A consistent algorithmic framework for structured machine learning

Lorenzo Rosasco
Universitá di Genova+ MIT + IIT
`lcsl.mit.edu`

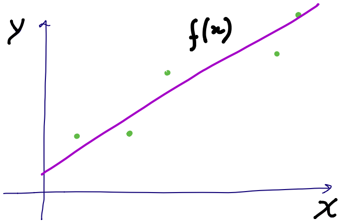
April 29th, 2019 -IPAM

joint work with C. Ciliberto (Imperial College), A. Rudi (INRIA-Paris)

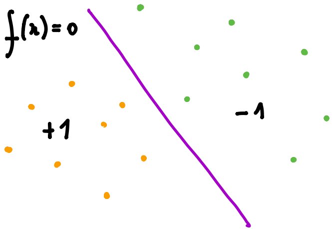
Classic supervised learning

given $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ **find** $f(x_{\text{new}}) \sim y_{\text{new}}$

Regression



Binary classification



Structured learning

''A domain of machine learning, in which the prediction must satisfy the additional constraints found in **structured data**, poses one of machine learning's greatest challenges: learning functional dependencies between arbitrary input and output domains.''

Baklr et al., Predicting structured data. MIT press, 2007. [1]

Structured learning applications

- Image segmentation [2],
- captioning [3],
- speech recognition [4, 5],
- protein folding [6],
- ordinal regression [7],
- ranking [8].

Examples of “structured” outputs

- Finite discrete alphabets (binary/multi-category classification, multilabel),
- strings,
- ordered lists,
- sequences.

<u>Classically only discrete possibly output spaces.</u>

Classical approaches

Likelihood estimation models

- General approaches (Struct-SVM [9], Conditional Random Fields [10]),
- but limited guarantees (generalization bounds).

Surrogate approaches

- Strong theoretical guarantees,
- but ad hoc, e.g. classification [11], multiclass [12], ranking [8]...

We will try to take the best of both!

Outline

Framework

Algorithms

Theory

Experiments

Statistical learning

- $(\mathcal{X} \times \mathcal{Y}, \rho)$ probability space, such that $\rho(x, y) = \rho_{\mathcal{X}}(x)\rho(y|x)$.
- $\Delta : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$

Statistical learning

- $(\mathcal{X} \times \mathcal{Y}, \rho)$ probability space, such that $\rho(x, y) = \rho_{\mathcal{X}}(x)\rho(y|x)$.
- $\Delta : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$

Problem Solve

$$\min_{f \in \mathcal{Y}^{\mathcal{X}}} \int d\rho(x, y)\Delta(f(x), y)$$

given $(x_i, y_i)_{i=1}^n$ i.i.d. samples of ρ.

Empirical risk minimization (ERM)

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \Delta(f(x_i), y_i)$$

- Statistically sound

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \Delta(f(x_i), y_i) - \int d\rho(x, y) \Delta(f(x), y) \right|$$

- Impractical: how to pick $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ if $\mathcal{Y}$ is not linear?

Inner risk

Lemma (Ciliberto, Rudi, R. '17)

Let

$$f_* = \operatorname*{argmin}_{f \in \mathcal{Y}^{\mathcal{X}}} \int d\rho(x, y) \Delta(f(x), y)$$

then

$$f_*(x) = \operatorname*{argmin}_{y \in \mathcal{Y}} \int d\rho(y|x) \Delta(y, y').$$

Structured Encoding Loss Function (SELF)

Definition (SELF)

The loss function $\Delta : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ is such that there exists

- a real separable Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ and
- maps $\Psi, \Phi : \mathcal{Y} \to \mathcal{H}$

such that $\forall y, y' \in \mathcal{Y}$

$$\Delta(y, y') = \langle \Psi(y), \Phi(y') \rangle$$

Examples of SELF

- In any finite output spaces $|\mathcal{Y}| = T$

$$\Delta(y, y') \;=\; e_y^\top \, V \, e_{y'}, \quad V \in \mathbb{R}^{T \times T}.$$

- Symmetric positive definite loss functions, Kernel Dependency Estimator [16].

- Smooth loss functions with $\mathcal{Y} = [0, 1]^d$.

- Restriction of SELF are SELF, and SELF can be composed.

Structured statistical learning

$$(\mathcal{Y}, \Delta)$$

- The output space might not be a linear space and can be continuous.
- Structure encoded by the loss function.

Beyond finite, discrete spaces to include continuous output spaces, e.g.

- Manifold regression [14],
- prediction of probability distributions [15].

Inner SELF (risk)

$$\int d\rho(y|x)\Delta(f(x), y) = \int d\rho(y|x)\, \langle \Psi(y), \Phi(y') \rangle = \left\langle \underbrace{\int d\rho(y|x)\Psi(y)}_{g_*(x)}, \Phi(y') \right\rangle$$

Inner SELF (risk)

$$\int d\rho(y|x)\Delta(f(x),y) = \int d\rho(y|x)\,\langle \Psi(y), \Phi(y')\rangle = \left\langle \underbrace{\int d\rho(y|x)\Psi(y)}_{g_*(x)}, \Phi(y') \right\rangle$$

Lemma (Ciliberto, Rudi, R. '17)

$$f_*(x) = \operatorname*{argmin}_{y\in\mathcal{Y}} \langle g_*(x), \Phi(y)\rangle$$

$$g_* = \int d\rho(y|\cdot)\Psi(y) = \operatorname*{argmin}_{g\in\mathcal{H}^{\mathcal{X}}} \int d\rho(x,y)\|g(x) - \Psi(y)\|^2$$

Inner risk minimization (IRM)

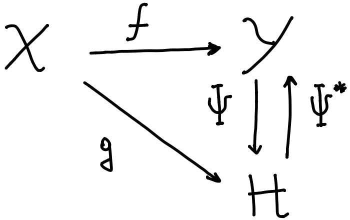
$$\hat{f}(x) = \operatorname*{argmin}_{y \in \mathcal{Y}} \langle \hat{g}(x), \Phi(y) \rangle$$

$$\hat{g} = \operatorname*{argmin}_{g \in \mathcal{G} \subset \mathcal{H}^{\mathcal{X}}} \frac{1}{n} \sum_{i=1}^{n} \| g(x_i) - \Psi(y_i) \|^2$$

IRM: a general surrogate approach

- encode $\Psi : \mathcal{Y} \to \mathcal{H}$
- learn $(x_i, \Psi(y_i))_{i=1}^n \mapsto \hat{g}$
- decode $\Psi^* : \mathcal{H} \to \mathcal{Y}$

$$\Psi^*(h) = \operatorname*{argmin}_{y \in \mathcal{Y}} \langle h, \Phi(y) \rangle, \qquad h \in \mathcal{H}.$$



Some questions

- A minimization over $\mathcal{Y}$ instead of $\mathcal{Y}^{\mathcal{X}}$: what we gained?

- Does a SELF exist?

Outline

Solving IRM with linear estimators

$$\hat{f}(x) = \operatorname*{argmin}_{y \in \mathcal{Y}} \langle \hat{g}(x), \Phi(y) \rangle, \qquad \hat{g} = \operatorname*{argmin}_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \|g(x_i) - \Psi(y_i)\|^2.$$

Solving IRM with linear estimators

$$\hat{f}(x) = \operatorname*{argmin}_{y \in \mathcal{Y}} \langle \hat{g}(x), \Phi(y) \rangle, \qquad \hat{g} = \operatorname*{argmin}_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \|g(x_i) - \Psi(y_i)\|^2.$$

Lemma (Ciliberto, Rudi, R. '17)

If $g(x) = Wx$*, then*

$$W = (\hat{X}^\top \hat{X})^{-1} \hat{X}^\top \hat{Y}, \qquad \hat{X} \in \mathbb{R}^n d, \quad \hat{Y} \in \mathcal{H}^n$$

and

$$\hat{g}(x) = \sum_{i=1}^{n} \alpha_i(x) \Psi(y_i), \qquad \alpha(x) = (\hat{X}\hat{X}^\top)^{-1} \hat{X} x \in \mathbb{R}^n$$

Implicit IRM for linear estimators

$$\hat{f}(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmin}} \langle \hat{g}(x), \Phi(y) \rangle, \qquad \hat{g} = \underset{g \in \mathcal{G}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \|g(x_i) - \Psi(y_i)\|^2.$$

Lemma (Ciliberto, Rudi, R. '17)

If

$$\hat{g}(x) = \sum_{i=1}^{n} \alpha_i(x) \Psi(y_i),$$

then

$$\hat{f}(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmin}} \sum_{i=1}^{n} \alpha_i(x) \Delta(y_i, y)$$

Other linear estimators

$$\hat{g}(x) = \sum_{i=1}^{n} \alpha_i(x)\Psi(y_i),$$

- Kernel methods $g(x) = W\gamma(x)$, where $\gamma : \mathcal{X} \to (\mathcal{H}_\Gamma, \langle\cdot,\cdot\rangle_\Gamma)$.
- Local kernel estimators.
- Spectral filters.
- Sketching/random features/Nÿstrom.

Computations: no free lunch

Training

$$\hat{g} = \operatorname*{argmin}_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \|g(x_i) - \Psi(y_i)\|^2.$$

Computing $(\alpha_i(x))_i$ depends only on the inputs and is efficient.

Prediction

$$\hat{f}(x) = \operatorname*{argmin}_{y \in \mathcal{Y}} \sum_{i=1}^{n} \alpha_i(x) \Delta(y_i, y).$$

Requires problem specific decoding and can be hard.

Outline

Consistency and excess risk bounds

Problem Solve

$$\min_{f \in \mathcal{Y}^{\mathcal{X}}} R(f), \qquad R(f) = \int d\rho(x,y)\Delta(f(x),y)$$

given $(x_i, y_i)_{i=1}^n$ i.i.d. samples of ρ.

Excess risk Convergence and rates on

$$R(\hat{f}) - R(f_*)$$

A relaxation error analysis

Let

$$L(g) = \int d\rho(x,y) \|g(x) - \Psi(y)\|^2$$

Theorem (Ciliberto, Rudi, R. '17)

The following hold:

- *Fisher consistency*

$$f_*(x) = \Psi^* g_*(x). \quad a.s.$$

- *Comparison inequality, for all g and $f(x) = \Psi^* g(x)$ a.s.*

$$R(f) - R(f_*) \le c_\Delta \sqrt{L(g) - L(g_*)}$$

where

$$c_\Delta = \sup_{y \in \mathcal{Y}} \|\Psi(y)\|$$

Consistency and rates for IRM-KRR

Let $\hat{g}_\lambda(x) = \hat{W}_\lambda \gamma(x)$ with

$$\hat{W}_\lambda = \operatorname*{argmin}_{W \in \mathcal{L}_2(\mathcal{H}_\Gamma, \mathcal{H})} \frac{1}{n} \sum_{i=1}^{n} \|W x_i - \Psi(y_i)\|^2 + \lambda \|W\|_2^2.$$

Theorem (Ciliberto, Rudi, R. '17)

Let $\kappa_\gamma = \sup_{x \in \mathcal{X}} \|\gamma(x)\|$. *Assume* $\exists W_* \in \mathcal{L}_2(\mathcal{H}_\Gamma, \mathcal{H})$ *such that* $g_*(x) = W_* x$. *If* $\lambda_n = O(1/\sqrt{n})$, *then with probability at least* $1 - 8e^{-\tau}$

$$\sqrt{L(\hat{g}) - L(g_*)} \;\leq\; 24\, \kappa_\gamma \, (1 + \|W\|_2)\, \tau^2 n^{-1/4}.$$

and for $\hat{f}(x) = \Psi^* \hat{g}_\lambda(x)$ *a.s.*

$$R(\hat{f}) - R(f_*) \;\leq\; 24\, \kappa_\gamma \, c_\Delta (1 + \|W\|_2)\, \tau^2 n^{-1/4}.$$

Remarks

- This is the first result establishing consistency and rates for structured prediction, see [13] for similar efforts.

- The bound on $L(\hat{g}) - L(g_*)$ extend results in [17] under weaker assumptions.

- The constant c_Δ is problem dependent. Finding a general estimate is an open problem [18].

Outline

Ranking

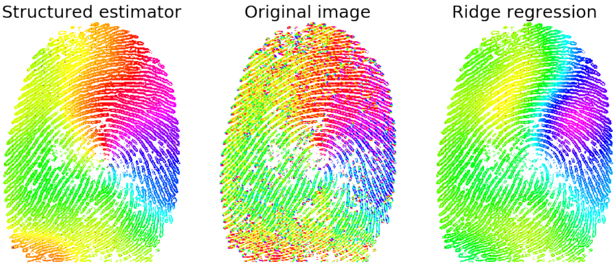
	Rank Loss
Linear [8]	0.430 ± 0.004
Hinge [19]	0.432 ± 0.008
Logistic [20]	0.432 ± 0.012
SVM Struct [9]	0.451 ± 0.008
IRM-KRR	$\mathbf{0.396 \pm 0.003}$

Ranking movies in the MovieLens dataset [21] (ratings (from 1 to 5) of 1682 movies by 943 users). The goal is predict preferences of a given user, i.e. an ordering of the 1682 movies, according to the user's partial ratings. We the loss [8]

$$\Delta_{rank}(y, y') = \frac{1}{2} \sum_{i,j=1}^{M} \gamma(y')_{ij} \; (1 - \mathrm{sign}(y_i - y_j)),$$

Fingerprints reconstruction

	Δ Deg.
KRLS	26.9 ± 5.4
MR[14]	22 ± 6
SP (ours)	$\mathbf{18.8 \pm 3.9}$



Average absolute error (in degrees) for the manifold structured estimator (SP), the manifold regression (MR) approach in [14] and the KRLS baseline. (Right) Fingerprint reconstruction of a single image where the structured predictor achieves 15.7 of average error while KRLS 25.3. The loss is the geodesic on $\mathcal{S}$

$$\Delta_{\mathcal{S}}(z, y) = \arccos\left(\langle z, y\rangle\right)^2$$

Summing up

- First consistent algorithmic framework for StructML.
- A general surrogate approach.
- TBD: decoding computations+ beyond linear estimators.

Openings

European Research Council
Executive Agency

Multiple openings for post-docs/PhD positions!

@lrntzrsc

$\rightarrow$ **Launching: Machine Learning Genova Center!**

Related papers

- Ciliberto, Rudi and Rosasco A consistent regularization approach for structured prediction. NIPS 2016.
- Ciliberto, Rudi and Rosasco, and Pontil. Consistent multitask learning with nonlinear output relations, NIPS 2017.
- Rudi, Ciliberto, Marconi, and Rosasco. Manifold structured prediction. NIPS 2018.
- Mroueh, Poggio, Rosasco, and Slotine. Multiclass learning with simplex coding. NIPS 2012.

Bakir Gökhan, Thomas Hofmann, Bernhard Schölkopf, Alexander J. Smola, Ben Taskar, and S.V.N Vishwanathan.
Predicting structured data.
MIT press, 2007.

Karteek Alahari, Pushmeet Kohli, and Philip HS Torr.
Reduce, reuse & recycle: Efficiently solving multi-label mrfs.
In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–8. IEEE, 2008.

Andrej Karpathy and Li Fei-Fei.
Deep visual-semantic alignments for generating image descriptions.
In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3128–3137, 2015.

Lalit Bahl, Peter Brown, Peter De Souza, and Robert Mercer.
Maximum mutual information estimation of hidden markov model parameters for speech recognition.
In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86., volume 11, pages 49–52. IEEE, 1986.

Charles Sutton, Andrew McCallum, et al.

An introduction to conditional random fields.
Foundations and Trends® in Machine Learning, 4(4):267–373, 2012.

Thorsten Joachims, Thomas Hofmann, Yisong Yue, and Chun-Nam Yu.
Predicting structured objects with support vector machines.
Communications of the ACM, 52(11):97–104, 2009.

Fabian Pedregosa, Francis Bach, and Alexandre Gramfort.
On the consistency of ordinal regression methods.
The Journal of Machine Learning Research, 18(1):1769–1803, 2017.

John C Duchi, Lester W Mackey, and Michael I Jordan.
On the consistency of ranking algorithms.
In Proceedings of the 27th International Conference on Machine Learning (ICML-10), pages 327–334, 2010.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun.
Large margin methods for structured and interdependent output variables.
In Journal of Machine Learning Research, pages 1453–1484, 2005.

Sebastian Nowozin, Christoph H Lampert, et al.
Structured learning and prediction in computer vision.
Foundations and Trends® in Computer Graphics and Vision, 6(3–4):185–365, 2011.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe.
Convexity, classification, and risk bounds.
Journal of the American Statistical Association, 101(473):138–156, 2006.

Youssef Mroueh, Tomaso Poggio, Lorenzo Rosasco, and Jean-Jacques Slotine.
Multiclass learning with simplex coding.
In Advances in Neural Information Processing Systems (NIPS) 25, pages 2798–2806, 2012.

Anton Osokin, Francis Bach, and Simon Lacoste-Julien.
On structured prediction theory with calibrated convex surrogate losses.
In Advances in Neural Information Processing Systems, pages 302–313, 2017.

Florian Steinke, Matthias Hein, and Bernhard Schölkopf.
Nonparametric regression between general riemannian manifolds.
SIAM Journal on Imaging Sciences, 3(3):527–563, 2010.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio.
Learning with a wasserstein loss.
In Advances in Neural Information Processing Systems, pages 2053–2061, 2015.

Jason Weston, Olivier Chapelle, Vladimir Vapnik, André Elisseeff, and Bernhard Schölkopf.
Kernel dependency estimation.

In Advances in neural information processing systems, pages 873–880, 2002.

Andrea Caponnetto and Ernesto De Vito.
Optimal rates for the regularized least-squares algorithm.
Foundations of Computational Mathematics, 7(3):331–368, 2007.

Alex Nowak-Vila, Francis Bach, and Alessandro Rudi.
Sharp analysis of learning with discrete losses.
arXiv preprint arXiv:1810.06839, 2018.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer.
Large margin rank boundaries for ordinal regression.
Advances in neural information processing systems, pages 115–132, 1999.

Ofer Dekel, Yoram Singer, and Christopher D Manning.
Log-linear models for label ranking.
In Advances in neural information processing systems, page None, 2004.

F Maxwell Harper and Joseph A Konstan.
The movielens datasets: History and context.
ACM Transactions on Interactive Intelligent Systems (TiiS), 5(4):19, 2015.