SVM and SMO

Instructor: Prof. Ganesh Ramakrishnan

Joachims' SVM^{light}

Choosing the working set in *SVM*^{light}

- Let

$$f(\alpha) = \frac{1}{2}\alpha^\top Q\alpha - e^\top \alpha$$

- *SVM*^{light} chooses working set B by solving for:

$$\Delta\alpha = \min_d \nabla^\top f(\alpha^k) d$$

where d is the descent direction and $\Delta\alpha = \alpha^{k+1} - \alpha^k$
s.t.

  - $|\{d_i : d_i \neq 0\}| \leq q$

    Intuitively, if q non-zero d_i's are possible, they *will* be picked up since such a set will reduce the objective further as compared to a smaller set
  - $y^\top d = 0$

    $(\alpha^k)^\top y = 0$, and $(\alpha^{k+1})^\top y = 0 \implies (\alpha^k + d)^\top y = 0$

    Thus, $y^\top d = 0$
  - $d_i \in [-1, 1]$
  - $d_i \geq 0$, for $(\alpha^k)_i = 0$
  - $d_i \leq 0$, for $(\alpha^k)_i = C$

Solving for d in SVM^{light}

The intuition is that:

- The descent directions d_i's for the most violating $(\alpha^k)_i$'s correspond to the $(\nabla f(\alpha^k))_i$'s that are farthest from 0,
- taking care that we also want $y^\top d = 0$, ie. $\sum y_i d_i = 0$, for all i's chosen as per above
(s.t. $|\{d_i : d_i \neq 0\}| \leq q$)

Solving for d in SVM^{light}

1. Sort $y_i(\nabla f(\alpha^k))_i$ in decreasing order
2. Symmetrically do:
   From the top, sequentially set $d_i = -y_i$
   From the bottom, sequentially set $d_i = y_i$
   - Until either
     - $\frac{q}{2}$ '$d_i = -y_i$'s have been selected from the top, and $\frac{q}{2}$ '$d_i = y_i$'s have been selected from the bottom
     - we cannot find $d_i = -y_i$ from the top and $d_i = y_i$ from the bottom at the same time
   - At any point,
     if $(\alpha_k)_i = 0$ and $d_i = -1$, set $d_i = 0$ and bypass it, and
     if $(\alpha_k)_i = C$ and $d_i = 1$, set $d_i = 0$ and bypass it
   - The goal is to achieve a balancing between the two signs from the opposite ends, ie. $\sum y_i d_i = 0$
3. d_i's not yet considered are assigned 0

If $\frac{q}{2}$ '$d_i = -y_i$'s from the top and $\frac{q}{2}$ '$d_i = y_i$'s from the bottom could not be selected (or if q is large enough), the algorithm will stop at i_t from the top and i_b from the bottom

One of the following will happen:

- i_t is just before i_b
- There is one position i between i_t and i_b with $0 < (\alpha^k)_i < C$

When the algorithm stops, d is an optimal solution for

$$\Delta\alpha = \min_{d} \nabla^\top f(\alpha^k) d$$

s.t.

- $|\{d_i : d_i \neq 0\}| \leq q$
- $y^\top d = 0$
- $d_i \in [-1, 1]$
- $d_i \geq 0$, for $(\alpha^k)_i = 0$
- $d_i \leq 0$, for $(\alpha^k)_i = C$

When the algorithm stops at i_t, if the next index in the sorted list of $y_i(\nabla f(\alpha^k))_i$ is $\bar{i}_t$, there are three possible situations:

- $(\alpha^k)_{\bar{i}_t} \in (0, C)$
- $(\alpha^k)_{\bar{i}_t} = 0$ and $y_{\bar{i}_t} = -1$
- $(\alpha^k)_{\bar{i}_t} = C$ and $y_{\bar{i}_t} = 1$

If the last two do not hold, we can move down further by assigning $d_{\bar{i}_t} = 0$

Decomposition in Joachims' SVM^{light} (continued)

Choice of the working set size q

- In the decomposition algorithm, a working set size $q \leq l$ must be chosen
- There is a tradeoff between q and the number of iterations needed for the algorithm to converge
  - The higher the working set size q, the lower will be the number of iterations needed
  - However, with a larger q, individual iterations become extremely expensive

Correctness of the algorithm

- Verify that the algorithm actually minimizes the objective[1]
- When an iteration of the algorithm stops, d is an optimal solution for

$$\Delta\alpha = \min_{d} \nabla^\top f(\alpha^k) d$$

s.t.

  - $|\{d_i : d_i \neq 0\}| \leq q$
  - $y^\top d = 0$
  - $d_i \in [-1, 1]$
  - $d_i \geq 0$, for $(\alpha^k)_i = 0$
  - $d_i \leq 0$, for $(\alpha^k)_i = C$

[1] Full proof at `http://www.csie.ntu.edu.tw/~cjlin/papers/conv.pdf`

When an iteration of the algorithm stops, the following KKT conditions are satisfied, showing that d is an optimal solution:

- $\nabla f(\alpha^k) = -by + \lambda_i - \xi_i$
- $y^\top d = 0$
- $\lambda_i(d_i + 1) = 0$, if $0 < \alpha_i^k \leq C$
- $\lambda_i d_i = 0$, $\alpha_i^k = 0$
- $\xi_i(1 - d_i) = 0$, if $0 \leq \alpha_i^k < C$
- $\xi_i d_i = 0$, if $\alpha_i^k = C$
- $\lambda_i \geq 0$, $\xi_i \geq 0$, $\forall i = 1, \ldots, l$

- Assume that B is the working set at the kth iteration, and $N = 1, \ldots, l \backslash B$
- If we define $s = \alpha^{k+1} - \alpha^k$, then $s_N = 0$ and
    - $f(\alpha^{k+1}) - f(\alpha^k)$
      $= \frac{1}{2} s^\top Q s + s^\top Q \alpha^k - e^\top s$
      $= \frac{1}{2} s_B^\top Q_{BB} s_B + s_B^\top (Q\alpha^k)_B - e_B^\top s_B$

Thus, in the kth iteration, we solve the following problem with the variable s_B:

$$\min_{s_B} \frac{1}{2} s_B^\top Q_{BB} s_B + s_B^\top (Q\alpha^k)_B - e_B^\top s_B$$

s.t.

- $0 \le (\alpha_k + s)_i \le C,\ i \in B$
- $y_B^\top s_B = 0$

This is written purely in terms of the basis B components, ignoring the function of s_N in the objective which does not depend on s_B

Using the KKT conditions that the optimal solution s_B must satisfy, we show a sufficient decrease of $f(\alpha)$ over the iterations:

- $f(\alpha_{k+1}) \leq f(\alpha^k) - \frac{\sigma}{2}\left\|\alpha^{k+1} - \alpha^k\right\|^2$
  - where, $\sigma = \min_I(\min(eig(Q_{II})))$
- At every step, the function decreases by an amount that does not become insignificant

Convergence of SMO

- **SVM Dual objective**:

$$\min_{\alpha} \frac{1}{2}\alpha^\top Q\alpha - e^\top \alpha$$

s.t.

  - $0 \leq \alpha_i \leq C, \forall i$
  - $y^\top \alpha = 0$

- where:
  - Q is positive-semidefinite, and $Q_{ij} = y_i y_j \phi^\top(x_i)\phi(x_j)$
  - $e = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbf{R}^n$

- We split the constraint $0 \leq \alpha_i \leq C, \forall i$ into:
  - $-\alpha_i \leq 0$, with the Lagrange multiplier θ_i
  - $\alpha_i \leq C$, with the Lagrange multiplier Γ_i
- and, consider $y^\top \alpha = 0$, with the Lagrange multiplier β
- Thus, we can write the Lagrangian as:
  $L(\alpha, \theta, \Gamma, \beta) = \frac{1}{2}\alpha^\top Q\alpha - e^\top \alpha - \theta^\top \alpha + \Gamma^\top(\alpha - C) + \beta y^\top \alpha$
  s.t. $\forall i$,
  - $\theta_i \geq 0$
  - $\Gamma_i \geq 0$
  - $\theta_i \alpha_i = 0$
  - $\Gamma_i(\alpha_i - C) = 0$
- Taking $\nabla_\alpha L = 0$, we get:
  $Q\alpha - e - \theta + \Gamma + \beta y = 0$

- If $\alpha_i = 0$, $\alpha_i - C \neq 0$ and thus $\Gamma_i = 0$
  - $(Q\alpha)_i - 1 - \theta_i + \beta y_i = 0$
    $\implies \theta_i = (Q\alpha)_i - 1 + \beta y_i$
  - As $\theta_i \geq 0$,
  $$(Q\alpha)_i - 1 + \beta y_i \geq 0$$
- If $\alpha_i = C$, $\theta_i = 0$
  - $(Q\alpha)_i - 1 + \Gamma_i + \beta y_i = 0$
    $\implies -\Gamma_i = (Q\alpha)_i - 1 + \beta y_i$
  - As $\Gamma_i \geq 0$,
  $$(Q\alpha)_i - 1 + \beta y_i \leq 0$$
- If $\alpha_i \in (0, C)$, $\theta_i = 0$ and $\Gamma_i = 0$
  - Thus,
  $$(Q\alpha)_i - 1 + \beta y_i = 0$$

Let us define the following sets of indices i

- $I_0(\alpha) = \{i : 0 < \alpha_i < C\}$
- $I_1(\alpha) = \{i : y_i = +1, \alpha_i = 0\}$
- $I_2(\alpha) = \{i : y_i = -1, \alpha_i = C\}$
- $I_3(\alpha) = \{i : y_i = +1, \alpha_i = C\}$
- $I_4(\alpha) = \{i : y_i = -1, \alpha_i = 0\}$

Let us now consider

1. $I_0(\alpha) \cup I_1(\alpha) \cup I_2(\alpha)$
   $= \{i : y_i = +1, \alpha_i < C\} \cup \{i : y_i = -1, \alpha_i > 0\}$
   - $\big((Q\alpha)_i - 1\big)\, y_i \geq -\beta$
2. $I_0(\alpha) \cup I_3(\alpha) \cup I_4(\alpha)$
   $= \{i : y_i = +1, \alpha_i > 0\} \cup \{i : y_i = -1, \alpha_i < C\}$
   - $\big((Q\alpha)_i - 1\big)\, y_i \leq -\beta$

Here, $\big((Q\alpha)_i - 1\big)\, y_i$ is equivalent to $\big(\nabla f(\alpha)\big)_i\, y_i$ from the decomposition algorithm in Joachims' SVM^{light}

Thus, we have:

$$\min_{i \in I_0 \cup I_1 \cup I_2} \big((Q\alpha) - 1\big)\, y_i \geq \max_{i \in I_0 \cup I_3 \cup I_4} \big((Q\alpha) - 1\big)\, y_i$$

We get:

$$\min\left(\min_{y_i=+1,\alpha_i>0} -\big(\nabla f(\alpha)\big)_i,\ \min_{y_i=-1,\alpha_i<C} \big(\nabla f(\alpha)\big)_i\right)$$

$$\geq$$

$$\max\left(\max_{y_i=+1,\alpha_i<C} -\big(\nabla f(\alpha)\big)_i,\ \max_{y_i=-1,\alpha_i>0} \big(\nabla f(\alpha)\big)_i\right)$$

Let the min be attained at index l, and max be attained at index j.
If for (l, j), the inequality is violated, the KKT conditions are violated.

We need to prove that for all such choices of l and j across iterations, $\forall k$,

$$f(\alpha^{k+1}) \leq f(\alpha^k) - \frac{\sigma}{2}\left\|\alpha^{k+1} - \alpha^k\right\|^2$$

s.t. $\sigma > 0$, and $\alpha^{k+1} \neq \alpha^k$

Once we find l and j, we will find closed form solutions for
$\alpha_l^{k+1} = g(\alpha_l^k, \alpha_j^k, \alpha_N^k)$
$\alpha_j^{k+1} = \bar{g}(\alpha_l^k, \alpha_j^k, \alpha_N^k)$
(which have been discussed before)

- Whatever be the values of α_l^{k+1} and α_j^{k+1}, we will have:
  - $y_l\alpha_l^{k+1} + y_j\alpha_j^{k+1} = -y_N^\top \alpha_N^k$ (constant)
- Thus, we can say that if α_l changes linearly, then α_j also changes linearly
  - We can replace α_l^{k+1} and α_j^{k+1} as:
    $\alpha_l(t) \leftarrow \alpha_l^{k+1}$
    $\alpha_j(t) \leftarrow \alpha_j^{k+1}$
- α_l and α_j vary linearly with t
  - $\alpha_l(t) \equiv \alpha_l^k + ty = \alpha_l^k + \frac{t}{y_l}$
  - $\alpha_j(t) \equiv \alpha_j^k + ty = \alpha_j^k + \frac{t}{y_j}$

- Let $f(\alpha) = \psi(t)$
  - ψ is a function of α_N, $\alpha_l(t)$, and $\alpha_j(t)$
- We need to analyze w.r.t. $\bar{t}$ that minimizes $\psi(t)$ subject to constraints
  - $\sum \alpha_i y_i = 0$
  - $\alpha_i \in [0, C]$
- That would give
  - $\alpha^k = \begin{bmatrix} \alpha_N^k \\ \alpha_j^k \\ \alpha_l^k \end{bmatrix}$, and $\alpha^{k+1} = \begin{bmatrix} \alpha_N^k \\ \alpha_j^k + \frac{\bar{t}}{y_j} \\ \alpha_l^k + \frac{\bar{t}}{y_l} \end{bmatrix}$
  - $\alpha^{k+1} - \alpha^k = \begin{bmatrix} 0 \\ \frac{\bar{t}}{y_j} \\ \frac{\bar{t}}{y_l} \end{bmatrix}$
- Taking norm on both sides, we get:
  $\left\|\alpha^{k+1} - \alpha^k\right\| = 2\bar{t}^2$
  $\implies |\bar{t}| = \frac{1}{\sqrt{2}}\left\|\alpha^{k+1} - \alpha^k\right\|$

- Now, $\psi(t)$ is a quadratic function on t
- Thus, $\psi(t) = \psi(0) + \psi'(0) + \psi''(0)\frac{t^2}{2}$
- $\psi'(t) = \sum_{i=1}^{m} \Big(\nabla f\big(\alpha(t)\big)\Big)_i \frac{d\alpha_i(t)}{dt}$
  $= y_l \Big(\nabla f\big(\alpha(t)\big)\Big)_l - y_j \Big(\nabla f\big(\alpha(t)\big)\Big)_j$
  $= y_l \left(\sum_{i=1}^{m} Q_{li}\alpha_i(t) - 1\right) - y_j \left(\sum_{i=1}^{m} Q_{ji}\alpha_i(t) - 1\right)$
  - $\psi'(0) = y_l \Big(\nabla f(\alpha^k)\Big)_l - y_j \Big(\nabla f(\alpha^k)\Big)_j$
- $\psi''(t) = Q_{ll} + Q_{jj} - 2y_l y_j Q_{lj}$
  $= \phi^\top(x_l)\phi(x_l) + \phi^\top(x_j)\phi(x_j) - 2\phi^\top(x_l)\phi(x_j)$
  - $\psi''(0) = \left\|\phi(x_l) - \phi(x_j)\right\|^2$

- $\bar{t}$ minimizes $\psi(t)$ s.t. $\sum \alpha_i y_i = 0$ and $\alpha_i \in [0, C]$, $\forall i$
  - $|\bar{t}| = \frac{1}{\sqrt{2}} \left\| \alpha^{k+1} - \alpha^k \right\|$
- Suppose t^* minimizes $\psi(t)$ without constraints
  - Solving for $\psi'(t^*) = 0$, we get: $t^* = -\frac{\psi'(0)}{\psi''(0)}$
- $\psi(\bar{t}) \geq \psi(t^*)$
- We can say that $\bar{t} = \gamma t^*$, where $\gamma \in [0, 1]$
  (you could have gone till t^* but had to halt at $\bar{t}$ due to constraints)

- $\psi(\bar{t}) = \psi(\gamma t^*) = \psi(-\gamma \frac{\psi'(0)}{\psi''(0)})$
$= \psi(0) + \psi'(0)\left(-\gamma \frac{\psi'(0)}{\psi''(0)}\right) + \frac{\psi''(0)}{2}\left(-\gamma \frac{\psi'(0)}{\psi''(0)}\right)^2$
$= \psi(0) - \gamma \frac{\left(\psi'(0)\right)^2}{\psi''(0)} + \frac{\gamma^2}{2}\frac{\left(\psi'(0)\right)^2}{\psi''(0)}$
- Since $\gamma \in [0,1]$, $\gamma^2 \leq \gamma$, and
$\frac{\gamma^2}{2} - \gamma \leq -\frac{\gamma^2}{2}$
- Thus, $\psi(\bar{t}) \leq \psi(0) - \frac{\gamma^2}{2}\frac{\left(\psi'(0)\right)^2}{\psi''(0)}$
$\implies \psi(\bar{t}) - \psi(0) \leq -\frac{\gamma^2}{2}\frac{\left(\psi'(0)\right)^2}{\psi''(0)}$
$\implies \psi(\bar{t}) - \psi(0) \leq -\frac{\psi''(0)}{4}\left\|\alpha^{k+1} - \alpha^k\right\|^2$
- This becomes:
$f(\alpha^{k+1}) - f(\alpha^k) \leq -\frac{\sigma}{2}\left\|\alpha^{k+1} - \alpha^k\right\|^2$
  - where, $\sigma = \frac{\psi''(0)}{2} = \frac{1}{2}\left\|\phi(x_l) - \phi(x_j)\right\|^2$
  - $\sigma > 0$ except when feature vector $\phi(x_l)$ is the same as $\phi(x_j)$

- We assume Q to be positive-semidefinite so that $\psi''(0) \geq 0$
- But in the analysis of general decomposition, we assumed Q_{II} to be positive-semidefinite for any submatrix of Q, which is a stronger assumption