Lecture 9 - scribbles - SVMstruct dual

Friday, February 3, 2017
15:03

primal: $\min_{w,\xi_i} \quad \frac{\lambda\|w\|^2}{2} + \frac{1}{n}\sum_{i=1}^{n}\xi_i \Big\} p(w,\xi)$

$\xi_i \geq \underbrace{\ell_i(y) - w^T\psi_i(y)}_{\triangleq H_i(y;w)} \quad \forall y \in \mathcal{Y}_i$

dual: $\max_{\alpha_i \in \Delta_{|\mathcal{Y}_i|}} \quad -\frac{\lambda}{2}\|\overbrace{A\alpha}^{w(\alpha)}\|^2 + b^T\alpha \Big\} d(\alpha)$

$b \in \mathbb{R}^m \quad m = \sum_i m_i \quad m_i \triangleq |\mathcal{Y}_i|$

$b_i(y) \triangleq \frac{1}{n}\ell_i(y)$

$A \in \mathbb{R}^{d\times m} \quad \underset{[i,y^{th}\text{ column}]}{A_i(y)} \triangleq \psi_i(y)$

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

$\ell_i(y) \triangleq \ell(y^{(i)}, y)$

$H_i(w) = \max_{\tilde{y}\in\mathcal{Y}_i} H_i(\tilde{y};w)$

(also look: Danskin's thm)

subdifferential

$$\boxed{\begin{gathered}\partial_w H_i(w) = \mathrm{conv}\left(\{-\psi_i(y) : y \in \mathcal{Y}_i^*(w)\}\right) \\ \mathcal{Y}_i^*(w) = \underset{\tilde{y}\in\mathcal{Y}_i}{\arg\max}\ H_i(\tilde{y};w) \quad \text{as a set}\end{gathered}}$$

$\Leftarrow$ Let $v_i \in \partial_w H_i(w)$

$\Rightarrow v_i = -\sum_{y\in\mathcal{Y}_i^*(w)} \alpha_i(y)\psi_i(y)$ where $\alpha_i \in \Delta_{|\mathcal{Y}_i^*(w)|}$

$\longrightarrow$ so KKT like finding expansion for a zero subgradient of $p(w^*)$

KKT conditions:

$$\boxed{w^* = w(\alpha^*) = \frac{1}{\lambda n}\sum_i\sum_{y\in\mathcal{Y}_i} \alpha_i^*(y)\psi_i(y)}$$

$\xi_i^* = \xi_i(\alpha^*) \triangleq \max_y H_i(y;w^*)$
$= H_i(w^*)$

and if $H_i(y;w^*) < H_i(w^*)$

$\Rightarrow \alpha_i^*(y) = 0$

complementary slackness

not a support vector

primal-dual gap: (using "KKT mapping")

$\mathrm{gap}(\alpha) \triangleq p(w(\alpha),\xi(\alpha)) - d(\alpha)$
$\geq p(w(\alpha),\xi(\alpha)) - p(w^*,\xi^*)$

$= \frac{\lambda\|w\|^2}{2} + \frac{1}{n}\sum_i H_i(w) + \frac{\lambda\|w\|^2}{2} - \frac{1}{n}\sum_i\sum_y \alpha_i(y)\ell_i(y)$

$= \frac{1}{n}\sum_{i=1}^{n}\left[H_i(w) - \sum_y \alpha_i(y)\ell_i(y)\right] + \lambda\langle w, \underbrace{w(\alpha)}_{\frac{1}{\lambda n}\sum_i\sum_y \alpha_i(y)\psi_i(y)}\rangle$

$= \frac{1}{n}\sum_{i=1}^{n}\left[H_i(w) - \sum_y \alpha_i(y)\underbrace{\left[\ell_i(y) - \langle w, \psi_i(y)\rangle\right]}_{H_i(y;w)}\right]$

$$\boxed{\mathrm{gap}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{H_i(w)}_{\max} - \sum_{y\in\mathcal{Y}_i}\alpha_i(y)H_i(y;w)\right)}$$

Support of solution: $\alpha_i^*(y) > 0$ only for $y \in \mathcal{Y}_i^*(w^*(\alpha^*))$

$\psi_i(y)$ for $\alpha_i^*(y) > 0$, is called a "structured support vector" when $y \neq y^{(i)}$

"margin idea" is want $w^T\varphi(x^{(i)},y^{(i)}) \geq w^T\varphi(x^{(i)},\tilde{y}) + \ell_i(\tilde{y}) \quad \forall \tilde{y}$

predicting correctly: $w^T\varphi(x^{(i)},y^{(i)}) > \max_{\tilde{y}} w^T\varphi(x^{(i)},\tilde{y})$ $\Big\}$ still have $H_i(w) \geq 0$ is ...

with margin $\; > \max_{y \neq y^{(i)}} \left\{ w^T \varphi(x^{(i)}, y) + \ell_i(y) \right\} \Longrightarrow H_i(w) = 0$

binary SVM: $y \in \{\pm 1\}$, $\varphi(x^{(i)}, y) \triangleq \frac{y\, x^{(i)}}{2}$

$$\Psi_i(y) = y^{(i)} \frac{x^{(i)}}{2} - y \frac{x^{(i)}}{2} = \begin{cases} 0 & \text{if } y = y^{(i)} \\ y^{(i)} x^{(i)} & \end{cases}$$

$$H_i(y; w) = \begin{cases} 0 & \text{if } y = y^{(i)} \\ 1 - w^T \Psi_i(-y^{(i)}) = 1 - y^{(i)} w^T x^{(i)} & \end{cases}$$

$$H_i(w) = \left[1 - y_i \langle w, x_i \rangle\right]_+ \quad \text{(binary hinge loss)}$$

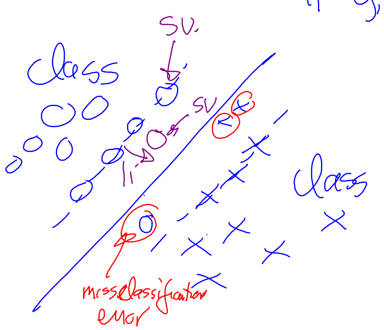
$$w^* = \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i y_i x_i \qquad \alpha_i \to \alpha_i(-y^{(i)}) \qquad 0 \le \alpha_i \le 1$$

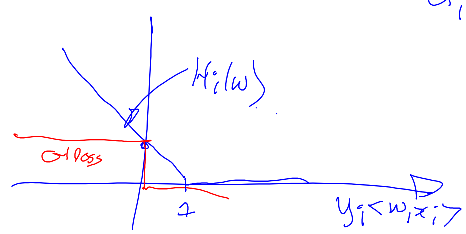
$$g(\alpha) = \frac{1}{n} \sum_{i=1}^{n} H_i(w) - \alpha_i \left(1 - y_i \langle w, x_i \rangle\right)$$

$H_i(w) \to (1 - y_i \langle w, x_i \rangle)_+$ $\;<\;$ $1 - y_i\langle w, x_i\rangle \ge 0$ / $\ldots \le 0$

$(1-\alpha_i)\left[1 - y_i \langle w, x_i \rangle\right]_+ - \alpha_i \left[1 - y_i \langle w, x_i \rangle\right]$

$\alpha_i > 0 \Rightarrow$ either $1 - y_i \langle w, x_i \rangle = 0$ (i.e. "on the margin") or $H_i(w) > 0$





properties of solution: $w^* = \frac{1}{\lambda} \left(\frac{1}{n} \sum_i \overbrace{\sum_y \alpha_i(y)}^{\text{convex combo}} \Psi_i(y) \right)$

1) $\boxed{\|w^*\| \le \frac{\bar{R}}{\lambda}}$

let $R_i \triangleq \max_y \|\Psi_i(y)\|$ then $\|w^*\| \le \frac{1}{\lambda} \frac{1}{n} \sum_i \underbrace{\sum_y \alpha_i(y)}_{1} \overbrace{\|\Psi_i(y)\|}^{\le R_i}$

$\bar{R} = \frac{1}{n} \sum_{i=1}^{n} R_i$ $\qquad \le \frac{1}{\lambda} \left(\sum_{i=1}^{n} R_i \right) = \frac{\bar{R}}{\lambda}$

kernel trick

$$\langle w, \varphi(x, y) \rangle = \frac{1}{\lambda n} \sum_i \sum_y \alpha_i(y) \langle \Psi_i(y), \varphi(x, y) \rangle$$

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

$\langle \varphi(x^{(i)}, y^{(i)}), \varphi(x, y) \rangle$

$k((x^{(i)}, x; y^{(i)}, y))$ Kernel trick?

2) suppose scale features: $\boxed{\tilde{\Psi} = b\Psi}$ $\quad \tilde{H}_i(y; w) = \ell_i(y) - \langle \tilde{w}, \tilde{\Psi}_i(y) \rangle$ $\quad \tilde{w} = \frac{1}{\lambda} \frac{1}{n} \sum_i \sum_y \alpha_i \underbrace{\tilde{\Psi}_i(y)}_{b\Psi_i}$

w

$H_i(y;w) = \ell_i(y) - \langle \tilde{w}, \psi_i(y) \rangle$ $\tilde{w} = \frac{1}{\tilde{\lambda}} \frac{1}{n} \sum_{i,y} \alpha_i \psi_i(y)$ $K(x^{(i)}, x_j; y^{(i)}, y)$ Kernel trick?

want $\tilde{w} = \frac{w}{b}$ want $\frac{w}{b} = \frac{1}{\tilde{\lambda}} \lambda \left(\overbrace{\frac{1}{\lambda n} \sum_{i,y} \alpha_i(y) \psi_i}^{w} \right) b$

$\Rightarrow \boxed{\tilde{\lambda} = b^2 \lambda}$

3) Scale loss: $\tilde{\ell} = b \cdot \ell$ $\Rightarrow \tilde{\lambda} = \frac{\lambda}{b}$

4) etc...

M^3-net example: suppose $\psi(y) = \sum_c \psi_c(y_c)$ $\left[\text{score}(y) = \langle w, \varphi(y) \rangle = \sum_c \langle w, \varphi(y_c) \rangle \right]$

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

$\lambda n A_i \alpha_i = \sum_y \alpha_i(y) \psi_i(y) = \sum_y \alpha_i(y) \left[\sum_c \psi_{i,c}(y_c) \right]$

marginal variable

$$= \sum_c \sum_{y_c} \psi_{i,c}(y_c) \overbrace{\left[\sum_{y' \sim y_c} \alpha_i(y) \right]}^{\triangleq \mu_{i,c}(y_c)}$$

$y' \sim y_c$ y' s.t. $\hookrightarrow y'_c = y_c$

similarly,

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

$\tilde{b}_{i,c}(y_c) \triangleq \underline{\ell_{i,c}(y_c)}$ ... $A_i \alpha_i = \tilde{A}_i \mu_i$ where $(\tilde{A}_i)_{:,(c,y_c)} = \frac{\psi_{i,c}(y_c)}{\lambda n}$ (c, y_c) column of A_i matrix vs. $\{ y_c \in \mathcal{Y}_c \}$

$\langle b, \alpha \rangle = \langle \tilde{b}, \mu(\alpha) \rangle$ $\mu_i \in \mathcal{M}_i$ marginal polytope

$\# \text{columns} = \sum_c |\mathcal{Y}_c|$ $|\mathcal{Y}_c| \approx 2^{|c|}$

vs. $(A_i)_{:,y} = \frac{\psi_i(y)}{\lambda n}$ $y \in \mathcal{Y}_i$ $|\mathcal{Y}| \approx 2^{\# \text{ pts}}$

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

$A = \underbrace{[\cdots \overset{m_i}{A_i} \cdots]}_{m = \sum_i m_i}$

1-slack vs n-slack for SVMstruct constraint generation

n-slack: $\frac{\lambda \|w\|^2}{2} + \frac{1}{n} \sum_i \xi_i$ equivalent $\leftrightarrow$ | 1-slack $\frac{\lambda \|w\|^2}{2} + \xi$ [ML 2009]

[JMLR 2005] $\xi_i \geq H_i(y;w) \; \forall y \in \mathcal{Y}_i$

paper ~

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

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

$O\left(\frac{1}{\varepsilon^2}\right)$ # iterations

here adding constraint (# of oracle calls) can take $O(1)$ if lucky or $O(n)$ if unlucky

$\xi \geq \frac{1}{n}\sum_{i=1}^{n} H_i(y_i; w) \quad \forall (y_1,\dots,y_n) \in \underbrace{\mathcal{Y}_1 \times \mathcal{Y}_2 \times \dots \times \mathcal{Y}_n}_{\bigtimes_{i=1}^{n} \mathcal{Y}_i}$

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

constraints $\prod_{i=1}^{n} |\mathcal{Y}_i|$

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

$\to$ need less constraints in QP to get good solution...,

$O\left(\frac{1}{\varepsilon}\right)$ # iterations

$O(n)$ per iteration always

$w^* = \frac{1}{\lambda} \sum_{(y_1,\dots,y_n) \in \bigtimes_{i=1}^{n} \mathcal{Y}_i} \alpha(y_1,\dots,y_n) \left(\frac{1}{n}\sum_{i=1}^{n} \psi_i(y_i)\right)$

code online for SVMstruct:
https://www.cs.cornell.edu/people/tj/svm_light/svm_struct.html

papers:

- n-slack formulation:
  I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun
  Large Margin Methods for Structured and Interdependent Output Variables, JMLR 2005 (link)
- improved version (1-slack formulation):
  T. Joachims, T. Finley, Chun-Nam Yu, Cutting-Plane Training of Structural SVMs, Machine Learning Journal, 77(1):27-59, 2009 (link)