Maximum Margin Multi-Label Structured Prediction

Christoph H. Lampert
Institute of Science and Technology Austria (IST Austria)

Summary

Task: structured output learning when each input can have several correct outputs.

Examples: *protein folds, graph matchings with symmetry, object locations in images, natural language translations,...*

Contribution: max-margin formulation

- convex,
- efficient by use of working sets,
- tractable even for exponentially sized label sets,
- PAC-Bayesian generalization bound.

$\rightarrow$ essentially same as single-label SSVM [1].

Idea: build on one-versus-rest SVM instead of Crammer-Singer multiclass SVM, as SSVM does.

Common Notation

- $\mathcal{X}$ inputs (arbitrary)
- $\mathcal{Y}$ structured labels, (e.g. *graph-labelings*, *parse trees*, ...)
- $k : (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$: *joint kernel function*, $\phi : \mathcal{X} \times \mathcal{Y} \to \mathcal{H}$: *induced feature map*.
- $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, *compatibility function*, $f(x, y) = \langle w, \phi(x, y) \rangle_{\mathcal{H}}$

Single-Label Structured Prediction: CRFs, SSVMs, ...

Single-label prediction: learn a function $f : \mathcal{X} \to \mathcal{Y}$ from training data $(x^1, y^1), \dots, (x^n, y^n)$, with $x^i \in \mathcal{X}$, $y^i \in \mathcal{Y}$.

- $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$: task-dependent loss
  - $\Delta(y, \bar{y})$ cost of predicting $\bar{y}$ if y is correct

SSVM training problem (slack-rescaled) [1]:

$$(w^*, \xi^*) = \operatorname*{argmin}_{w \in \mathcal{H}, \xi^1, \dots, \xi^n \in \mathbb{R}^+} \frac{1}{2}\|w\|^2_{\mathcal{H}} + \frac{C}{n}\sum_{i=1}^{n} \xi^i,$$

subject to, for $i = 1, \dots, n$, for all $y \in \mathcal{Y}$,

$$\xi^i \geq \Delta(y^i, y)[1 + f(x^i, y) - f(x^i, y^i)]$$

Prediction function: $g(x, y) = \operatorname{argmax}_{y \in \mathcal{Y}} f(x, y)$

Obs: For $\mathcal{Y} = \{1, \dots, M\}$, $k((x, y), (\bar{x}, \bar{y})) = \tilde{k}(x, \bar{x})[\![y = \bar{y}]\!]$, $\Delta(y, \bar{y}) = [\![y \neq \bar{y}]\!]$, SSVM turns into Crammer-Singer MC-SVM [2].

Multi-Label Structured Prediction

Multi-label prediction: learn a function $f : \mathcal{X} \to \mathbb{P}(\mathcal{Y})$, from training set $(x^1, Y^1), \dots, (x^n, Y^n)$ with $Y^i \subset \mathcal{Y}$. ($Y^i = \emptyset$ is allowed)

- $\lambda : \mathbb{P}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}^+$: *task-dependent per-label misclassification cost*
  - if $\bar{y} \in Y$: $\lambda(Y, \bar{y})$ cost of *not predicting* $\bar{y}$,
  - if $\bar{y} \notin Y$: $\lambda(Y, \bar{y})$ cost of *predicting* $\bar{y}$,

First Idea: SSVM with output space $\mathbb{P}(\mathcal{Y})$ – $\mathbb{P}$-SSVM

- output set has $2^{|\mathcal{Y}|}$ elements $\rightarrow$ we might need $O(|\mathcal{Y}|)$ samples [3]
- working set training needs identifying most violated (x^i, Y) pair with $Y \subset \mathcal{Y}$, where Y could be as large as all of $\mathcal{Y}$

Main Insight:
$\mathbb{P}$-SSVM training does not scale to structured spaces with large $\mathcal{Y}$.

Proposed Method: MLSP

Define *indicator vectors*, $v^i \in \{\pm 1\}^{\mathcal{Y}}$, $v^i_y := \begin{cases} +1 & \text{if } y \in Y^i, \\ -1 & \text{otherwise.} \end{cases}$

MLSP training problem (slack-rescaled):

$$(w^*, \xi^*) = \operatorname*{argmin}_{w \in \mathcal{H}, \xi^1, \dots, \xi^n \in \mathbb{R}^+} \frac{1}{2}\|w\|^2_{\mathcal{H}} + \frac{C}{n}\sum_{i=1}^{n} \xi^i$$

subject to, for $i = 1, \dots, n$, for all $y \in \mathcal{Y}$,

$$\xi^i \geq \lambda(Y^i, y)[1 - v^i_y f(x^i, y)].$$

Prediction function: $g(x) = \{y \in \mathcal{Y} : f(x, y) > 0\}$. $(*)$

Obs: For $\mathcal{Y} = \{1, \dots, M\}$, $k((x, y), (\bar{x}, \bar{y})) = \tilde{k}(x, \bar{x})[\![y = \bar{y}]\!]$, $\lambda(Y, \bar{y}) \equiv 1$, MLSP turns into M binary SVMs, trained *one-vs-rest* [4].

Generalization Properties

Let $g_w(x) := \{y \in \mathcal{Y} : f_w(x, y) > 0\}$ for $f_w(x, y) = \langle w, \psi(x, y) \rangle$. Assume $|\mathcal{Y}| < r$ and $\|\psi(x, y)\| < s$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and $\lambda(Y, y) \leq \Lambda$ for all $(Y, y) \in \mathbb{P}(\mathcal{Y}) \times \mathcal{Y}$.

For any distribution Q_w over weight vectors, denote by $L(Q_w, P)$ the expected Δ-risk for P-distributed data,

$$L(Q_w, P) = \mathbb{E}_{\bar{w} \sim Q_w}\{\mathcal{R}_{P,\Delta}(g_{\bar{w}}(x))\} = \mathbb{E}_{\bar{w} \sim Q_w, (x,Y) \sim P}\{\Delta(Y, g_{\bar{w}}(x))\}.$$

where $\Delta(Y, \bar{Y}) := \max_{y \in Y \ominus \bar{Y}} \lambda(Y, y)$ is the *max*-loss over sets.

Theorem 1 (slack-rescaled version):
With probability at least $1 - \sigma$ over the sample S of size n, the following inequality holds simultaneously for all weight vectors w.

$$L(Q_w, P) \leq \frac{1}{n}\sum_{i=1}^{n} \ell(x^i, Y^i, f) + \frac{\|w\|^2}{n} + \left(\frac{s^2\|w\|^2 \ln(r\, n/\|w\|^2) + \ln\frac{n}{\sigma}}{2(n-1)}\right)^{1/2}$$

for $\ell(x^i, Y^i, f) := \max_{y \in \mathcal{Y}} \lambda(Y^i, y)[\![v^i_y f(x^i, y) < 1]\!]$ (=margin violation).

Proof: Based on [3], see paper.

Main Insight:
Training $g : \mathcal{X} \to \mathbb{P}(\mathcal{Y})$ requires $O(\log |\mathcal{Y}|)$ samples, not $O(|\mathcal{Y}|)$.

Experiments

Baselines: SSVM with multi-label prediction rule $(*)$, $\mathbb{P}$-SSVM, JKSE [5]

Evaluation measures:

- $\Delta_{max}(Y, \bar{Y}) := \max_{y \in Y \ominus \bar{Y}} \lambda(Y, y)$ *max*-loss (lower is better)
- $\Delta_{sum}(Y, \bar{Y}) := \sum_{y \in Y \ominus \bar{Y}} \lambda(Y, y)$ *sum*-loss (lower is better)
- mAUC: area under per-sample precision-recall curves (higher is better)
- prec/rec/F1: precision, recall, F1-score (higher is better).

Hierarchical Classification:

$\lvert\mathcal{Y}\rvert = 10$	Δ_{max}	Δ_{sum}	mAUC	F1 (prec / rec)
MLSP	0.73	1.59	0.82	0.42 (0.40 / 0.46)
JKSE	1.00	1.91	0.54	0.23 (0.14 / 0.76)
SSVM	0.88	3.86	0.84	0.37 (0.24 / 0.88)
$\mathbb{P}$-*SSVM*	0.75	1.11	0.83	0.44 (0.48 / 0.41)
indep.	0.73	1.07	0.84	0.46 (0.61 / 0.38)

Object Detection:

$\lvert\mathcal{Y}\rvert \approx 10^6$	Δ_{max}	Δ_{sum}	F1 (prec / rec)
MLSP	0.66	1.31	0.46 (0.60 / 0.52)
JKSE	0.99	7.29	0.09 (0.60 / 0.16)
SSVM	0.93	3.71	0.21 (0.79 / 0.33)
$\mathbb{P}$-*SSVM*	*infeasible*		
indep.	*infeasible*		

Main Insight:
MLSPs achieve similar results to $\mathbb{P}$-SSVMs, but scale to large $\mathcal{Y}$.

Summary

- maximum margin framework for predicting multiple structured labels,
- efficient through re-use of techniques/insights from single-label SSVM.

[1] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. *Large margin methods for structured and interdependent output variables.* JMLR, 6, 2006.
[2] K. Crammer, Y. Singer. *On the algorithmic implementation of multiclass kernel-based vector machines*, JMLR, 2, 2001.
[3] D. McAllester. *Generalization bounds and consistency for structured labeling*, In *Predicting Structured Data*, MIT Press, 2007.
[4] V. N. Vapnik. *Statistical learning theory*, Wiley-Interscience, 1998.
[5] C. H. Lampert, M. B. Blaschko. *Structured prediction by joint kernel support estimation*, Machine Learning, 2009.