Lecture 15 - cutting plane alg.

Thursday, March 11, 2021 13:27

Today:
- more SVMstruct properties
- M^3-net dual
- cutting plane alg.
- FW alg.

more properties of SVMstruct dual

primal-dual gap

$p(w) - d(\alpha) \geq 0 \quad \forall w, \alpha$ feasible

$p(w) \geq p(w^*) = d(\alpha^*) \geq d(\alpha)$

Certificate of primal or dual suboptimality

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

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

$$= \underbrace{\frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^n H_i(w(\alpha))}_{p(w(\alpha),\xi(\alpha))} + \underbrace{\frac{\lambda}{2}\|w\|^2 - \frac{1}{n}\sum_{i,\tilde{y}} a_i(\tilde{y}) l_i(\tilde{y})}_{-d(\alpha)}$$

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

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

$$\boxed{gap(\alpha) = \frac{1}{n}\sum_{i=1}^n \left[H_i(w(\alpha)) - \sum_{\tilde{y}} a_i(\tilde{y}) \underbrace{H_i(\tilde{y}; w(\alpha))}_{l_i(\tilde{y}) - w^T\psi_i(\tilde{y})} \right]}$$

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

← use to get a bound on dual subopt. of α

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

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

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

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

$\psi_i(\tilde{y}) = \varphi(x^{(i)}, y^{(i)}) - \varphi(x^{(i)}, \tilde{y})$

$K((x^{(i)}, y^{(i)}), (x,y)) - K((x^{(i)}, y^{(i)}), (x^{(i)}, \tilde{y}))$

$\|w(\alpha)\|^2 \rightsquigarrow \alpha^T K \alpha$

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

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

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

$\tilde{w}(\tilde{\alpha}) = \frac{1}{\tilde{\lambda}} \frac{1}{n} \sum_i \sum_{\tilde{y}} \tilde{\alpha}_i(\tilde{y}_i) \tilde{\Psi}_i(\tilde{y}_i)$

$b \Psi_i(\tilde{y}_i)$

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

if you use $\tilde{\alpha}_i \triangleq \alpha_i^*$

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

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

$\Rightarrow \tilde{\alpha}_i$ is really optimal for new problem with $\tilde{\Psi}$ & $\tilde{\lambda}$

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

M^3-net example (dual) & (get Hing small dual)

$w(\alpha) = A\alpha = \sum_i A_i \alpha_i$ $\quad$ suppose $\Psi(\tilde{y}) = \sum_c \Psi_c(\tilde{y}_c)$

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

$(\mathbb{R}^n)$ $A_i \alpha_i = \sum_{\tilde{y}} \alpha_i(\tilde{y}_i) \Psi_i(\tilde{y}_i) = \sum_{\tilde{y}} \alpha_i(\tilde{y}_i) \sum_c \Psi_{i,c}(\tilde{y}_c)$

$= \sum_c \sum_{\tilde{y}_c} \Psi_{i,c}(\tilde{y}_c) \left[\sum_{y' \text{ s.t. } y'_c = \tilde{y}_c} \alpha_i(\tilde{y}_i) \right]$

$\triangleq \mu_{i,c}(\tilde{y}_c)$ "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 $(\tilde{A}_i)_{o,(c,\tilde{y}_c)} = \frac{\Psi_{i,c}(\tilde{y}_c)}{n}$

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

similarly, suppose $\ell_i(\tilde{y}_i) = \sum_c \ell_{i,c}(\tilde{y}_c)$

define $\tilde{b}_{i,c}(\tilde{y}_c) \triangleq \frac{\ell_{i,c}(\tilde{y}_c)}{n}$ $\Rightarrow$ $\langle b_i, \alpha_i \rangle = \langle \tilde{b}_i, \mu_i(\alpha_i) \rangle$

(*) thus replace replace

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

$\hookrightarrow$ this is a tractable size QP if $\mathcal{M}_i$ is tractable

→ this is a tractable size QP if M_i is tractable

M^3-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)

14h29

constraint generation alg.:

[Tsochantaridis & al. JMLR 2005]

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

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

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

$\max -\frac{\lambda}{2}\|A a\|^2 + b^T a$ (D)

s.t. $a_i \in \Delta_{|\mathcal{Y}_i|}$

variables

n-slack version

vs.

1-slack version

[ML 2009 paper]

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

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

# of constraints $\to \prod_i |\mathcal{Y}_i|$

$\max -\frac{1}{2}\|A a\|^2 + b^T a$

$a \in \Delta_{|\prod_{i=1}^n \mathcal{Y}_i|}$

$w(a) = \frac{1}{\lambda n} \sum_{\tilde{y}_{1:n} \in \prod_{i=1}^n \mathcal{Y}_i} a(\tilde{y}_{1:n}) \left(\sum_{i=1}^n \psi_i(\tilde{y}_i)\right)$

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

$O(d)$ to store

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

n-slack SVMstruct alg.: cutting plane / constraint generation

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

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

[in 2005, show that alg. stop after $O(\frac{1}{\varepsilon^2})$ iterations]

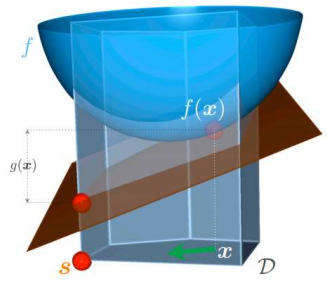
refined later [2009] to $O(\frac{1}{\varepsilon})$ for 1-slack version

Frank-Wolfe algorithm

↳ for smooth constrained opt. [motivation in our context: dual of SVM struct $\min_{\alpha_i \in \Delta_{|Y_i|}} \frac{\lambda \|A\alpha\|^2}{2} - b^T\alpha$]

1940s: simplex alg. to solve LPs

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



Setup: $\min_{s.t.\ x \in M} f(x)$

- f is L-smooth
- M is convex and bounded set

and assume we can solve effectively $\min_{s \in M} \langle s, d \rangle$ for any d — LMO

FW algorithm

start with $x_0 \in M$

for $t = 0, \dots,$ compute $s_t = \underset{s \in M}{\operatorname{argmin}} \langle s, \nabla f(x_t) \rangle$ — FW corner; "linear minimization oracle" LMO

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

min RHS w.r.t. s

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

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

$= x_t + \gamma_t (s_t - x_t)$, with $d_t = s_t - x_t$

end

output x_t

step size choice: $\gamma_t = \begin{cases} \text{universal} & \frac{2}{t+2} \\ \text{line search} & \gamma_t = \underset{\gamma \in [0,1]}{\operatorname{argmin}} f(x_t + \gamma(s_t - x_t)) \end{cases}$

adaptive choice: truncated $\left[\frac{g_t}{L\|s_t - x_t\|^2} \text{ or } \frac{g_t}{C} \right]$

(*) big motivation for FW

is LMO is often much cheaper than projections.

is LMO is often much cheaper
than projections
and cheap for many structured M appearing
in ML

adaptive choice: truncated of 1 $\left[\frac{g_t}{L\|x_t - x_0\|^2} \text{ or } \frac{g_t}{C_t} \right]$ affine invariant Const.