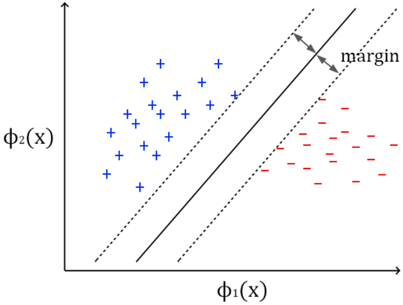
Support Vector Machines

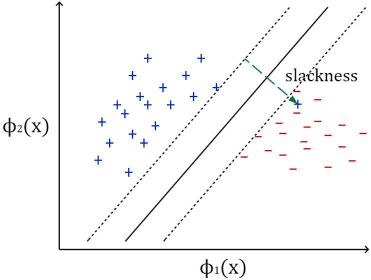
Instructor: Prof. Abir De



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

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

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

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