Lecture 12 - structured SVM

Thursday, February 13, 2020 13:37

Today: structured SVM optimization

Other approaches to optimize SVMstruct

\[
\begin{align*}
& \text{(unconstrained primal)} \\
& \text{min } \frac{1}{2} \|w\|^2 + \frac{1}{n} \sum_{i=1}^{n} H_i(w) \\
& \text{(constrained formulation)} \\
& \text{min } \frac{1}{2} \|w\|^2 + \frac{1}{n} \sum_{i=1}^{n} \xi_i \\
& \text{(smooth, convex QP)} \\
& \text{with } sp. \# of } \xi_i \text{ linear constraints}
\end{align*}
\]

1) General approach: Use convexity of loss-augmented decoding - [Taskar et al. ICML 2005]

Idea here, we suppose that loss-augmented decoding can be expressed as a "compact" maximization problem of a concave set:

\[ H_i(w) = \max_y \langle y, g_i(w) \rangle - \langle w, \phi(y) \rangle = \max_{y \in \mathcal{Y}} g_i(w; z) \]

\[ z \in \mathcal{Z}; \text{ should not depend on } w \]

\[ \mathcal{Z} \text{ is convex in } z \text{ and concave in } w \]

\[ (\text{convex-concave saddle point problem}) \]

\[ \min \max_{w \in \mathcal{W}} \frac{1}{2} \|w\|^2 + \frac{1}{n} \sum_{i=1}^{n} g_i(w; z) \]

\[ \min \max_{w \in \mathcal{W}} g(w, z) \text{ convex in } w \]

\[ \max_{z \in \mathcal{Z}} g(w, z) \text{ concave in } z \]

\[ \forall z \in \mathcal{Z}, \forall w \in \mathcal{W}, g(w, z) \leq g(w^*, z) \leq g(w^*, z^*) \]

\[ \text{under reg. conditions } \min \max = \max \min \rightarrow \text{saddle point} \]

\[ \forall z \in \mathcal{Z}, \forall w \in \mathcal{W}, g(w, z) \leq g(w^*, z) \]

\[ \text{w.r.t. } z \text{, argmin } g(w, z^*) \text{ depends on } z \]

A saddle point (in red) on the graph of z=x^2-y^2 (hyperbolic paraboloid)
Standard dual:

Extended dual algorithm:

\[
\begin{align*}
(w_{e+1}, z_{e+1}) &= (w_e, z_e) + \lambda_e \left( \nabla w_{e} \left( \bar{w}_{e}, \bar{z}_{e} \right) \right) \\
\lambda_{e+1} &= \frac{\sum_{i=1}^{n} \nabla z_i \left( \bar{w}_{e}, \bar{z}_{e} \right)}{\sum_{i=1}^{n} \nabla z_i \left( \bar{w}_{e}, \bar{z}_{e} \right)}
\end{align*}
\]

\[z_{e+1} = \bar{z}_{e} - \lambda_e F(z_{e+1})\]

1st-order approx. to implicit method:

\[z_{e+1} = \bar{z}_{e} - \lambda_e F(z_{e+1})\]

applied to structure SVM [Taskar et al. June 2006]

b) Small "completition" gap formulation for structured SVM

\[
H_i(w) = \max_{z_i \in \mathbb{Z}_i} \min_{v_i \in V_i(w)} \left\{ \phi_i(w_z) \right\}
\]

obtain: \min_{v_i \in V_i(w)} \left\{ \gamma_i(w_v) \right\}

\[
\min_{v_i \in V_i(w)} \left\{ \gamma_i(w_v) \right\}
\]

If \( \gamma_i \) is jointly convex in \( w \) and \( z \), we get a "tractable" convex min problem

- can solve with iterative convex min alg.

- if \( \gamma_i \) not convex, use interior point soln.
  e.g. MOSEK CPLEX (commercial), CVXOPT (free python)

Examples of \( \phi_i(w_z) \):

1) word alignment:

Example: English

Example: French

Example: Spanish

\[\text{Spanish to English}\]

\[\text{French to English}\]

\[\text{German to English}\]

\[\text{English to Spanish}\]

\[\text{English to French}\]

\[\text{English to German}\]
recall that score: \[ s(x, y; w) = \sum_{k, l} y_{k,l} \left[ w_j(x_k, x_l) \right] \]

Let \( y \in \mathbb{R}^{L \times L} \)

Let \( \mathbf{y}^* \) be \[
\begin{bmatrix}
\cdots & \mathbf{y}_e & \cdots
\end{bmatrix}
\]
\( \mathbf{d} \ll (L \times L) \)

\[ s(x, y; w) = w^T \mathbf{y} \]
\[ s(x^{10}, y; w) = w^T \mathbf{y}^* \]

Decoding: \( \hat{w}(x_{10}) = \arg \max_{y_{10}} s(x_{10}, y; w) \)

\[ \begin{array}
\mathbf{h}(x_{10}) = \arg \max_{y_{10}} s(x_{10}, y; w) \\
\text{subject to} \\
\begin{align}
\sum y_{10,k} &\leq \mathbf{y}_e \\
\sum y_{k,10} &\leq \mathbf{y}_e \\
\mathbf{y}_{10} &\leq \mathbf{y}_e
\end{align}
\end{array} \]

\( \mathbf{c} = \mathbf{y}_e \times 2 \Rightarrow \mathbf{C} = \begin{pmatrix}
1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1
\end{pmatrix} \]

\( \mathbf{y}^* = \mathbf{y}^* \times 2 \)

\( \text{relaxation is tight} \)

\( \exists \mathbf{y} \Rightarrow \mathbf{y}^* \in \text{convex hull} \{ \mathbf{y} \} \)

Reasons that relaxation is tight:

1. Write \( z_{10} \) as \( A_{10} z_b \), \( z_b \geq 0 \)

Main: \( A_{10} \) here is "blocky modular"

Which means any subdeterminant of \( A \) has value \( \frac{1}{2} \)

\( \Rightarrow \) that if \( z_b \) has integer entries

Then all vertices of \( \mathbb{Z}^n : A_{10} z_b, z_b \geq 0 \)

Have integer coordinates

\( \Rightarrow \) relaxation is tight for any linear cost

Idea: \( \tilde{A}_{10} \tilde{z} = \tilde{b} \) a corner of the obtained "blocky"

\( \tilde{A}_{10} \tilde{z} = b_{10} \) for \( A_{10} \) infeasible
Conclusion: can write decoding as $\max_{z \in M_i} w^T z$

What about loss?

Hamming loss example

$$L(y, \hat{y}) = \frac{1}{k} \sum I(y_i \neq \hat{y}_i)$$

$$= \frac{1}{k} \left( y_i - \hat{y}_i \right)^2$$

$$= \frac{1}{k} \left( y_i^2 - 2y_i \hat{y}_i + \hat{y}_i^2 \right) = \left( y_i^2 + 1 - 2y_i \hat{y}_i \right) \frac{1}{k}$$

$$q_i | y) = a_i + \frac{(1 - 2y_i \hat{y}_i)}{c_i}$$

Loss-augmented decoding:

$$\max_{\substack{y \in \{0,1\} \in M_i \in \mathbb{R}}} a_i + \frac{(1 - 2y_i \hat{y}_i)}{c_i} - \frac{(w^T y_i) - w^T \hat{y}_i}{w^T \hat{y}_i}$$

$$= a_i - w^T y_i + \max_{\substack{y \in \{0,1\} \in M_i \in \mathbb{R}}} \left( f_i^T w + c_i \right)^T y$$

$$= \max_{z \in M_i} \left( f_i^T w + c_i \right)^T z + a_i - w^T y_i$$

1st duality:

$$\max_{A \succeq b} c^T z = \min_{A^T y \succeq c} b^T z$$

$$\max_{A^T y \succeq c} c^T z = \min_{b \geq 0} a_i^T (w_i z)$$

where $c_i \succeq f_i w + c_i$

$A_i$ is $2L \times L^2$
Sym short objective becomes

\[
\min \min \limits_{u \in \mathcal{U}} \min \limits_{v \in \mathcal{V}} \frac{1}{2} \sum_{i=1}^{n} w_i^2 + \sum_{i=1}^{n} [a_i - w_i^T u_i] v_i^T v_i
\]

s.t. \( A^T u \geq \alpha_{\bar{\epsilon}} w + c \)  
(weakly complementarity CP)

\[
A^T (\bar{v}) \leq b
\]

\[
\min \max \limits_{w \in \mathcal{W}} \frac{1}{2} \sum_{i=1}^{n} w_i^2 + \sum_{i=1}^{n} [a_i - w_i^T u_i] v_i^T v_i
\]

\[
\bar{v} \in \mathcal{V}
\]

*simple contains*