

Learning Max-Sum classifiers by Structured Output SVM

Vojtěch Franc

February 24, 2020

- ◆ Learning Max-Sum classifiers on acyclic graphs
- ◆ Learning Max-Sum classifiers with super-modular functions
- ◆ Learning generic Max-Sum classifiers via LP relaxation

XEP33SML – Structured Model Learning, Summer 2020

Structured Output Support Vector Machines

- ◆ Given $\mathcal{T} = \{(x^i, y^i) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \dots, m\}$ and a feature map $\phi: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^n$, we want to learn $\mathbf{w} \in \mathbb{R}^n$ of a classifier

$$h(\mathbf{x}; \mathbf{w}) = \operatorname{argmax}_{y \in \mathcal{Y}} \langle \mathbf{w}, \phi(x, y) \rangle$$

- ◆ SO-SVM with margin rescaling loss find \mathbf{w} by solving a convex problem

$$\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^n} \left(\frac{\lambda}{2} \|\mathbf{w}\|^2 + R^\psi(\mathbf{w}) \right)$$

where

$$R^\psi(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^m \max_{y \in \mathcal{Y}} \left(\ell(y^i, y) + \langle \mathbf{w}, \phi(x^i, y) \rangle - \langle \mathbf{w}, \phi(x^i, y^i) \rangle \right)$$

- ◆ For every loss $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ such that $\ell(y, y') = 0 \iff y = y'$, it holds that $R^\psi(\mathbf{w}) \geq R_{\mathcal{T}^m}(h) = \frac{1}{m} \sum_{i=1}^m \ell(y^i, h(x^i; \mathbf{w}))$.

SO-SVM solved via Cutting Plane Method

- ◆ The first order oracle computes the risk

$$R^\psi(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^m \max_{y \in \mathcal{Y}} \left(\ell(y^i, y) + \langle \mathbf{w}, \phi(x^i, y) \rangle - \langle \mathbf{w}, \phi(x^i, y^i) \rangle \right)$$

and one of its sub-gradient $\mathbf{g} \in \partial R^\psi(\mathbf{w})$ at any $\mathbf{w} \in \mathbb{R}^n$, e.g.

$$\mathbf{g} = \frac{1}{m} \sum_{i=1}^m \left(\phi(x^i, \hat{y}^i) - \phi(x^i, y^i) \right)$$

- ◆ To this end, we need to solve the loss augmented classification problem

$$\hat{y}^i = \operatorname{argmax}_{y \in \mathcal{Y}} \left(\ell(y^i, y) + \langle \mathbf{w}, \phi(x^i, y) \rangle \right)$$

Max-sum classifier and Hamming loss

- ◆ The max-sum classifier

$$\hat{\mathbf{y}} = \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}^{\mathcal{V}}} \langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \rangle := \sum_{v \in \mathcal{V}} q_v(x_v, y_v) + \sum_{\{v, v'\} \in \mathcal{E}} g_{vv'}(y_v, y_{v'})$$

where $q_v(x, y) = \langle \mathbf{w}, \phi_v(x, y) \rangle$ and $g_{vv'}(y, y') = \langle \mathbf{w}, \phi_{vv'}(y, y') \rangle$

- ◆ Hamming loss $\ell(\mathbf{y}, \mathbf{y}') = \sum_{v \in \mathcal{V}} [y_v \neq y'_{v'}]$
- ◆ The loss Augmented Classification Problem

$$\begin{aligned} \hat{\mathbf{y}}^i &= \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}^{\mathcal{V}}} \left[\ell(\mathbf{y}^i, \mathbf{y}) + \langle \mathbf{w}, \phi(\mathbf{x}^i, \mathbf{y}) \rangle \right] \\ &= \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}^{\mathcal{V}}} \left[\sum_{v \in \mathcal{V}} \left([y_v^i \neq y_v] + q_v(x^i, y_v) \right) + \sum_{\{v, v'\} \in \mathcal{E}} g_{vv'}(y_v, y_{v'}) \right] \end{aligned}$$

- ◆ The ACP is tractable for acyclic graph $(\mathcal{V}, \mathcal{E})$.

Super-modular Max-sum classifier and Hamming loss

- ◆ The max-sum classifier

$$\hat{\mathbf{y}} = \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}^{\mathcal{V}}} \langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \rangle := \sum_{v \in \mathcal{V}} q_v(x_v, y_v) + \sum_{\{v, v'\} \in \mathcal{E}} g_{vv'}(y_v, y_{v'})$$

where $g_{vv'}(y, y') = \langle \mathbf{w}, \phi_{vv'}(y, y') \rangle$ is super-modular.

- ◆ SO-SVM leads to

$$\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^n} \left(\frac{\lambda}{2} \|\mathbf{w}\|^2 + R^\psi(\mathbf{w}) \right)$$

subject to

$$g_{vv'}(y, y') + g_{vv'}(y + 1, y' + 1) - g_{vv'}(y, y' + 1) - g_{vv'}(y + 1, y') \geq 0, \\ \{v, v'\} \in \mathcal{E}, y, y' \in \{1, \dots, K - 1\}$$

- ◆ Provided the solver maintains intermediate solution \mathbf{w} feasible the ACPs are sub-modular and thus tractable.

BMRM with constraints

- ◆ Constrained regularized convex risk minimization

$$\mathbf{w}^* \in \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^n} \left(\frac{\lambda}{2} \|\mathbf{w}\|^2 + R(\mathbf{w}) \right) \quad \text{s.t.} \quad \mathbf{A}\mathbf{w} \leq \mathbf{b}$$

- ◆ The BMRM algorithm:

1. Init: $t \leftarrow 0$, $\mathbf{w}_0 \in \mathbb{R}^n$
2. Compute $R(\mathbf{w}_t)$ and $\mathbf{g}_t \in \partial R(\mathbf{w}_t)$
3. Solve the constrained reduced problem

$$\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^n} \left(\frac{\lambda}{2} \|\mathbf{w}\|^2 + R_t(\mathbf{w}) \right) \quad \text{s.t.} \quad \mathbf{A}\mathbf{w} \leq \mathbf{b}$$

where

$$R_t(\mathbf{w}) = \max_{i=0, \dots, t} \left[R(\mathbf{w}_i) + \langle \mathbf{g}_i, \mathbf{w} - \mathbf{w}_i \rangle \right]$$

4. if $\min_{i=1, \dots, t} F(\mathbf{w}_i) - F_t(\mathbf{w}_{t+1}) \leq \varepsilon$ stop else $t \leftarrow t + 1$ go to 2.

General max-sum classifier learned via LP relaxation

- ◆ The ACP leads to

$$\hat{\mathbf{y}}^i = \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}^{\mathcal{V}}} f^i(\mathbf{y}, \mathbf{w}) := \sum_{v \in \mathcal{V}} \left([y_v^i \neq y_v] + q_v(x^i, y_v) \right) + \sum_{\{v, v'\} \in \mathcal{E}} g_{vv'}(y_v, y_{v'})$$

- ◆ The value of ACP can be upper bounded via the LP relaxation:

$$\max_{\mathbf{y} \in \mathcal{Y}^{\mathcal{V}}} f^i(\mathbf{y}, \mathbf{w}) \leq \min_{\boldsymbol{\varphi}} E^i(\boldsymbol{\varphi}, \mathbf{w})$$

where $\boldsymbol{\varphi} \in \mathbb{R}^{2|\mathcal{E}||\mathcal{Y}|}$ is composed of $\varphi_{vv'}, \varphi_{v'v} : \mathcal{Y} \rightarrow \mathbb{R}$, $\{v, v'\} \in \mathcal{E}$ and

$$E^i(\boldsymbol{\varphi}, \mathbf{w}) = \sum_{v \in \mathcal{V}} \max_{y \in \mathcal{Y}} q_v^{\boldsymbol{\varphi}, \mathbf{w}}(y, x^i, y_v^i) + \sum_{\{v, v'\} \in \mathcal{E}} \max_{(y, y') \in \mathcal{Y}^2} g_{vv'}^{\boldsymbol{\varphi}, \mathbf{w}}(y, y')$$

$$q_v^{\boldsymbol{\varphi}, \mathbf{w}}(y, x^i, y_v^i) = [y_v^i \neq y_v] + q_v(x^i, y_v) - \sum_{v' \in \mathcal{N}(v)} \varphi_{vv'}(y), \quad v \in \mathcal{V}, y \in \mathcal{Y}$$

$$g_{vv'}^{\boldsymbol{\varphi}, \mathbf{w}}(y, y') = g_{vv'}(y, y') + \varphi_{vv'}(y) + \varphi_{v'v}(y'), \quad \{v, v'\} \in \mathcal{E}, y, y' \in \mathcal{Y}$$

General max-sum classifier learned via LP relaxation

- ◆ The LP-relaxed margin-rescaling loss:

$$\begin{aligned}
 \psi(\mathbf{x}^i, \mathbf{y}^i, \mathbf{w}) &= \max_{\mathbf{y} \in \mathcal{Y}^{\mathcal{V}}} \left(\ell(\mathbf{y}^i, \mathbf{y}) + \langle \mathbf{w}, \phi(\mathbf{x}^i, \mathbf{y}) \rangle \right) - \langle \mathbf{w}, \phi(\mathbf{x}^i, \mathbf{y}^i) \rangle \\
 &\leq \min_{\varphi} E^i(\varphi, \mathbf{w}) - \langle \mathbf{w}, \phi(\mathbf{x}^i, \mathbf{y}^i) \rangle \\
 &= \psi_{\text{LP}}(\mathbf{x}^i, \mathbf{y}^i, \mathbf{w})
 \end{aligned}$$

- ◆ SO-SVM leads to

$$\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^n} \left(\frac{\lambda}{2} \|\mathbf{w}\|^2 + R^\psi(\mathbf{w}) \right)$$

where

$$R^\psi(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^m \psi_{\text{LP}}(\mathbf{x}^i, \mathbf{y}^i, \mathbf{w})$$

Sub-gradient of minimum of convex function

Theorem 1. *Let $u: \mathbb{R}^{n_x+n_y} \rightarrow \mathbb{R}$ be a convex function. Then*

$$f(\mathbf{x}) = \min_{\mathbf{y} \in \mathbb{R}^{n_y}} u(\mathbf{x}, \mathbf{y}).$$

is a convex function and its sub-differential at $\hat{\mathbf{x}}$ reads

$$\partial f(\hat{\mathbf{x}}) = \partial_{\mathbf{x}} u(\mathbf{x}, \hat{\mathbf{y}}) \big|_{\mathbf{x}=\hat{\mathbf{x}}} \quad \text{where} \quad \hat{\mathbf{y}} \in \underset{\mathbf{y} \in \mathbb{R}^{n_y}}{\operatorname{argmin}} u(\hat{\mathbf{x}}, \mathbf{y}).$$

General max-sum classifier learned via LP relaxation

- ◆ We need the first order oracle of the LP loss:

$$\psi_{\text{LP}}(\mathbf{x}^i, \mathbf{y}^i, \mathbf{w}) = \min_{\varphi} E^i(\varphi, \mathbf{w}) - \langle \mathbf{w}, \phi(\mathbf{x}^i, \mathbf{y}^i) \rangle$$

- ◆ Solve the LP

$$\varphi^i = \underset{\varphi}{\operatorname{argmin}} E^i(\varphi, \mathbf{w})$$

- ◆ Compute the sub-gradient

$$\mathbf{g}^i \in \partial \psi_{\text{LP}}(\mathbf{x}^i, \mathbf{y}^i, \mathbf{w}) = \partial_{\mathbf{w}} E(\varphi^i, \mathbf{w}) - \phi(\mathbf{x}^i, \mathbf{y}^i)$$