Lecture 12 - structured SVM

Tuesday, February 23, 2021 14:27

today: structured SVM optimization

other approaches to optimize SVMstruct

(UP) unconstrained primal

$$\min_{w} \frac{\lambda \|w\|^2}{2} + \frac{1}{n}\sum_{i=1}^{n} H_i(w)$$

[unconstrained / non-smooth]

(PQP) primal QP ↳ quadratic problem

$$\min_{w,(\xi_i)_{i=1}^n} \frac{\lambda\|w\|^2}{2} + \frac{1}{n}\sum_{i=1}^{n} \xi_i$$

[constrained formulation]

$$\xi_i \geq H_i(w;\tilde{y}) \quad \forall \tilde{y} \in \mathcal{Y}_i \ \& \ \forall i$$

(smooth) strongly convex QP with exp # of linear constraints

1) generic approach to use convexity of loss augmented decoding: [Taskar & al. ICML 2005]

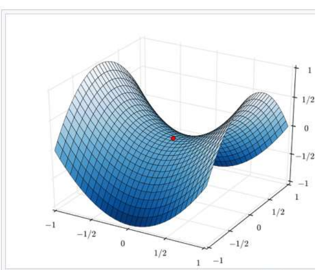
idea: here, we suppose that loss-augmented decoding can be expressed as a "compact" maximization problem of a convex fct.

i.e. $H_i(w) = \max_{\tilde{y}\in\mathcal{Y}_i} \ell_i(\tilde{y}) - \langle w, \psi_i(\tilde{y})\rangle = \max_{z\in\mathcal{Z}} g_i(w;z)$

$\tilde{y}$: discrete; z: cts.; $\mathcal{Z}$: convex set

where g_i is concave in z and convex in w

$\mathcal{Z}$: · should not depend on w
· " have a tractable description


A saddle point (in red) on the graph of $z=x^2-y^2$ (hyperbolic paraboloid)

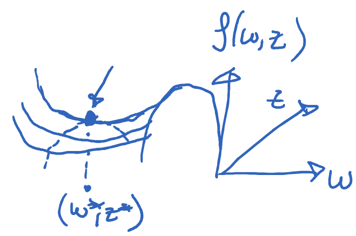
a) saddle point formulation:

$$\min_{w} \max_{\substack{z_i\in\mathcal{Z}_i \\ i=1,\dots,n}} \frac{\lambda\|w\|^2}{2} + \frac{1}{n}\sum_{i=1}^{n} g_i(w,z_i)$$

$$\min_{w}\max_{z} \mathcal{L}(w,z)$$

· convex in w
· concave in z

convex-concave saddle problem



(under reg. conditions $\min\max = \max\min$) → "saddle point"

$$\forall z \ f(w^*,z) \leq f(w^*,z^*) \leq f(w,z^*) \ \forall w$$

$w^* \in \arg\min f(w,z^*)$
$z^* \in \arg\max f(w^*,z)$

in general:

$$\min_{w}\max_{z} \geq \max_{z}\min_{w}$$

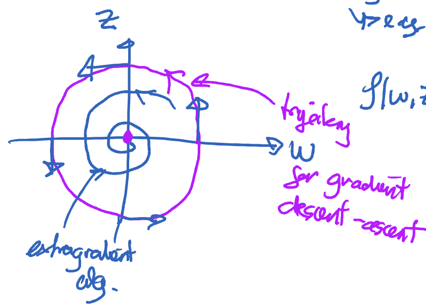
$$\min_{w}\max_{z} f(w,z) \geq \max_{z}\min_{w'} f(w',z)$$

$z^* \in \arg\max_z f(w^*, z)$ $\qquad$ w t $\qquad$ t w'

$\rightsquigarrow$ might not exist? $\qquad$ [but it always exists for convex-concave + reg. condition]
$\quad$ $\hookrightarrow$ e.g. bounded convex constraints

<u>standard alg.</u>: $\quad$ converges $O\left(\frac{1}{t}\right)$ for convex-concave game

<u>extragradient alg</u>:



$f(w,z) = \langle w, z \rangle$

$$\begin{pmatrix} -\nabla_w f \\ \nabla_z f \end{pmatrix} = \begin{pmatrix} -z \\ w \end{pmatrix}$$

$$\begin{pmatrix} \tilde{w}_{t+1} \\ \tilde{z}_{t+1} \end{pmatrix} = \begin{pmatrix} w_t \\ z_t \end{pmatrix} + \gamma_t \begin{pmatrix} -\nabla_w f(w_t, z_t) \\ \nabla_z f(w_t, z_t) \end{pmatrix}$$

"lookahead step"

$$\begin{pmatrix} w_{t+1} \\ z_{t+1} \end{pmatrix} = \begin{pmatrix} w_t \\ z_t \end{pmatrix} + \gamma_t \begin{pmatrix} -\nabla_w f(\tilde{w}_{t+1}, \tilde{z}_{t+1}) \\ \nabla_z f(\;" \quad "\;) \end{pmatrix}$$

$$x = \begin{pmatrix} w \\ z \end{pmatrix}$$

$$\tilde{x}_{t+1} = x_t - \gamma_t F(x_t) \qquad F(x) = \begin{pmatrix} \nabla_w f \\ -\nabla_z f \end{pmatrix}$$

$$x_{t+1} = x_t - \gamma_t F(\tilde{x}_{t+1})$$

1^{st} order approx. to the <u>implicit</u> method

$$x_{t+1} = x_t - \gamma_t F(x_{t+1}) \qquad x^* = x^* - \gamma F(x^*)$$

extragradient to structured SVM [Taskar et al. JMLR 2006]

b) <u>small "complicated" QP formulation</u> (for SVMstruct)

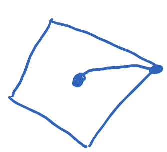
$$H_i(w) = \max_{z_i \in Z_i} g_i(w; z_i) \underset{\text{use strong duality}}{=} \min_{v_i \in V_i(w)} \overbrace{\tilde{g}_i(w; \underbrace{v_i}_{\text{dual variables}})}^{\text{convex dual of } \max_{z_i \in Z_i} g_i(w; z_i)}$$

<u>obtain</u> $\displaystyle \min_{w} \min_{\substack{v_i \in V_i(w) \\ \forall i}} \frac{\lambda \|w\|^2}{2} + \frac{1}{n} \sum_{i=1}^{n} \tilde{g}_i(w; v_i)$

if $\tilde{g}_i$ is jointly convex in w & v_i
we get a "tractable" convex min. problem

$\rightarrow$ can solve with favorite convex min alg.

<u>if</u> d & n too big, use interior point solver
e.g. Mosek, Cplex (commercial)
CVXopt (free, python)

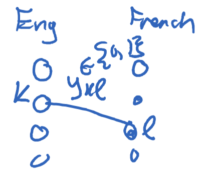


15h33

<u>examples $g_i(w; z_i)$</u> $\qquad$ Eng $\quad$ French
$0 \;\; \in \{0,1\}$

examples $Q_i(w; z_i)$

I) word alignment :



features on pair (x^E_k, x^F_ℓ)

recall that score $s(x,y;w) = \sum_{k,\ell} y_{k,\ell} \left[w^T \varphi(x^E_k, x^F_\ell) \right]$

let $y \in \{0,1\}^{L_E \times L_F}$

let matrix F be $d \times L_E \cdot L_F$ $\quad \left[\cdots \varphi_{k\ell} \cdots \right]$

$s(x,y;w) = w^T F y$

$S(x^{(i)}, \tilde{y}; w) = w^T F_i \tilde{y}$

decoding: $h_w(x^{(i)}) = \underset{\tilde{y} \in \mathcal{Y}_i}{\arg\max}\; s(x^{(i)}, \tilde{y}; w) \overset{\text{becomes}}{\leadsto} \underset{\substack{y_{k\ell} \in \{0,1\} \\ y \in M_i}}{\max} w^T F_i y$ linear integer program

$M_i \triangleq \left\{ y \in \mathbb{R}^{L^{(i)}_E \times L^{(i)}_F} : \begin{array}{l} \sum_\ell y_{k,\ell} \le 1 \;\forall k \\ \sum_k y_{k,\ell} \le 1 \;\forall \ell \\ 0 \le y_{k\ell} \le 1 \end{array} \right\}$ matching constraints

constraints $L^{(i)}_E + L^{(i)}_F$

$M_i = \{ z : Az \le \mathbb{1},\; z \ge 0 \}$

$$A : \begin{array}{cccc} y_{11} & y_{12} & y_{21} & y_{22} \end{array}$$
$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ & & I & \end{pmatrix}$$

$(2L + L^2) \times L^2$

(*) here, turns out that can remove integer constraint to get a "relaxed LP" give the same opt. obj. value

i.e. relaxation is tight

[actually here, M_i = convex-hull$(\mathcal{Y}_i)$]

why is relaxation tight?

write $z \in M_i$ as $\begin{array}{l} Az \le b_i \\ z \ge 0 \end{array}$; matrix A here is "totally unimodular"

which means that any subdeterminant of A has value $\begin{cases} +1 \\ -1 \\ 0 \end{cases}$

$\Rightarrow$ that if b has integer entries then all vertices of $\{z : Az \le b,\; z \ge 0\}$ have integer coordinates

$\Rightarrow$ relaxation is tight for any linear cost

idea: $\tilde{A} z \le \tilde{b}$, a corner of this is obtained by solving $\tilde{A}_{I:} \tilde{z} = \tilde{b}_I$ for $\tilde{A}_{I:}$ is invertible

$|I| = \dim(\tilde{z})$ $\Downarrow$ $\tilde{z}^* = \tilde{A}_{I:}^{-1} b_I$

use Cramer's rule
$\rightarrow$ ratio of subdeterminant
$\Rightarrow$ get integer values //

conclusion: can write decoding as $\max_{z \in M_i} w^T F_i z$

what about loss?

Hamming loss example

$$\ell(y, \tilde{y}) = \sum_{k,l} \mathbb{1}\{y_{kl} \neq \tilde{y}_{kl}\}$$

$$= \sum_{k,l} (y_{kl} - \tilde{y}_{kl})^2$$

$$(y_{kl}^2 - 2\tilde{y}_{kl} y_{kl} + \underbrace{\tilde{y}_{k,l}^2}_{=\tilde{y}_{kl}}) = (y_{kl}^2 + (1 - 2y_{kl})\tilde{y}_{kl})$$

cool trick:

$y^2 = y$

when $y \in \{0,1\}$

$$\ell_i(\tilde{y}) = \overbrace{a_i}^{\sum_{k,l} y_{kl}^{(i)}} + \underbrace{(1 - 2y^{(i)})}_{c_i}{}^T \tilde{y}$$

loss-augmented decoding:

$$\max_{\substack{\tilde{y} \in \{0,1\} \\ \tilde{y} \in M_i}} \underbrace{a_i + c_i^T \tilde{y}}_{\ell_i(\tilde{y})} - (\underbrace{w^T F_i y^{(i)} - w^T F_i \tilde{y}}_{w^T \psi_i(\tilde{y})})$$

$$\overbrace{a_i - w^T F_i y^{(i)}}^{\tilde{c}_i} + \max_{\substack{\tilde{y}_{k,l} \in \{0,1\} \\ \tilde{y} \in M_i}} (F_i^T w + c_i)^T \tilde{y}$$

$$= \max_{z \in M_i} \underbrace{(F_i^T w + c_i)^T z + a_i - w^T F_i y^{(i)}}_{g_i(w; z)}$$

LP duality:

$$\max_{\substack{A_i z \leq b_i \\ z \geq 0}} \tilde{c}_i^T = \min_{\substack{A_i^T \nu \geq \tilde{c}_i \\ \nu \geq 0}} b_i^T \nu \qquad \max_{z \in M_i} g_i(w; z) = \min_{\nu \in V_i(w)} \tilde{g}_i(w; z)$$

here $\tilde{c}_i \triangleq F_i^T w + c_i$

A_i is $(2L + L^2) \times L^2$

SVM struct obj. becomes (small QP)

$$\min_{w} \min_{\{v_i\}_{i=1}^n} \frac{1}{2}\lambda\|w\|^2 + \frac{1}{n}\sum_{i=1}^n \left[a_i - w^T F_i y^{(i)} + b_i^T v_i\right]$$

$$\text{s.t. } A_i^T v_i \geq F_i^T w + c_i$$

$$v_i \geq 0$$

"small complicated QP"

Compare with saddle pt. formulation

$$\min_{w} \max_{\{z_i\}_{i=1}^n} \frac{1}{2}\lambda\|w\|^2 + \frac{1}{n}\sum_{i=1}^n \left[a_i - w^T F_i y^{(i)}\right] + \left[(F_i^T w + c_i)^T z_i\right]$$

$$z_i \in M_i$$

simpler constraints