

Lecture 19.

SVM (III): Kernel Formulation

Outline

- Kernel representation
- Mercer's Theorem
- SVM using Kernels

Inner Product Kernels

In general, if the input is first transformed via a set of nonlinear functions $\{\phi_i(x)\}$ and then subject to the hyperplane classifier

$$g(x) = \sum_{j=1}^p w_j \phi_j(x) + b = \sum_{j=0}^p w_j \phi_j(x) = \underline{w}^T \underline{\phi} \quad b = w_0; \quad \phi_0(x) = 1$$

Define the inner product kernel (a scalar) as

$$k(x, y) = \sum_{j=0}^p \phi_j(x) \phi_j(y) = \underline{\phi} \underline{\phi}^T$$

one may obtain a dual optimization problem formulation as:

$$Q(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j d_i d_j k(x_i, x_j)$$

Often, \dim of $\underline{\phi}$ ($=p+1$) \gg \dim of x !

Polynomial Kernel

Consider a polynomial kernel

$$k(x, y) = (1 + x^T y)^2 = 1 + 2 \sum_{i=1}^m x_i y_i + 2 \sum_{i=1}^m \sum_{j=i+1}^m x_i y_i x_j y_j + \sum_{i=1}^m x_i^2 y_i^2$$

Let $K(x, y) = [k(x, y)] = \underline{\varphi}^T(x) \underline{\varphi}(y)$, then

$$\underline{\varphi}(x) = [1 \underbrace{x^2(1) x^2(2) \cdots x^2(m)}_{m \text{ terms}} \underbrace{\sqrt{2}x(1) \cdots \sqrt{2}x(m)}_{m \text{ terms}} \underbrace{\sqrt{2}x(1)x(2) \cdots \sqrt{2}x(m-1)x(m)}_{(m^2-m)/2 \text{ terms}}] = [\varphi_1(x) \cdots \varphi_p(x)]$$

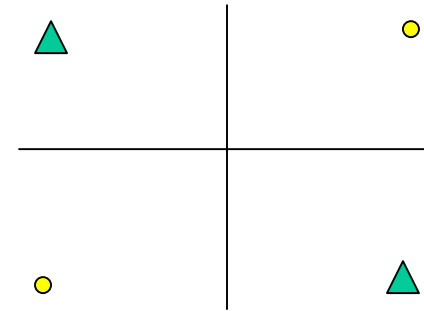
where $p = 1 + m + m + (m-1) + (m-2) + \dots + 1 = (m+2)(m+1)/2$
Hence, using a kernel, a low dimensional pattern classification problem (with dimension m) is solved in a higher dimensional space (dimension $p+1$). But only $\phi_j(x)$ corresponding to support vectors are used for pattern classification!

Numerical Example: XOR Problem

Training samples:

$$(-1 \ -1; -1), (-1 \ 1 \ +1),$$

$$(1 \ -1 \ +1), (1 \ 1 \ -1)$$



$\mathbf{x} = [x(1) \ x(2)]^T$. Use $k(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}^T \mathbf{y})^2$ one has
 $\varphi(\mathbf{x}) = [1 \ x^2(1) \ x^2(2) \ \sqrt{2} x(1), \ \sqrt{2} x(2), \ \sqrt{2} x(1)x(2)]^T$

$$\Phi = \begin{bmatrix} 1 & 1 & 1 & -\sqrt{2} & -\sqrt{2} & \sqrt{2} \\ 1 & 1 & 1 & -\sqrt{2} & \sqrt{2} & -\sqrt{2} \\ 1 & 1 & 0 & \sqrt{2} & -\sqrt{2} & -\sqrt{2} \\ 1 & 1 & 1 & \sqrt{2} & \sqrt{2} & \sqrt{2} \end{bmatrix} \quad K(\mathbf{x}_i, \mathbf{x}_j) = \Phi \Phi^T = \begin{bmatrix} 9 & 1 & 1 & 1 \\ 1 & 9 & 1 & 1 \\ 1 & 1 & 9 & 1 \\ 1 & 1 & 1 & 9 \end{bmatrix}$$

Note $\dim[\varphi(\mathbf{x})] = 6 > \dim[\mathbf{x}] = 2!$

$\text{Dim}(K) = N_s = \#$ of support vectors.

XOR Problem (Continued)

Note that $K(x_i, x_j)$ can be calculated directly without using Φ !

$$\text{E.g. } k_{1,1} = \left(1 + [-1 \ -1] \begin{bmatrix} -1 \\ -1 \end{bmatrix} \right)^2 = 9; \quad K_{1,2} = \left(1 + [-1 \ -1] \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right)^2 = 1$$

The corresponding Lagrange multiplier $\alpha = (1/8)[1 \ 1 \ 1 \ 1]^T$.

$$\begin{aligned} W &= \sum_{i=1}^N \alpha_i d_i \varphi(\mathbf{x}_i) = \Phi^T [\alpha_1 d_1 \quad \alpha_2 d_2 \quad \cdots \quad \alpha_N d_N]^T \\ &= \frac{1}{8}(-1)\varphi(\mathbf{x}_1) + \frac{1}{8}(1)\varphi(\mathbf{x}_2) + \frac{1}{8}(1)\varphi(\mathbf{x}_3) + \frac{1}{8}(-1)\varphi(\mathbf{x}_4) \\ &= \left[0 \quad 0 \quad 0 \quad 0 \quad 0 \quad -\frac{1}{\sqrt{2}} \right]^T \end{aligned}$$

Hence the hyper-plane is: $y = \mathbf{w}^T \varphi(\mathbf{x}) = -x_1 x_2$

(x_1, x_2)	$(-1, -1)$	$(-1, +1)$	$(+1, -1)$	$(+1, +1)$
$y = -1 x_1 x_2$	-1	$+1$	$+1$	-1

Other Types of Kernels

type of SVM	$K(\mathbf{x}, \mathbf{y})$	Comments
Polynomial learning machine	$(\mathbf{x}^T \mathbf{y} + 1)^p$	p : selected a priori
Radial basis function	$\exp\left(-\frac{1}{2\sigma^2} \ \mathbf{x} - \mathbf{y}\ ^2\right)$	σ^2 : selected a priori
Two-layer perceptron	$\tanh(\beta_0 \mathbf{x}^T \mathbf{y} + \beta_1)$	only some β_0 and β_1 values are feasible.

What kernel is feasible? It must satisfy the "Mercer's theorem"!

Mercer's Theorem

Let $K(\mathbf{x}, \mathbf{y})$ be a continuous, symmetric kernel, defined on $a \leq \mathbf{x}, \mathbf{y} \leq b$. $K(\mathbf{x}, \mathbf{y})$ admits an eigen-function expansion

$$K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{y})$$

with $\lambda_i > 0$ for each i . This expansion converges absolutely and uniformly if and only if

$$\int_b^a \int_b^a K(\mathbf{x}, \mathbf{y}) \psi(\mathbf{x}) \psi(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0$$

for all $\psi(\mathbf{x})$ such that $\int_b^a \psi^2(\mathbf{x}) d\mathbf{x} < \infty$

Testing with Kernels

For many types of kernels, $\varphi(\mathbf{x})$ can not be explicitly represented or even found. However,

$$W = \sum_{i=1}^N \alpha_i d_i \varphi(\mathbf{x}_i) = \Phi^T [\alpha_1 d_1 \quad \alpha_2 d_2 \quad \cdots \quad \alpha_N d_N]^T = \Phi^T f$$

$$y(x) = W^T \varphi(\mathbf{x}) = (\Phi^T f)^T \varphi(\mathbf{x}) = f^T K(\mathbf{x}_i, \mathbf{x}) = K(\mathbf{x}, \mathbf{x}_i) f$$

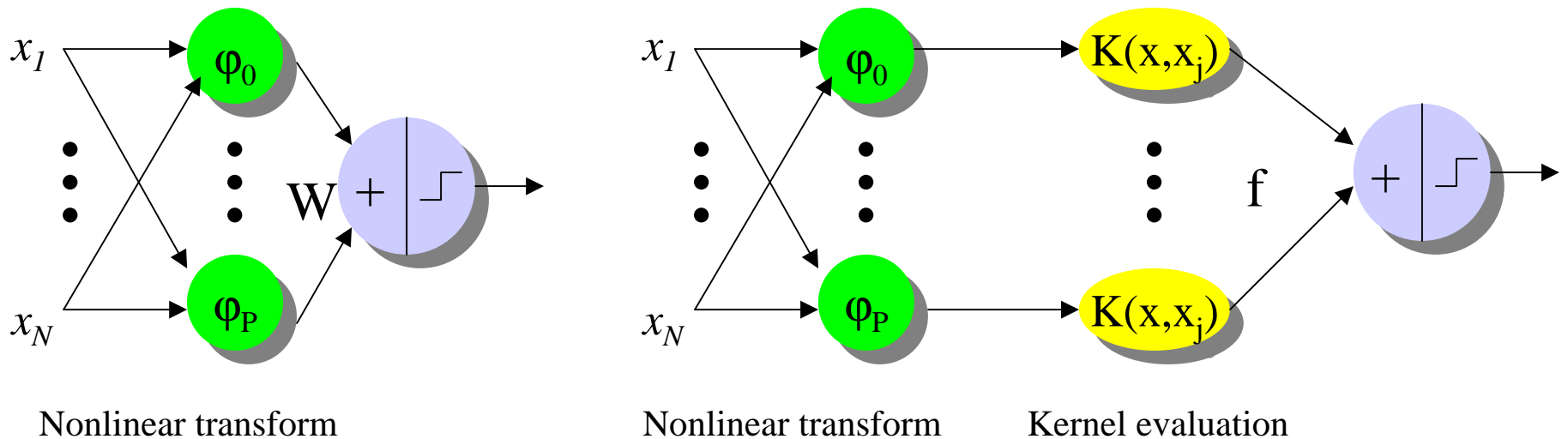
Hence there is no need to know $\varphi(\mathbf{x})$ explicitly! For example, in the XOR problem, $f = (1/8)[-1 \quad +1 \quad +1 \quad -1]^T$. Suppose that $\mathbf{x} = (-1, +1)$, then

$$y(x) = K(x, x_j) f$$

$$= \begin{bmatrix} 1 + [-1 \quad -1] \begin{bmatrix} -1 \\ 1 \end{bmatrix} \\ 1 + [-1 \quad 1] \begin{bmatrix} -1 \\ 1 \end{bmatrix} \\ 1 + [1 \quad -1] \begin{bmatrix} -1 \\ 1 \end{bmatrix} \\ 1 + [1 \quad 1] \begin{bmatrix} -1 \\ 1 \end{bmatrix} \end{bmatrix} \begin{bmatrix} -1/8 \\ 1/8 \\ 1/8 \\ -1/8 \end{bmatrix}$$

$$= [1 \quad 9 \quad 1 \quad 1] [-1/8 \quad 1/8 \quad 1/8 \quad -1/8]^T = 1$$

SVM Using Nonlinear Kernels



Using kernel, low dimensional feature vectors will be mapped to high dimensional (may be infinite dim) kernel feature space where the data are likely to be linearly separable.