Lecture 15 - cutting plane alg.

Tuesday, March 10, 2020 14:21

today: • more SVMstruct properties
• $M^3$-net dual
• cutting plane alg.

## more properties of SVMstruct dual

primal-dual gap:

$$gap(\alpha) = p(w(\alpha), \xi(\alpha)) - d(\alpha) \geq 0$$

certificate of subopt. (*)

$$= \underbrace{p(w(\alpha), \xi(\alpha)) - p(w^*, \xi^*)}_{\text{primal subopt. } w(\alpha)} + \underbrace{p(w^*, \xi^*)}_{= d(\alpha^*)} - d(\alpha)$$

$d(\alpha^*)$: dual subopt.

$$\rightarrow \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} H_i(w) + \frac{\lambda}{2}\|w\|^2 - \frac{1}{n}\sum_{i=1}^{n}\sum_{\tilde{y}} a_i(\tilde{y})\, l_i(\tilde{y})$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[H_i(w) - \sum_{\tilde{y}} a_i(\tilde{y})\psi_i(\tilde{y})\right] + \lambda \langle w, w(\alpha)\rangle$$

$\lambda \langle w, w(\alpha)\rangle = \frac{1}{n}\sum_{i}\sum_{\tilde{y}} a_i(\tilde{y})\langle \psi_i(\tilde{y}), w\rangle$

$$gap(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left[H_i(w(\alpha)) - \sum_{\tilde{y}} a_i(\tilde{y})\, H_i(\tilde{y}; w(\alpha))\right]$$

$H_i(\tilde{y}; w(\alpha)) = l_i(\tilde{y}) - w^T\psi_i(\tilde{y})$

$\max_{\tilde{y}} H_i(\tilde{y}; w)$

← use to get a bound on dual subopt. of $\alpha$

$$w(\alpha) = \frac{1}{\lambda n}\sum_{i=1}^{n}\left(\sum_{\tilde{y}} a_i(\tilde{y})\,\psi_i(\tilde{y})\right)$$

let $R_i \triangleq \max_{\tilde{y}} \|\psi_i(\tilde{y})\|_2$

$\bar{R} \triangleq \frac{1}{n}\sum_{i=1}^{n} R_i$

then 1) $\|w^*\|_2 \overset{\Delta\text{-ineq.}}{\leq} \frac{1}{\lambda}\frac{1}{n}\sum_i \underbrace{\sum_{\tilde{y}} a_i(\tilde{y})}_{=1} \underbrace{\|\psi_i(\tilde{y})\|_2}_{\leq R_i}$

$$\leq \frac{1}{\lambda}\left(\frac{1}{n}\sum_i R_i\right) = \frac{\bar{R}}{\lambda}$$

2) kernel trick: $\langle w(\alpha), \varphi(x,y)\rangle = \frac{1}{\lambda n}\sum_i\sum_{\tilde{y}} a_i(\tilde{y})\langle \psi_i(\tilde{y}), \varphi(x,y)\rangle$

$\psi_i(\tilde{y}) = \varphi(x_i, y_i) - \varphi(x_i, \tilde{y})$

$\langle \psi_i(\tilde{y}), \varphi(x,y)\rangle = k(x_i, y_i; x, y)$

$\|w(\alpha)\|^2 \rightarrow \alpha^T K \alpha$ $\quad - k(x_i, \tilde{y}; x, y)$

↳ $K_{iy;jy'} \triangleq \langle \psi_i(y), \psi_j(y') \rangle$

3) suppose scale features $\boxed{\tilde{\psi} \triangleq b\psi}$

$\tilde{H}_i(\tilde{y}; w) = \ell_i(\tilde{y}) - \langle \tilde{w}, \tilde{\psi}_i(\tilde{y}) \rangle$

$\tilde{w}^* = \frac{1}{\tilde{\lambda}} \frac{1}{n} \sum_i \sum_{\tilde{y}} \tilde{\alpha}_i^*(\tilde{y}) \underbrace{\tilde{\psi}_i(\tilde{y})}_{b\psi_i(\tilde{y})}$

let $\boxed{\tilde{\lambda} = b^2 \lambda}$ $\quad \tilde{w}^*(\tilde{\alpha}^*) = \frac{1}{b}\left[\frac{1}{\lambda}\frac{1}{n}\sum_{i,\tilde{y}} \tilde{\alpha}_i^*(\tilde{y})\psi_i(\tilde{y})\right]$

if use $\tilde{\alpha}^* = \alpha^* \Rightarrow \tilde{w}^*(\tilde{\alpha}^*) = \frac{w^*}{b}$

$\Rightarrow \tilde{H}_i(\tilde{y}; \tilde{w}^*) = \ell_i(\tilde{y}) - \langle \frac{w^*}{b}, b\psi_i(\tilde{y}) \rangle = H_i(\tilde{y}; w^*)$

ie. $\tilde{\alpha}^*$ is really optimal for new problem with $\tilde{\psi}, \tilde{\lambda}$?

4) similarly, can show $\tilde{\ell} = b \cdot \ell \Rightarrow \tilde{\lambda} = \frac{\lambda}{b}$ get same solution

**M³-net example (dual):** (getting small dual)

$w(\alpha) = A\alpha = \sum_i A_i \alpha_i$ $\quad$ suppose $\psi(y) = \sum_c \psi_c(y_c)$

$\alpha_i \in \Delta_{|\mathcal{Y}_i|}$

$\lambda n A_i \alpha_i = \sum_{\tilde{y}} \alpha_i(\tilde{y}) \psi_i(\tilde{y}) = \sum_{\tilde{y}} \alpha_i(\tilde{y}) \sum_c \psi_{i,c}(\tilde{y}_c)$

$= \sum_c \sum_{\tilde{y}_c} \psi_{i,c}(\tilde{y}_c) \underbrace{\left[\sum_{y' \text{ s.t. } y'_c = \tilde{y}_c} \alpha_i(y')\right]}_{\triangleq \mu_{i;c}(\tilde{y}_c)}$

"marginal variables"

$\alpha_i \in \Delta_{|\mathcal{Y}_i|} \Rightarrow \mu_i \in \mathcal{M}_i$

marginal polytope

thus $A_i \alpha_i = \tilde{A}_i \mu_i$ where $(\tilde{A})_{:,c,\tilde{y}_c} = \frac{\psi_{i,c}(\tilde{y}_c)}{\lambda n}$

↳ # of columns is $\sum_c |\mathcal{Y}_c|$

similarly, suppose $\ell(y) = \sum_c \ell_{i,c}(y_c)$

define $\tilde{b}_{i,c}(y_c) \triangleq \frac{b_{i,c}(y_c)}{n}$

$\langle b_i, \alpha_i \rangle = \langle \tilde{b}_i, \mu_i(\alpha_i) \rangle$

(*) thus we can replace

$$\max_{\alpha_i \in \Delta_{|\mathcal{Y}_i|}} -\lambda \frac{\|A\alpha\|^2}{2} + b^T\alpha \quad \text{with} \quad \boxed{\max_{\mu_i \in M_i} -\lambda \frac{\|\tilde{A}\mu\|^2}{2} + \tilde{b}^T\mu}$$

→ this is a tractable size QP if $M_i$ is tractable

M3-net paper

used "structured SMO" algorithm

block-coordinate ascent using pair of variables on this QP [similar to "pairwise FW"]

If $G_i$ is triangulated then $M_i = L_i$ (local consistency polytope)

15h23

**Constraint generation alg.:**

[Tsochantaridis et al. JMLR 2005]

want to solve $\min_{w,\xi} \frac{\lambda\|w\|^2}{2} + \frac{1}{n}\sum_{i=1}^n \xi_i$ (P)

$$\max_{\alpha_i \in \Delta_{|\mathcal{Y}_i|}} -\lambda\|A\alpha\|^2 + b^T\alpha \quad \text{(D)}$$

n-slack version

s.t. $\xi_i \geq H_i(\vec{y}, w) \quad \forall \vec{y} \in \mathcal{Y}_i$
$\xi_i \geq 0$

esp # of constraints $\sum_i |\mathcal{Y}_i|$

s.t. $\alpha_i \in \Delta_{|\mathcal{Y}_i|}$ — # variables

Vs,

1-slack version [ML 2009 paper]

$\min_{w,\xi} \frac{\lambda\|w\|^2}{2} + \xi$ (P)

s.t. $\xi \geq \frac{1}{n}\sum_{i=1}^n H_i(\vec{y}_i, w) \quad (\forall \vec{y}_i \in \mathcal{Y}_i)_{i=1,\dots,n}$

$\prod_i |\mathcal{Y}_i|$

$\frac{1}{n}\sum_{i=1}^n \delta_i(\vec{y}_i) - \langle w, \frac{1}{n}\sum_{i=1}^n \psi_i(\vec{y}_i)\rangle$

$\max_{\alpha \in \Delta_{|\prod_i \mathcal{Y}_i|}} -\lambda\frac{\|A\alpha\|^2}{2} + \langle b, \alpha\rangle$ (D)

$w(\alpha) = \frac{1}{\lambda n}\sum_{(\vec{y}_1,\dots,\vec{y}_n) \in \prod_i \mathcal{Y}_i} \alpha(\vec{y}_1,\dots,\vec{y}_n) \sum_{i=1}^n \psi_i(\vec{y}_i)$

$O(d)$ instead of $O(d \cdot n)$ in n-slack formulation ⇒ big memory saving

**n-slack SVM struct alg.:**

iterate solving QP with more & more constraints

1) start with no constraint on $w$ $\Rightarrow$ $w^{(0)}=0$, $\xi^{(0)}=0$

2) repeat: for each $i$, find $\hat{y}_i = \arg\max_{\tilde{y}\in\mathcal{Y}_i} H_i(\tilde{y}; w^{(t)})$ [loss-augmented decoding]

- add $\xi_i \geq H_i(\hat{y}_i; w)$ constraint to QP (if not already there)

↳ then resolve QP$(w,\xi)$ with these constraints to get $w^{(t+1)}, \xi^{(t+1)}$ [e.g. using CVXopt.]

stop when primal-dual gap $< \varepsilon$ } takes $O(n)$ time

[in 2005, showed that poly. step after $O(\frac{1}{\varepsilon^2})$ iterations]

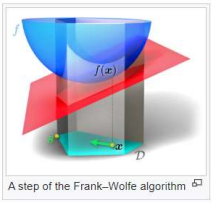
refined later to $O(\frac{1}{\varepsilon})$ ← at least for 1-slack version [2009]

## Frank-Wolfe algorithm

↳ for smooth constrained opt. [motivation in our context: dual of SVM start $\min_{\alpha_i \in \Delta_{|\mathcal{Y}_i|}} \underbrace{\frac{\lambda}{2}\|A\alpha\|^2 - b^T\alpha}_{\geq}$ ]

1940s: simplex alg. to solve LPs

1956: Marguerite Frank & Phil Wolfe → non-linear opt. by iterated LPs



A step of the Frank–Wolfe algorithm

setup: $\min f(x)$ s.t. $x \in M$

- $f$ is $L$-smooth ie. ($\nabla f$ is $L$-Lipschitz) and $f$ is convex
- $M$ is convex and bounded set

## FW algorithm

by convexity: $f(s) \geq f(x_t) + \langle \nabla f(x_t), s - x_t \rangle \quad \forall s \in M$ (linear approx of $f$ at $x_t$)

start with $x_0 \in M$

for $t = 0, \ldots$

"linear minimization oracle" LMO

FW corner

min. RHS w.r.t. $s$

compute $s_t = \arg\min_{s\in M} \langle s, \nabla f(x_t)\rangle$

stopping criterion

[let $g_t \triangleq \langle s_t - x_t, -\nabla f(x_t)\rangle$ FW gap if $g_t \leq \varepsilon$; output $x_t$ ]

$x_{t+1} = (1-\gamma_t)x_t + \gamma_t s_t$ $\quad \gamma_t \in [0,1]$ (convex combo)

step-size

$= x_t + \gamma_t(s_t - x_t)$

$x_{t+1} = (1-\alpha_t)x_t + \alpha_t z_t \qquad \alpha_t \in [0,1]$ (convex combo)

$= x_t + \gamma_t(\underbrace{z_t - x_t}_{d_t})$

end
output $x_t$

step-size choice: $\gamma_t =$
- universal choice $\frac{2}{t+2}$
- line search $\gamma_t = \arg\min_{\gamma \in [0,1]} f(x_t + \gamma(z_t - x_t))$
- adaptive: $\left[ \frac{g_t}{L\|z_t - x_t\|^2} \text{ or } \frac{g_t}{C_f} \right]$ truncate at 1
  - $L$: affine invariant const.

⊛ big motivation for FW
is LMO is often much cheaper than projections
and cheap for many sets $M$ appearing in ML

properties:

1) $f(x_t) - \min_{x \in M} f(x) \leq O(\frac{1}{t})$, $\min_{x \in M} f(x) = f^*$

2) FW-gap $g_t \geq f(x_t) - f^* \rightarrow$ certificate of subopt.
$\min_{s \leq t} g_s \leq O(\frac{1}{t})$ [i.e. we will stop in $O(\frac{1}{\varepsilon})$ iterations?]

3) $x_t = p_0^t x_0 + \sum_{u=1}^{t} p_u^t s_{u-1}$ $\rightsquigarrow x_t$ has "sparse" expansion in terms of the FW-corners $\{s_u\}_{u=1}^{t-1}$
where $\sum_{u=0}^{t} p_u^t = 1$, $p_u^t \geq 0$

⊛⊛ "sparse method" → popular in ML

→ see later how to apply on dual of suboptimality

4) FW is affine covariant (like Newton's method)

5) there is $\Omega(\frac{1}{t})$ lower bound for FW-like methods for $t \leq d$