

# Support Vector Machine

## Primal Form

- A hyperplane that separates a n-dimensional space into two half-spaces.
- Prediction rule:  $y = \text{sign}(\omega^T x + b)$
- Margin
  - Geometric margin ( $\geq 0$ ):  $\gamma^{(i)} = y^{(i)}((\frac{\omega}{\|\omega\|})^T x^{(i)} + \frac{b}{\|\omega\|})$
  - Whole training set, the margin is  $\gamma = \min_i \gamma^{(i)}$
- Goal: Learn  $\omega$  and  $b$  that achieves the maximum margin  $\max_{\omega, b} \min_i \gamma^{(i)}$

$$\begin{aligned} \max_{\gamma, \omega, b} \quad & \gamma \\ \text{s.t.} \quad & y^{(i)}(\omega^T x^{(i)} + b) \geq \gamma \|\omega\|, \quad \forall i \end{aligned}$$

- Scaling  $(\omega, b)$  such that  $\gamma \|\omega\| = 1$ , the problem becomes

$$\begin{aligned} \max_{\omega, b} \quad & \frac{1}{\|\omega\|} \Leftrightarrow \min_{\omega, b} \omega^T \omega \Leftrightarrow \min_{\omega, b} \frac{1}{2} \|\omega\|^2 \\ \text{s.t.} \quad & y^{(i)}(\omega^T x^{(i)} + b) \geq 1, \quad \forall i \end{aligned}$$

- $\max_{\omega, b} \frac{1}{\|\omega\|}$  is equivalent to  $\min_{\omega, b} \omega^T \omega$

Def. The primal problem

$$\begin{aligned} \min_{\omega, b} \quad & \frac{1}{2} \|\omega\|^2 \\ \text{s.t.} \quad & y^{(i)}(\omega^T x^{(i)} + b) \geq 1, \quad \forall i \end{aligned}$$

## Duality of SVM

- The Lagrangian problem for SVM

$$\min_{\omega, b, \alpha} \mathcal{L}(\omega, b, \alpha) = \frac{1}{2} \|\omega\|^2 + \sum_{i=1}^m \alpha_i (1 - y^{(i)}(\omega^T x^{(i)} + b))$$

- The Lagrangian dual problem for SVM is  $\max_{\alpha} \mathcal{G}(\alpha) = \inf_{\omega, b} \mathcal{L}(\omega, b, \alpha)$

$$\begin{aligned} \max_{\alpha} \quad & \mathcal{G}(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)} \\ \text{s.t.} \quad & \sum_{i=1}^m \alpha_i y^{(i)} = 0 \\ & \alpha_i \geq 0 \quad \forall i \end{aligned}$$

- Proof.
  - $\frac{\partial}{\partial \omega} \mathcal{L}(\omega, b, \alpha) = \omega - \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} = 0$  and  $\frac{\partial}{\partial b} \mathcal{L}(\omega, b, \alpha) = -\sum_{i=1}^m \alpha_i y^{(i)} = 0$
  - $\mathcal{L}$  is a convex function.

- It suffices **Slarter's Condition**. Thus, the problem can be solved by QP solver (MATLAB, ...)
- Since we have the solution  $\alpha^*$  for the dual problem, we can calculate the solution for the primal problem.

$$\omega^* = \sum_{i=1}^m \alpha^* y^{(i)} x^{(i)} b^* = y^{(i)} - \omega^{*T} x^{(i)}, \text{ if } \alpha^* > 0$$

- For robustness, the optimal value for  $b$  is calculated by taking the averages across all  $b^*$

$$b^* = \frac{\sum_{i:\alpha_i^* > 0} (y^{(i)} - \omega^{*T} x^{(i)})}{\sum_{i=1}^m \mathbf{1}(\alpha_i^* > 0)}$$

- However, according to **Complementary Slackness**,  $\alpha_i^* [1 - y^{(i)} (\omega^{*T} x^{(i)} + b^*)] = 0$ .
- $\alpha_i^*$  is non-zero only if  $x^{(i)}$  lies on the margin, i.e.,  $y^{(i)} (\omega^{*T} x^{(i)} + b^*) = 1$ . (**Support Vector**,  $\mathcal{S}$ ).

$$\therefore \omega = \sum_{s \in \mathcal{S}} \alpha_s y^{(s)} x^{(s)}$$

## Kernel

- Basic idea: mapping data to higher dimensions where it exhibits linear patterns.
- Each kernel  $K$  has an associated feature mapping  $\phi : \mathcal{X} \rightarrow \mathcal{F}$  from input to feature space.
  - e.g., quadratic mapping  $\phi : x \rightarrow \{x_1^2, x_2^2, \dots, x_1 x_2, \dots, x_1 x_n, \dots, x_{n-1} x_n\}$
- Kernel  $K(x, z) = \phi(x)^T \phi(z)$ ,  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  **takes two inputs and gives their similarity** in  $\mathcal{F}$ .

Thereom. **Mercer's Condition**.

For  $K$  to be a kernel function if  $K$  is a positive definite function.

$$\int \int f(x) K(x, z) f(z) dx dz > 0 \forall f, \text{ s.t. } \int_{-\infty}^{\infty} f^2(x) dx < \infty$$

- Composing rules
  - Direct sum  $K(x, z) = K_1(x, z) + K_2(x, z)$
  - Scalar product  $K(x, z) = \alpha K_1(x, z)$
  - Direct product  $K(x, z) = K_1(x, z) K_2(x, z)$

Def. Kernel Matrix.

$$K_{i,j} = K(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^T \phi(x^{(j)})$$

### Example Kernel

- Linear (trivial) Kernel  $K(x, z) = x^T z$
- Quadratic Kernel  $K(x, z) = (x^T z)^2$  or  $(1 + x^T z)^2$
- Polynomial Kernel (of degree  $d$ )  $K(x, z) = (x^T z)^d$  or  $(1 + x^T z)^d$
- Gaussian Kernel  $K(x, z) = \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$
- Sigmoid Kernel  $K(x, z) = \tanh(\alpha x^T z + c)$

### Applicable Algorithm

- SVM, linear regression, etc.
- K-means, PCA, etc.
- For SVM

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)} \\ \text{s.t.} \quad & \sum_{i=1}^m \alpha_i y^{(i)} = 0 \\ & \alpha_i \geq 0 \quad \forall i \end{aligned} \quad \Rightarrow \quad \begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j K_{i,j} \\ \text{s.t.} \quad & \sum_{i=1}^m \alpha_i y^{(i)} = 0 \\ & \alpha_i \geq 0 \quad \forall i \end{aligned}$$

### Soft Margin

- Relax the constraints from  $y^{(i)}(\omega^T x^{(i)} + b) \geq 1$  to  $y^{(i)}(\omega^T x^{(i)} + b) \geq 1 - \xi_i$
- $\xi_i \geq 0$  is called slack variable

Def. **Soft Margin SVM**

$$\begin{aligned} \min_{\omega, b, \xi} \quad & \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & y^{(i)}(\omega^T x^{(i)} + b) \geq 1 - \xi_i, \quad \forall i = 1, \dots, m \\ & \xi_i \geq 0, \quad \forall i = 1, \dots, m \end{aligned}$$

- $C$  is a hyper-parameter that controls the relative weighting between  $\frac{1}{2} \|\omega\|^2$  for **larger margins** and  $\sum_{i=1}^m \xi_i$  for **fewer misclassified examples**.
- Lagrangian function

$$\mathcal{L}(\omega, b, \xi, \alpha, r) = \frac{1}{2} \omega^T \omega + C \sum_{i=1}^m \xi_i - \sum_{i=1}^m \alpha_i [y^{(i)} (\omega^T x^{(i)} + b) - 1 + \xi_i] - \sum_{i=1}^m r_i \xi_i$$

- KKT conditions (the optimal values of  $\omega, b, \xi, \alpha$ , and  $r$  should satisfy the following conditions)
  - $\nabla_{\omega} \mathcal{L}(\omega, b, \xi, \alpha, r) = 0 \Rightarrow \omega^* = \sum_{i=1}^m \alpha_i^* y^{(i)} x^{(i)}$
  - $\nabla_b \mathcal{L}(\omega, b, \xi, \alpha, r) = 0 \Rightarrow \sum_{i=1}^m \alpha_i^* y^{(i)} = 0$
  - $\nabla_{\xi_i} \mathcal{L}(\omega, b, \xi, \alpha, r) = 0 \Rightarrow \alpha_i^* + r_i^* = C$ , for  $\forall i$

- $\alpha_i^*, r_i^*, \xi_i^* \geq 0$ , for  $\forall i$
- $y^{(i)} (\omega^{*T} x^{(i)} + b^*) + \xi_i^* - 1 \geq 0$ , for  $\forall i$
- $\alpha_i^* (y^{(i)} (\omega^{*T} x^{(i)} + b^*) + \xi_i^* - 1) = 0$ , for  $\forall i$
- $r_i^* \xi_i^* = 0$ , for  $\forall i$

- Dual problem

$$\begin{aligned} \max_{\alpha} \quad & \mathcal{J}(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j < x^{(i)}, x^{(j)} > \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C, \quad \forall i = 1, \dots, m \\ & \sum_{i=1}^m \alpha_i y^{(i)} = 0 \end{aligned}$$

- Solution

$$\begin{aligned} - \omega^* &= \sum_{i=1}^m \alpha_i^* y^{(i)} x^{(i)} \\ - b^* &= \frac{\sum_{i: 0 < \alpha_i^* < C} (y^{(i)} - \omega^{*T} x^{(i)})}{\sum_{i=1}^m 1(0 < \alpha_i^* < C)} \end{aligned}$$

- Proof.

$$\begin{aligned} \because r_i^* \xi_i^* &= 0 \Leftrightarrow (C - \alpha_i^*) \xi_i^* = 0 \\ \therefore \forall i, \alpha_i^* &\neq C \Rightarrow \xi_i^* = 0 \Rightarrow \alpha_i^* (y^{(i)} (\omega^{*T} x^{(i)} + b^*) - 1) = 0 \\ \therefore \forall i, \alpha_i^* &\in (0, C) \Rightarrow y^{(i)} (\omega^{*T} x^{(i)} + b^*) = 1 \Rightarrow \omega^{*T} x^{(i)} + b^* = y^{(i)} \end{aligned}$$

- Corollaries of KKT conditions for soft-margin SVM

- When  $\alpha_i^* = 0$ ,  $y^{(i)} (\omega^{*T} x^{(i)} + b^*) \geq 1$ , correctly classified.
- When  $\alpha_i^* = C$ ,  $y^{(i)} (\omega^{*T} x^{(i)} + b^*) \leq 1$ , misclassified.
- When  $0 < \alpha_i^* < C$ ,  $y^{(i)} (\omega^{*T} x^{(i)} + b^*) = 1$ , support vector.