Lecture 12 - structured SVM

Thursday, February 14, 2019 13:35

today: SVMstruct → small QP formulation

sidenote on saddle point optimization:

necessary stationary conditions on (w^*, z^*)

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

directional derivative of $\mathcal{L}$ with respect to w in $w-w^*$ direction is non-neg

$$\begin{cases} \langle \nabla_w \mathcal{L}(w^*, z^*), w-w^* \rangle \geq 0 & \forall w \in \mathcal{W} \\ \langle -\nabla_z \mathcal{L}(w^*, z^*), z-z^* \rangle \geq 0 & \forall z \in \mathcal{Z} \end{cases}$$

let $u \triangleq (w,z)$ $\quad F: \mathbb{R}^p \to \mathbb{R}^p$

$$F(u) \triangleq \begin{pmatrix} \nabla_w \mathcal{L}(w,z) \\ -\nabla_z \mathcal{L}(w,z) \end{pmatrix}$$

becomes

$$\boxed{\langle F(u^*), u-u^* \rangle \geq 0 \quad \forall u \in \mathcal{U} = \mathcal{W} \times \mathcal{Z}}$$

"variational inequality"

generalizes:
- minimization
- saddle point
- equilibrium problems
- multi-player games or non zero sum 2-player game

extragradient algorithm

$$\tilde{u}_{t+1} = u_t - \gamma_t F(u_t)$$

$$u_{t+1} = u_t - \gamma_t F(\tilde{u}_{t+1})$$

→ solves variational inequality

b) small "complicated" QP formulation (for structured SVM)

$$H_i(w) = \max_{z_i \in Z} g_i(w, z_i) = \min_{v_i \in V_i(w)} \hat{g}_i(w; v_i)$$

use duality

dual variables

convex dual of $\max_{z_i \in Z} g_i(w, z_i)$

obtain: $\min_{w \in W} \min_{\forall i,\, v_i \in V_i(w)} \frac{\lambda \|w\|^2}{2} + \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(w; v_i)$

if $\hat{g}_i$ is jointly convex in w & v_i

we get a "tractable" convex min. problem

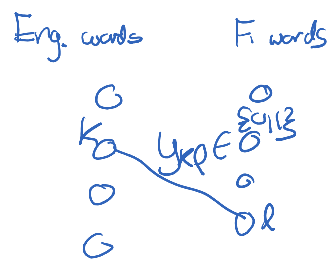
→ can solve with favorite convex min. alg.;

if d & n not too big, use interior point solver

e.g. Mosek, Cplex (commercial)

CVXopt (free python)

examples of $g_i(w; z_i)$



I) word alignment:

features on pair (x_k^E, x_ℓ^F)

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

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

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

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

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

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

$h_w(x^{(i)}) = \underset{y \in \mathcal{Y}_i}{\arg\max}\ s(x^{(i)}, y; w)$

decoding / inference problem

$\max\ w^T F_i y$
$y_{k\ell} \in \{0,1\}$
$y \in M_i$

matching constraints

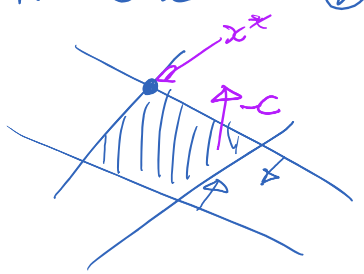
$M_i = \{ y \in \mathbb{R}^{L_E^{(i)} \cdot L_F^{(i)}} : \sum_\ell y_{k\ell} \leq 1,\ 0 \leq y_{k\ell} \leq 1 \}$

$\sum_k y_{k\ell} \leq 1,$

(# of constraints $L_E^{(i)} + L_F^{(i)}$)

"linear integer program"

$\min\ c^T x$



(*) here, turns out can remove the integer constraint to get a "relaxed LP" give the same optimal objective value

i.e. relaxation is tight!

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

14h36

reasons that relaxation is tight:

a) write $y \in M_i$ as $Ay \leq b_i$ — matrix A here is "totally unimodular"

a) write $y \in M_i$ as $\begin{array}{l} Ay \leq b_i \\ y \geq 0 \end{array}$

matrix A here is "totally unimodular"
which means that
any subdeterminant of A has value $\{+1, -1, 0\}$

$\Rightarrow$ that if b has integer entries

then all vertices of $\{y : Ay \leq b, y \geq 0\}$

have integer coordinates

$\Rightarrow$ relaxation is tight for any linear costs

idea: $\tilde{A}\tilde{y} \leq \tilde{b}$, a corner of this polytope is obtained by solving

$\tilde{A}_{I:}\tilde{y} = b_I$ for $\tilde{A}_{I:}$ invertible

$|I| = \dim(y)$

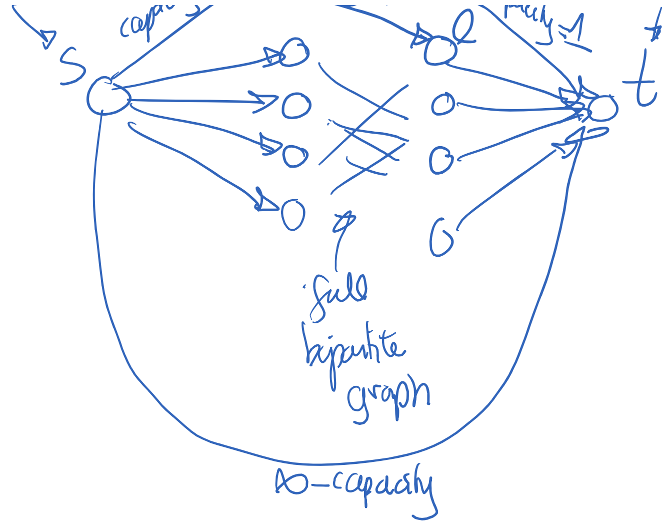
$\tilde{y} = \tilde{A}_I^{-1} b_I$

Cramer's rule : ratio of subdeter.
$\Rightarrow$ integers etc...

//

b) equivalent to min cost network flow problem with integer capacities

[which has integer solutions by min-cut/max-flow thm.]



<u>Conclusion:</u> can write decoding as $\max_{z \in \mathcal{M}_i} w^T F_i z$

<u>What about loss?</u>

<u>Hamming loss example:</u>

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

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

$$= \sum_{kl} \Big(y_{kl}^2 - 2 y_{kl} \tilde{y}_{kl} + \underbrace{\tilde{y}_{kl}^2}_{= \tilde{y}_{kl}} \Big)$$

$$\ell_i(\tilde{y}) = a_i + \underbrace{(\mathbb{1} - 2y^{(i)})^T}_{c_i} \tilde{y}$$

<u>cool trick:</u> $y^2 = y$ when $y \in \{0,1\}$

$$\underbrace{}_{c_i}$$

Loss augmented inference becomes:

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

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

LP duality:

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

$$\max_{\substack{A_i z \leq b_i \\ z \geq 0}} \tilde{c}_i^T z = \min_{\substack{A_i^T v \geq \tilde{c}_i \\ v \geq 0}} b_i^T v$$

$$\max_{z \in M_i} g_i(w, z) = \min_{v \in V_i(w)} \hat{g}_i(w; v)$$

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

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

SVM struct objective becomes

$$\boxed{\min_{w} \min_{\substack{\{v_i\}_{i=1}^n \\ A_i^T v_i \geq F_i^T w + c_i \\ v_i \geq 0}} \frac{\lambda \|w\|^2}{2} + \frac{1}{n} \sum_{i=1}^{n} \left[a_i - w^T F_i y^{(i)} + b_i^T v_i\right]}$$

"small complicated QP"

$$\underbrace{}_{A' \binom{w}{v} \leq b'}$$

compare with saddle pt. formulation

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

$\leftarrow$ simpler constraints

$$\underset{\substack{w \\ Z_i \in M_i}}{\min} \ \ldots \ \frac{1}{n}\sum_{i=1} \ldots$$

$Z_i \in M_i$ ← simplex constraints

⊛ note: primal error vs. dual error not the same!

II) M^3net example (CRF score)

* graph $G = (V, E)$ $\quad y_p,\ p \in V \quad y_c \triangleq (y_p)_{p \in C}$

  set of cliques $\mathcal{C}$

  MRF model

$$p(y|x;w) = \frac{1}{Z(x;w)} \prod_{c \in \mathcal{C}} \Psi_c(y_c; x, w)$$

$$= \frac{1}{Z} \exp\Big(\overbrace{\sum_c \underbrace{\log \Psi_c(y_c; x, w)}_{w^T \varphi_c(y_c;x)}}^{s(x,y;w)}\Big)$$

example: OCR
→ chain graph on y

inference: $\max\limits_{\tilde{y} \in \mathcal{Y}} \sum\limits_{c \in \mathcal{C}} w^T \varphi_c(y_c; x)$

goal: let's rewrite as a LP

* fix x & w; let s be a vector of size $|\mathcal{Y}(x)|$ of scores $\qquad e_k = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{pmatrix}$ ← k^{th} position

$$\max_{\tilde{y} \in \mathcal{Y}(x)} s(x, \tilde{y}; w) = \max_k s_k = \max_{\{e_k\}} \langle e_k, s \rangle$$

tight → $= \max \langle \qquad\quad \langle \alpha, s \rangle$

$y \in \mathcal{Y}(x)$ $\quad k \quad \{e_k\}$

trick? (tight convex cts. relaxation) $\overset{?}{=} \max\limits_{\alpha \in \underbrace{\mathrm{conv}(\{e_k\})}_{\cong \Delta_{|\mathcal{S}|}}} \langle \alpha, s \rangle$

note any comb. opt. problem
can be written as a convex opt. (LP) [but with exp #. of dimensions]

much smaller problem

$\star$ insight of M^3-net paper $\longrightarrow$ use MRF structure to transform $\max\limits_{\alpha \in \Delta_{|\mathcal{S}|}} \langle \alpha, s \rangle = \max\limits_{\mu \in \mathcal{M}} \sum\limits_{c} \langle \mu_c, \xi_c \rangle$