] Harish G. Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *The Journal of Machine Learning Research*, 17(1):397–441, 2016.