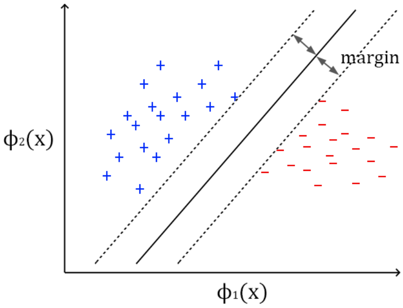
SVM and SMO

Instructor: Prof. Ganesh Ramakrishnan

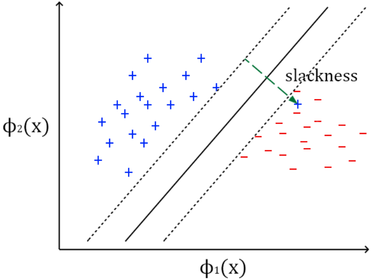
Support Vector Machines



$$w^\top \phi(x) + b \geq +1 \ \mathit{for}\ y = +1$$
$$w^\top \phi(x) + b \leq -1 \ \mathit{for}\ y = -1$$
$$w, \phi \in \mathbb{R}^m$$

There is large margin to seperate the $+$ve and -ve examples

Overlapping examples



When the examples are not linearly seperable, we need to consider the slackness ξ_i of the examples x_i (how far a misclassified point is from the seperating hyperplane, always $+$ve):

$w^\top \phi(x_i) + b \geq +1 - \xi_i$ (*for* $y_i = +1$)

$w^\top \phi(x_i) + b \leq -1 + \xi_i$ (*for* $y_i = -1$)

Multiplying y_i on both sides, we get:

$$y_i(w^\top \phi(x_i) + b) \geq 1 - \xi_i, \forall i = 1, \ldots, n$$

Maximize the margin

- We maximize the margin given by $(\phi(x^+) - \phi(x^-))^\top [\frac{w}{\|w\|}]$
- Here, x^+ and x^- lie on boundaries of the margin.
- We can verify that w is perpendicular to the seperating surface: at the seperating surface, the dot product of w and $\phi(x)$ is 0 (with b captured), which is only possible if w and $\phi(x)$ are perpendicular.
- We project the vectors $\phi(x^+)$ and $\phi(x^-)$ on w, and normalize by w as we are only concerned with the direction of w and not its magnitude.

Simplifying the margin expression

- Maximize the margin $(\phi(x^+) - \phi(x^-))^\top[\frac{w}{\|w\|}]$
- At x^+: $y^+ = 1$, $\xi^+ = 0$ hence, $(w^\top\phi(x^+) + b) = 1$ — 1
  At x^-: $y^- = 1$, $\xi^- = 0$ hence, $-(w^\top\phi(x^-) + b) = 1$ — 2
- Adding 2 to 1,
  $w^\top(\phi(x^+) - \phi(x^-)) = 2$
- *Thus, the margin expression to maximize is:* $\frac{2}{\|w\|}$

Formulating the objective

- Problem at hand: Find w^*, b^* that maximize the margin.
- $(w^*, b^*) = \arg\max_{w,b} \frac{2}{\|w\|}$
  s.t. $y_i(w^\top \phi(x_i) + b) \geq 1 - \xi_i$ and
  $\xi_i \geq 0$, $\forall i = 1, \ldots, n$
- However, as $\xi_i \to \infty$, $1 - \xi_i \to -\infty$
- Thus, with arbitrarily large values of ξ_i, the constraints become easily satisfiable for any w, which defeats the purpose.
- *Hence, we also want to minimize the ξ_i's. ie. minimize $\sum \xi_i$*

Objective

- $(w^*, b^*, \xi_i^*) = \text{argmin}_{w,b,\xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i$
  s.t. $y_i(w^\top \phi(x_i) + b) \geq 1 - \xi_i$ and
  $\xi_i \geq 0, \forall i = 1, \ldots, n$
- Instead of maximizing $\frac{2}{\|w\|}$, minimize $\frac{1}{2}\|w\|^2$
  ($\frac{1}{2}\|w\|^2$ *is monotonically decreasing with respect to* $\frac{2}{\|w\|}$)
- C determines the trade-off between the error $\sum \xi_i$ and the margin $\frac{2}{\|w\|}$

More on the Objective

- $(w^*, b^*, \xi_i^*) = \text{argmin}_{w,b,\xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i$
  s.t. $y_i(w^\top \phi(x_i) + b) \geq 1 - \xi_i$ and
  $\xi_i \geq 0$, $\forall i = 1, \ldots, n$
- Converting the constraints to the form $g_i(x) \leq 0$:
  $1 - \xi_i - y_i(w^\top \phi(x_i) + b) \leq 0$
  $-\xi_i \leq 0$
- $L(w, b, \alpha, \mu, \xi_i) =$
  $\frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i + \sum_{i=1}^{n} \alpha_i(1 - \xi_i - y_i(w^\top \phi(x_i) + b)) + \sum_{i=1}^{n} \mu_i(-\xi_i)$
- We want: $\nabla_{w,b,\xi_i} L(w^*, b^*, \alpha^*, \mu^*, \xi_i^*) = 0$

Gradient of the SVM Lagrangian

$$\nabla L(w^*, b^*, \alpha^*, \mu^*, \xi_i^*) = 0$$

- w.r.t. w:
  $$w^* + \sum_{i=1}^{n} \alpha_i^* (-y_i) \phi(x_i) = 0$$
  $$\implies w^* = \sum_{i=1}^{n} \alpha_i^* y_i \phi(x_i)$$
- w.r.t. b:
  $$\sum_{i=1}^{n} \alpha_i^* y_i = 0$$
- w.r.t. ξ_i, $\forall i$:
  $$C - \alpha_i^* - \mu_i^* = 0$$
  $$\implies \alpha_i^* + \mu_i^* = C, \forall i = 1, \ldots, n$$

Necessary conditions for optimality

1. $y_i(w^{*\top}\phi(x_i) + b^*) \geq 1 - \xi_i^*, \forall i$
2. $\xi_i^* \geq 0, \forall i$
3. $w^* = \sum_{i=1}^{n} \alpha_i^* y_i \phi(x_i)$
4. $\sum_{i=1}^{n} \alpha_i^* y_i = 0$
5. $\alpha_i^* \geq 0, \forall i$
6. $\mu_i^* \geq 0, \forall i$
7. $\alpha_i^* + \mu_i^* = C, \forall i$
8. $\alpha_i^*(1 - \xi_i^* - y_i(w^{*\top}\phi(x_i) + b^*)) = 0, \forall i$
9. $\mu_i^* \xi_i^* = 0, \forall i$

For SVM, since the original objective and the constraints are convex, any $(w^*, b^*, \alpha^*, \mu^*, \xi_i^*)$ that satisfies the necessary conditions gives optimality (conditions are also sufficient)

Some observations

- $\alpha_i^* \geq 0$, $\mu_i^* \geq 0$, and $\alpha_i^* + \mu_i^* = C$
  Thus, $\alpha_i^*, \mu_i^* \in [0, C]$, $\forall i$
- If $0 < \alpha_i^* < C$, then $0 < \mu_i^* < C$
  (as $\alpha_i^* + \mu_i^* = C$)
- $\mu_i^* \xi_i^* = 0$ and $\alpha_i^*(1 - \xi_i^* - y_i({w^*}^\top \phi(x_i) + b^*)) = 0$ are complementary slackness conditions
  If $\xi_i^* = 0$ and $1 - \xi_i^* - y_i({w^*}^\top \phi(x_i) + b^*) = 0$, then $y_i({w^*}^\top \phi(x_i) + b^*) = 1$
  - All such points lie on a margin
  - Using any point on a margin, we can recover b^* as: $b^* = y_i - {w^*}^\top \phi(x_i)$

Dual function

- Let $L^*(\alpha, \mu) = \min_{w,b,\xi} L(w, b, \xi, \alpha, \mu)$
- By weak duality theorem, we have:
  $L^*(\alpha, \mu) \leq \min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$
  s.t. $y_i(w^\top \phi(x_i) + b) \geq 1 - \xi_i$, and
  $\xi_i \geq 0$, $\forall i = 1, \ldots, n$
- The above is true for any $\alpha_i \geq 0$ and $\mu_i \geq 0$
- Thus,

$$\max_{\alpha,\mu} L^*(\alpha, \mu) \leq \min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$

Dual objective

- In case of SVM, we have a convex objective and linear constraints – therefore, strong duality holds:

$$\max_{\alpha,\mu} L^*(\alpha, \mu) = \min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$

- This value is precisely obtained at the $(w^*, b^*, \xi^*, \alpha^*, \mu^*)$ that satisfies the necessary (and sufficient) optimality conditions
- Assuming that the necessary and sufficient conditions (KKT or Karush–Kuhn–Tucker conditions) hold, our objective becomes:

$$\max_{\alpha,\mu} L^*(\alpha, \mu)$$

- $L(w, b, \xi, \alpha, \mu) =$
  $\frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha_i(1 - \xi_i - y_i(w^\top \phi(x_i) + b)) - \sum_{i=1}^n \mu_i \xi_i$
- We obtain w, b, ξ in terms of α and μ by setting $\nabla_{w,b,\xi} L = 0$:
  - **w.r.t.** w**:** $w = \sum_{i=1}^n \alpha_i y_i \phi(x_i)$
  - **w.r.t.** b**:** $-b\sum_{i=1}^n \alpha_i y_i = 0$
  - **w.r.t.** ξ_i**:** $\alpha_i + \mu_i = C$
- Thus, we get:
  $L(w, b, \xi, \alpha, \mu)$
  $= \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \phi^\top(x_i)\phi(x_j) + C\sum_i \xi_i + \sum_i \alpha_i - \sum_i \alpha_i \xi_i -$
  $\sum_i \alpha_i y_i \sum_j \alpha_j y_j \phi^\top(x_j)\phi(x_i) - b\sum_i \alpha_i y_i - \sum_i \mu_i \xi_i$
  $= -\frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \phi^\top(x_i)\phi(x_j) + \sum_i \alpha_i$

- The dual optmimization problem becomes:

$$\max_{\alpha} -\frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \phi^\top(x_i)\phi(x_j) + \sum_i \alpha_i$$

  s.t.
  $\alpha_i \in [0, C]$, $\forall i$ and
  $\sum_i \alpha_i y_i = 0$
- Deriving this did not require the complementary slackness conditions
- Conveniently, we also end up getting rid of μ

Solving SVMs

- *Dual objective:* $\max_\alpha \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j)$ s.t. $\sum \alpha_i y_i = 0$ and $\alpha_i \in [0, C], \forall i$
- We have standard solvers available such as LCQP (linearly constrained quadratic program) solvers like:
  - Projected gradient ascent
  - Active set
  - Ellipsoid
  - Cutting plane
  - etc.
- We will discuss a fast "Active set"-like algorithm known as **Sequential minimal optimization (SMO)**
- SMO algorithm comprises of Projected gradient ascent and Active set

Coordinate Ascent algorithm

- Optimize over one α_i at a time
- However, $\sum \alpha_i y_i = 0$
- Therefore, we consider a *Block Coordinate Ascent* which will optimize over a subset of $\alpha_1, \ldots, \alpha_n$

SMO's Block coordinate acsent (blocksize 2)

- *Objective:*
  $\max_\alpha \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j)$
  s.t. $\sum_i \alpha_i y_i = 0$ and $\alpha_i \in [0, C], \forall i$
- w.l.o.g, we say that α_1 and α_2 are the α's to be updated
  - $\alpha_3^{new} = \alpha_3^{old}, \alpha_4^{new} = \alpha_4^{old}, \ldots, \alpha_n^{new} = \alpha_n^{old}$
  - $\alpha_1^{new} \neq \alpha_1^{old}, \alpha_2^{new} \neq \alpha_2^{old}$
    (equality may hold true under certain conditions like convergence but does not hold by design)

Solving for α_1^{new}, α_2^{new}

- Re-writing the objective in terms of α_1^{new}, α_2^{new}:
  $(\alpha_1^{new}, \alpha_2^{new}) =$
  $\text{argmax}_{\alpha_1,\alpha_2}\ \alpha_1 + \alpha_2 + \sum_{i=3}^{n} \alpha_i^{old} - \frac{1}{2}[\alpha_1^2 y_1^2 K(x_1, x_1) + \alpha_2^2 y_2^2 K(x_2, x_2) + 2\alpha_1 \sum_{j=3}^{n} \alpha_j^{old} y_1 y_j K(x_1, x_j) + 2\alpha_2 \sum_{j=3}^{n} \alpha_j^{old} y_2 y_j K(x_2, x_j) + 2\alpha_1\alpha_2 y_1 y_2 K(x_1, x_2)]$
  - s.t. $\alpha_1 y_1 + \alpha_2 y_2 = -\sum_{j=3}^{n} \alpha_j^{old} y_j$
- Multiplying the constraint by y_2, we have:
  $\alpha_2 = -\alpha_1 y_1 y_2 - \sum_{j=3}^{n} \alpha_j^{old} y_j y_2$
  Let $\sum_{j=3}^{n} \alpha_j^{old} y_j$ be β^{old}
- Thus, $\alpha_2 = -\alpha_1 y_1 y_2 - \beta^{old} y_2$

Substituting the values for α_2 and β^{old} in the SMO objective

- $\alpha_1^{new} = \text{argmax}_{\alpha_1} \frac{1}{2}(2K(x_1, x_2) - K(x_1, x_1) - K(x_2, x_2))\alpha_1^2 + (1 - y_1 y_2 - y_1 K(x_1, x_1)\beta_{old} + y_1 K(x_1, x_2)\beta_{old} + y_1 \sum_{j=3}^{n} \alpha_j^{old} y_j K(x_1, x_j) - y_1 \sum_{j=3}^{n} \alpha_j^{old} y_j K(x_2, x_j))\alpha_1 + \gamma$
  where γ is a constant term
- Simplifying the above expression and taking θ_1 and θ_2 as the coefficients of α_1 and α_1^2 respectively, we get:
  $\alpha_1^{new} = \text{argmax}_{\alpha_1} \theta_1 \alpha_1 + \theta_2 \alpha_1^2 + \gamma$

For more information, see
`http://www.cs.iastate.edu/~honavar/smo-svm.pdf`

- $\alpha_1^{new} = \text{argmax}_{\alpha_1}\ \theta_1\alpha_1 + \theta_2\alpha_1^2 + \gamma$
- For this objective to be upper convex, $\frac{\partial^2}{\partial\alpha_1^2}(\theta_1\alpha_1 + \theta_2\alpha_1^2 + \gamma) \leq 0$
  - Thus $\theta_2 \leq 0$ must hold
  - We can see that $\theta_2 = \frac{1}{2}(2K(x_1, x_2) - K(x_1, x_1) - K(x_2, x_2)) \leq 0$
  - If $K(x_1, x_2) = x_1^\top x_2$, then
    $\theta_2 = \frac{1}{2}(2x_1^\top x_2 - x_1^\top x_1 - x_2^\top x_2)$
    $= -\frac{1}{2}(x_2 - x_1)^\top(x_2 - x_1)$
    $= -\frac{1}{2}\|x_2 - x_1\|^2 \leq 0$
- If $\theta_2 < 0$, the expression gives us the unconstrained maximum point α_1^{new}
- Here, $\frac{\partial}{\partial\alpha_1}(\theta_1\alpha_1 + \theta_2\alpha_1^2 + \gamma) = 0$
  $\implies \alpha_1^{new} = \frac{-\theta_1}{2\theta_2}$

The SMO algorithm

1. Initialise $\alpha_1, \ldots, \alpha_n$ to some value $\in [0, C]$
2. Pick α_i, α_j to estimate next (i.e. estimate α_i^{new}, α_j^{new})
3. $\alpha_i^{new} = \frac{-\theta_1}{2\theta_2}$
   - if $\alpha_i^{new} < 0$ then $\alpha_i^{new} = 0$
   - if $\alpha_i^{new} > C$ then $\alpha_i^{new} = C$
4. $\alpha_j^{new} = -\alpha_i y_i y_j - \beta^{old} y_j$
   - if $\alpha_j^{new} < 0$ then $\alpha_j^{new} = 0$
   - if $\alpha_j^{new} > C$ then $\alpha_j^{new} = C$
5. Check if all the KKT conditions are satisfied
   - $\alpha_i(1 - y_i(w^\top \phi(x_i) + b)) = 0, \forall i$
   - If not, choose α_i and α_j that worst violate the KKT conditions (i.e. max value of $\alpha_i(1 - y_i(w^\top \phi(x_i) + b))$), and reiterate

The SMO procedure has been proved to converge, and is therefore an algorithm

SMO-type decomposition methods for SVMs

- **Dual objective (vectorized)**:

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

s.t.

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

- where:
  - $Q_{ij} = y_i y_j \phi^\top(x_i)\phi(x_j)$
    Thus, Q is like a 'signed' kernel matrix, carrying the dot products of feature fectors $y_i\phi(x_i)$
  - $e = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$

- SMO can be shown to converge asymptotically to a minimum if Q is positive-semidefinite (ie. $\forall x \in \mathbf{R}^n$, $x^\top Qx \geq 0$)

The general decomposition method

1. Fix a working set size $q \leq n$, where n is the number of examples; Let α^1 be the initial solution at iteration counter value $k = 1$
2. If α^k satisfies KKT conditions, stop;
   else, find a working set $B \subset \{1, \ldots, n\}$ s.t. $|B| = q$
   Let $N = \{1, \ldots, n\} \backslash B$, and $\begin{bmatrix} \alpha_B^k \\ \alpha_N^k \end{bmatrix}$ be a partition of α^k
3. Solve the following subproblem (for α_B):

$$\min_{\alpha_B} \frac{1}{2} \alpha_B^\top Q_{BB} \alpha_B - (e_B - Q_{BN} \alpha_N^k)^\top \alpha_B$$

   s.t.
   - $0 \leq (\alpha_B)_i \leq C, \forall i = 1, \ldots, q$
   - $y_B^T \alpha_B = -y_N^T \alpha_N^k$

   where $\begin{bmatrix} Q_{BB} & Q_{BN} \\ Q_{NB} & Q_{NN} \end{bmatrix}$ is a permutation of the matrix Q.
4. Set α_B^{k+1} to be the optimal solution of 3, and $\alpha_N^{k+1} = \alpha_N^k$. Set $k \leftarrow k + 1$ and go to 2

- *w.l.o.g.*, $\alpha = \begin{bmatrix} \alpha_B^k \\ \alpha_N^k \end{bmatrix}$ is obtained by permuting the examples.
  B is often chosen as the maximal KKT violating set.
- For SMO, $q = 2$

In SVM^{light}, Joachims chooses B by solving another (smaller) optimization problem[1]

[1] http://www.cs.cornell.edu/people/tj/publications/joachims_99a.pdf