Lecture 15 - cutting plane alg.

Thursday, March 14, 2024 9:35 AM

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

more properties of SVM struct dual

primal-dual gap

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

certificate of primal or dual suboptimality

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

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

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

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

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

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

$$\sum_{\tilde{y}} \left(-H_i(\tilde{y}; w(\alpha)) \right)$$

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

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

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

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

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

Let $R_i \triangleq \max_{\tilde{y}} \|\psi_i(\tilde{y})\|_2$ $\quad \bar{R} \triangleq \frac{1}{n}\sum_{i=1}^n R_i$

Then 1) $\|w^*\|_2^2 \overset{\text{triang}}{\leq} \frac{1}{\lambda n}\sum_{i,\tilde{y}} \alpha_i(\tilde{y}) \underbrace{\|\psi_i(\tilde{y})\|_2}_{\leq R_i}$

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

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

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

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

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

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

4) [illegible]

↳ $K_{i,y_i,j,y'} \triangleq \langle \psi_i(y), \psi_j(y') \rangle$

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

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

$$\tilde{\omega}(\tilde{\alpha}) = \frac{1}{\tilde{\lambda}} \frac{1}{n} \sum_i \sum_{\tilde{y}} \tilde{\alpha}_i(\tilde{y}) \frac{\tilde{\psi}_i(\tilde{y})}{b\psi_i(\tilde{y})}$$

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

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

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

$= 0$ as w^* is optimal

$\Rightarrow \tilde{\alpha}_i$ 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

<u>M^3-net example (dual)</u> : (getting small dual)

$$w(\alpha) = A\alpha = \sum_i A_i \alpha_i \qquad \alpha_i \in \Delta_{|\mathcal{Y}_i|}$$

suppose $\psi(\tilde{y}) = \sum_c \psi_c(\tilde{y}_c)$

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

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

"marginal variable"

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

marginal polytope

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

↳ # 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

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

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

If G_i is triangulated, then $\mathcal{M}_i = \mathcal{L}_i$ (local consistency polytope)

M^3-net paper: (sequential minimal optimization)

used "structured SMO algorithm"

block-coordinate ascent on

pair of variables on this QP

[similar to "pairwise FW"] (see next classes)

10h 28

Constraint generation alg.:

[Tsochantaridis & al. JMLR 2005]

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

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

$\xi_i \geq 0$

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

$\alpha_i \in \Delta_{|\mathcal{Y}_i|}$ → # variables

n-slack version

vs.

1-slack version

[ML 2009 paper]

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

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

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

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

$$\alpha \in \bigtriangleup_{|\prod_{i=1}^n \mathcal{Y}_i|}$$

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

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

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

instead of $O(d \cdot n)$ storage in n-slack formulation → 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 $\hat{y}_i = \underset{\tilde{y}\in\mathcal{Y}_i}{\text{argmax}}\ H_i(\tilde{y}; w^{(t)})$ [loss-augmented decoding]

- add $\xi_i \geq H_i(\hat{y}_i; w)$ constraint to QP (if not already there)
- then resolve QP(w,ξ) with these constraints to get $w^{(t+1)}, \xi^{(t+1)}$

stop when primal-dual gap $< \varepsilon$ [e.g. CVXopt]

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

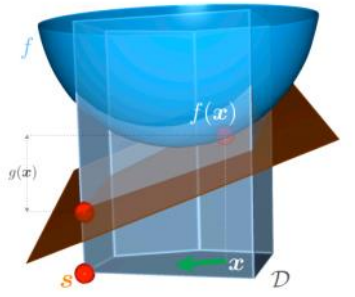
refined later [2009] to $O(\frac{1}{\varepsilon})$ for 1-slack formulation

Frank-Wolfe algorithm

↳ for smooth constrained opt. ("feasible pt. method") [motivation in our context: dual of SVMstruct $\min \frac{\lambda}{2}\|A\alpha\|^2 - b\alpha$ s.t. $\alpha_{iy}\geq 0, \sum_y \alpha_{iy}=1$]

1940s: simplex alg. to solve LPs

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



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

- f is L-smooth
- $\mathcal{M}$ is convex and bounded set

and assume we can solve efficiently $\min_{s\in\mathcal{M}} \langle s, d\rangle$ for any d ("linear minimization oracle" LMO)

FW algorithm

start with $x_0 \in \mathcal{M}$ — FW corner

for $t=0,\dots,$

compute $s_t = \underset{s\in\mathcal{M}}{\text{argmin}} \langle s, \nabla f(x_t)\rangle$ — LMO

[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.)

by convexity $f(s) \geq f(x_t) + \langle \nabla f(x_t), s - x_t\rangle$ ($\forall s$)

$f(x^*) \geq$ linear approx. of f at x_t

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

$= x_t + \gamma_t (s_t - x_t)$ — d_t

end

output x_t

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

adaptive choice: $\left[\frac{g_t}{L\|s_t - x_t\|^2} \text{ or } \frac{g_t}{C_f} \right]$ (truncate at 1)

C_f: affine invariant constant

big motivation for FW

is LMO is often much cheaper than projections

and cheap for many structured $\mathcal{M}$ appearing in ML