Project 3: List Prediction and Online SVM

Dawei Wang (daweiwan@andrew.cmu.edu)

Theoritical: List Prediction

- 1.2.1: Monotone Submodularity:

  MONOTONICITY: $\forall L_1, L_2, T \subseteq S$, $L_1 \cap T \subseteq (L_1 \cup L_2) \cap T$, [1] $\Rightarrow |L_1 \cap T| \leq |(L_1 \cup L_2) \cap T|$, therefore $f(L_1; T) = \min(1, |L_1 \cap T|) \leq \min(1, |(L_1 \cup L_2) \cap T|) = f(L_1 \cup L_2; T)$;

  SUBMODULARITY: $\forall L_1, L_2, T \subseteq S$, and $\forall s \in S$, we have

$$b(s|L_1) = f(L_1 \cup \{s\}; T) - f(L_1; T) \tag{1}$$
$$b(s|L_1 \cup L_2) = f(L_1 \cup L_2 \cup \{s\}; T) - f(L_1 \cup L_2; T) \tag{2}$$

  consider the following cases:

  - $s \in L_1 \Rightarrow L_1 \cup \{s\} = L_1, L_1 \cup L_2 \cup \{s\} = L_1 \cup L_2 \Rightarrow b(s|L_1) = 0 \geq 0 = b(s|L_1 \cup L_2)$;
  - $s \notin L_1$, $s \in L_2 \Rightarrow L_1 \cup L_2 \cup \{s\} = L_1 \cup L_2 \Rightarrow b(s|L_1) \geq 0$ (monotonicity) $= b(s|L_1 \cup L_2)$;
  - $s \notin L_1$, $s \notin L_2$, $s \in T$, from Equations (1) and (2):

$$\begin{aligned} b(s|L_1) &= \min(1, |(L_1 \cup \{s\}) \cap T|) - f(L_1; T) & (3)\\ &= \min(1, |(L_1 \cap T) \cup (\{s\} \cap T)|) - f(L_1; T) & (4)\\ &= \min(1, |L_1 \cap T| + 1) - f(L_1; T) = 1 - f(L_1; T) & (5)\\ b(s|L_1 \cup L_2) &= \min(1, |(L_1 \cap L_2 \cap T) \cup (\{s\} \cap T)|) - f(L_1 \cup L_2; T) & (6)\\ &= \min(1, |L_1 \cap L_2 \cap T| + 1) - f(L_1 \cup L_2; T) = 1 - f(L_1 \cup L_2; T) & (7) \end{aligned}$$

    since $f(L_1; T) \leq f(L_2 \cup L_2; T)$ (monotonicity), we have $b(s|L_1) \geq b(s|L_1 \cup L_2)$;
  - $s \notin L_1$, $s \notin L_2$, $s \notin T \Rightarrow \{s\} \cap T = \varnothing$, from Equations (4) and (6):

$$b(s|L_1) = \min(1, |L_1 \cap T|) - f(L_1; T) = 0 \tag{8}$$
$$b(s|L_1, L_2) = \min(1, |L_1 \cap L_2 \cap T|) - f(L_1 \cup L_2; T) = 0 \tag{9}$$

    we still have $b(s|L_1) = 0 \geq 0 = b(s|L_1 \cup L_2)$.

  therefore, the *multiple guess* function is monotone and submodular. ■

- 1.2.2: Greedy Guarantee:

$$\begin{aligned} \textbf{STEP1}: \Delta_i &= f(L^*) - f(L^G_{i-1}) \leq f(L^G_{i-1} \oplus L^*) - f(L^G_{i-1}) \quad \text{(monotonicity)} & (10)\\ &= \textstyle\sum_{j=1}^{k} \left[f(L^G_{i-1} \oplus L^*_j) - f(L^G_{i-1} \oplus L^*_{j-1})\right] & (11)\\ &= \textstyle\sum_{j=1}^{k} \left[f(L^G_{i-1} \oplus L^*_{j-1} \oplus l^*_j) - f(L^G_{i-1} \oplus L^*_{j-1})\right] & (12)\\ &\leq \textstyle\sum_{j=1}^{k} \left[f(L^G_{i-1} \oplus l^*_j) - f(L^G_{i-1})\right] \quad \text{(submodularity)} & (13)\\ \textbf{STEP2}: \Delta_i &\leq \textstyle\sum_{j=1}^{k} \left[f(L^G_{i-1} \oplus l^*_j) - f(L^G_{i-1})\right] & (14)\\ &\leq \textstyle\sum_{j=1}^{k} \left[f(L^G_{i-1} \oplus l^G_i) - f(L^G_{i-1})\right] \quad \text{(greediness)} & (15)\\ &= \textstyle\sum_{j=1}^{k} \left[f(L^G_i) - f(L^*)\right] + \left[f(L^*) - f(L^G_{i-1})\right] & (16)\\ &= \textstyle\sum_{j=1}^{k} \left[-\Delta_{i+1} + \Delta_i\right] = k(\Delta_i - \Delta_{i+1}) & (17)\\ \Delta_{i+1} &\leq (1 - 1/k)\,\Delta_i & (18)\\ \textbf{STEP3}: \Delta_{k+1} &\leq (1 - 1/k)\Delta_k \leq (1 - 1/k)^k \Delta_1 \leq (1/e)\Delta_1 & (19)\\ f(L^*) - f(L_G) &\leq (1/e) f(L^*) & (20)\\ f(L_G) &\geq (1 - 1/e) f(L^*) & (21) \end{aligned}$$

  therefore, the greedy optimization policy ensures near-optimal performance. ■

[1] $\forall x \in L_1 \cap T$, $x \in L_1$ and $x \in T$; therefore $x \in L_1 \cap L_2$, and $x \in (L_1 \cup L_2) \cap T$.

Theoritical: Online SVM

- 2.2.1: Formulation Equivalency: the constraints can be re-written as

$$\begin{aligned} \xi_i &\geq 0 \\ \xi_i &\geq 1 - y_i w^T f_i \end{aligned} \quad \Leftrightarrow \quad \xi_i \geq \max\left(0, 1 - y_i w^T f_i\right) \tag{22}$$

and the optimization can be performed sequentially with respect to ξ and w:

$$\min_{\xi,w}\left[\frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{T}\xi_i\right] = \min_{w}\left[\frac{\lambda}{2}\|w\|^2 + \min_{\xi}\sum_{i=1}^{T}\xi_i\right] = \min_{w}\left[\frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{T}\max(0, 1 - y_i w^T f_i)\right] \tag{23}$$

therefore the two problem formulations are equivalent. ■

- 2.2.2: Convexity: considering the objective function

$$J(w;\mathcal{D}) = \sum_{i=1}^{T}\left[\frac{\lambda}{2T}\|w\|^2 + \max(0, 1 - y_i w^T f_i)\right] \tag{24}$$

with $\forall w_1, w_2$ and $\gamma \in [0,1]$, we have [2]

$$\begin{aligned}
\max(0, 1 - y_i[(1-\gamma)w_1 + \gamma w_2]^T f_i) &= \max(0, (1-\gamma)(1 - y_i w_1^T f_i) + \gamma(1 - y_i w_2^T f_i)) && (25)\\
&\leq \max(0, (1-\gamma)(1 - y_i w_1^T f_i)) + \max(0, \gamma(1 - y_i w_2^T f_i)) && (26)\\
&= (1-\gamma)\max(0, 1 - y_i w_1^T f_i) + \gamma\max(0, 1 - y_i w_2^T f_i) && (27)\\
\|(1-\gamma)w_1 + \gamma w_2\|^2 &\leq (1-\gamma)^2\|w_1\|^2 + \gamma^2\|w_2\|^2 \quad \text{(triangle inequality)} && (28)\\
&\leq (1-\gamma)\|w_1\|^2 + \gamma\|w_2\|^2 && (29)
\end{aligned}$$

and combining both terms yields

$$\begin{aligned}
J[(1-\gamma)w_1 + \gamma w_2; \mathcal{D}] &\leq \textstyle\sum_{i=1}^{T}\left[\frac{\lambda}{2T}\|(1-\gamma)w_1 + \gamma w_2\|^2 + \max(0, 1 - y_i[(1-\gamma)w_1 + \gamma w_2]^T f_i)\right] && (30)\\
&\leq \textstyle\sum_{i=1}^{T}\frac{\lambda}{2T}((1-\gamma)\|w_1\|^2 + \gamma\|w_2\|^2)\\
&\quad + (1-\gamma)\max(0, 1 - y_i w_1^T f_i) + \gamma\max(0, 1 - y_i w_2^T f_i) && (31)\\
&= (1-\gamma)J(w_1;\mathcal{D}) + \gamma J(w_2;\mathcal{D}) && (32)
\end{aligned}$$

therefore this objective function is a convex function. ■

- 2.2.3: Sub-gradient Descent: for $\forall w, u$, if $1 - y_t w^T f_t > 0$

$$\begin{aligned}
l(w) + \nabla l_t(w)^T(u - w) &= l(w) + (\tfrac{\lambda}{T}w^T - y_t f_t^T)(u - w) && (33)\\
&= \tfrac{\lambda}{2T}\|w\|^2 + (1 - y_t w^T f_i) + \tfrac{\lambda}{T}w^T u - y_t f_t^T u - \tfrac{\lambda}{T}\|w\|^2 + y_t f_t^T w && (34)\\
&= -\tfrac{\lambda}{2T}\|w\|^2 + \tfrac{2\lambda}{2T}w^T u + (1 - y_t f_t^T u) && (35)\\
&= -\tfrac{\lambda}{2T}\|w - u\|^2 + \tfrac{\lambda}{2T}\|u\|^2 + (1 - y_t f_t^T u) && (36)\\
&\leq \tfrac{\lambda}{2T}\|u\|^2 + \max(0, 1 - y_t f_t^T u) = l(u) && (37)
\end{aligned}$$

otherwise, i.e., if $1 - y_t w^T f_t \leq 0$, $\max(0, 1 - y_t w^T f_t) = 0$, we have

$$\begin{aligned}
l(w) + \nabla l_t(w)^T(u - w) &= l(w) + (\tfrac{\lambda}{T}w^T)(u - w) && (38)\\
&= \tfrac{\lambda}{2T}\|w\|^2 - \tfrac{\lambda}{T}\|w\|^2 + \tfrac{2\lambda}{2T}w^T u && (39)\\
&= -\tfrac{\lambda}{2T}\|w - u\|^2 + \tfrac{\lambda}{2T}\|u\|^2 + 0 && (40)\\
&\leq \tfrac{\lambda}{2T}\|u\|^2 + \max(0, 1 - y_t f_t^T u) = l(u) && (41)
\end{aligned}$$

therefore the proposed sub-gradient $\nabla l_t(w)$ is valid. ■

[2] $\max(0, x+y) \leq \max(0, x) + \max(0, y)$, since if $x + y \leq 0$, $\max(0, x+y) = 0 \leq \max(0, x) + \max(0, y)$; otherwise, if $x + y > 0$, $\max(0, x) + \max(0, y) \geq x + y = \max(0, x+y)$. Hence this inequality always holds regardless of x and y.