Lecture 15 - cutting plane alg.

Tuesday, March 14, 2023 2:28 PM

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

more properties of SVMstruct dual

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

Certificate of primal or dual suboptimality

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

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

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

$$= \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_i H_i(w(\alpha)) + \frac{\lambda}{2}\|w\|^2 - \frac{1}{n}\sum_{i,\tilde{y}} \alpha_i(\tilde{y}) \ell_i(\tilde{y})$$

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

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

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

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

use this to get a bound dual subopt. of α

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

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

$\bar{R} = \frac{1}{n}\sum_i R_i$

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

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

2) kernel trick: $\langle w(\alpha), \varphi(x,y) \rangle = \frac{1}{\lambda n}\sum_i \sum_{\tilde{y}} \alpha_i(\tilde{y}) \langle \underbrace{\psi_i(\tilde{y})}_{\varphi(x^{(i)}, y^{(i)}) - \varphi(x^{(i)}, \tilde{y})}, \varphi(x,y) \rangle$

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

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

$\hookrightarrow K_{i,y;j,y'} \triangleq \langle \psi_i(\tilde{y}), \psi_j(y') \rangle$

$\hookrightarrow K_{i,y;j,y'} \triangleq \langle \psi_i(\tilde{y}), \psi_j(y') \rangle$

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

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

$$\tilde{w}(\tilde{\alpha}) = \frac{1}{\tilde{\lambda}} \frac{1}{n} \sum_{i,\tilde{y}} \tilde{\alpha}_i(\tilde{y}) \underbrace{\tilde{\psi}_i(\tilde{y})}_{b\psi(\tilde{y})}$$

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

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

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

$\Rightarrow \tilde{\alpha}_i$ 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) : (getting small dual)

$w(\alpha) = A\alpha = \sum_i A_i \alpha_i$

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

suppose $\psi(y) = \sum_c \psi_c(y_c)$

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

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

"marginal variable"

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

marginal polytope

thus $A_i \alpha_i = \tilde{A}_i \mu_i$ where

$$(\tilde{A})_{:,(c,\tilde{y}_c)} = \frac{\psi_{i,c}(\tilde{y}_c)}{\lambda n}$$

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

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

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

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

→ this is a tractable$^{\text{size}}$ QP

if M_i are tractable

if G_i is triangulated then $M_i = L_i$ (local consistency polytope) ⇒ tractable

M^3-net paper:

used "structured SMO algorithm"

block-coordinate ascent using pairs of variables on this QP

[similar to "pairwise FW"]

14h27

constraint generation alg.:

[Tsochantaridis & al. JMLR 2005]

want to solve $\min_{w,\xi} \frac{\lambda\|w\|^2}{2} + \frac{1}{n}\sum_i \xi_i$ (P) $\max_{\alpha} -\frac{\lambda}{2}\|A\alpha\|^2 + b^T\alpha$ (D)

n-slack version

s.t. $\xi_i \geq H_i(\tilde{y};w)\ \forall \tilde{y} \in \mathcal{Y}_i$ } exp # of constraints → # variables

$\xi_i \geq 0$

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

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

vs.

1-slack version

[ML 2009 paper]

$\min_{w,\xi} \frac{\lambda\|w\|^2}{2} + \xi$ (P) $\max -\frac{\lambda}{2}\|A\alpha\|^2 + b^T\alpha$ (D)

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

$\alpha \in \Delta_{|\times_{i=1}^n \mathcal{Y}_i|}$

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

$w(\alpha) = \frac{1}{\lambda n}\sum_{\tilde{y}_{1:n} \in \times_{i=1}^n \mathcal{Y}_i} \alpha(\tilde{y}_{1:n}) (\varphi$

$\frac{1}{n}\sum_i \ell_i(\tilde{y}_i) - \frac{1}{n}\langle w, \underbrace{\sum_{i=1}^n \psi_i(\tilde{y}_i)}_{O(d)\text{ to store}} \rangle$

$\sum_i \psi_i(\tilde{y}_i)$

$\frac{1}{n}\sum_i \ldots(y_i) \quad \frac{1}{n} \ldots \underbrace{\sum_{i=1} \ldots(y_i)}$

$O(d)$ to store

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

$\sum_{i=1}^{n} \psi_i(\tilde{y}_i)$

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

Iterate solving QP with more & more constraints

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

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

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

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

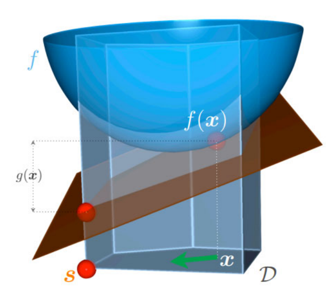
stop when primal-dual gap $\leq \varepsilon$

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

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

Frank-Wolfe algorithm

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



1940s: Simplex alg. to solve LPs

1956: Marguerite Frank & Phil Wolfe $\rightarrow$ non-linear opt. by iterating LPs

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

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

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

by convexity

$f(s) \geq \underbrace{f(x_t) + \langle \nabla f(x_t), s - x_t \rangle}_{\text{linear app. of } f \text{ at } x_t} \quad \forall s \in M$

$\min$

FW algorithm

start with $x_0 \in M$

for $t = 0, \ldots,$

FW corner

"linear minimization oracle" (LMO)

start with $x_0 \in M$

for $t = 0, \ldots,$

FW corner

linear minimization oracle (LMO)

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

min RHS w.r.t. s

linear app. of f at x_t

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]

step size

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

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

end

output x_t

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

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

truncate at 1

affine invariant const.

★ big motivation for FW

is LMO is often much cheaper than projections

and cheap for many structured M appearing in ML