CSE 5526: Introduction to Neural Networks

Support Vector Machines (SVMs)

Part 3: Kernels

Kernels are generalizations of inner products

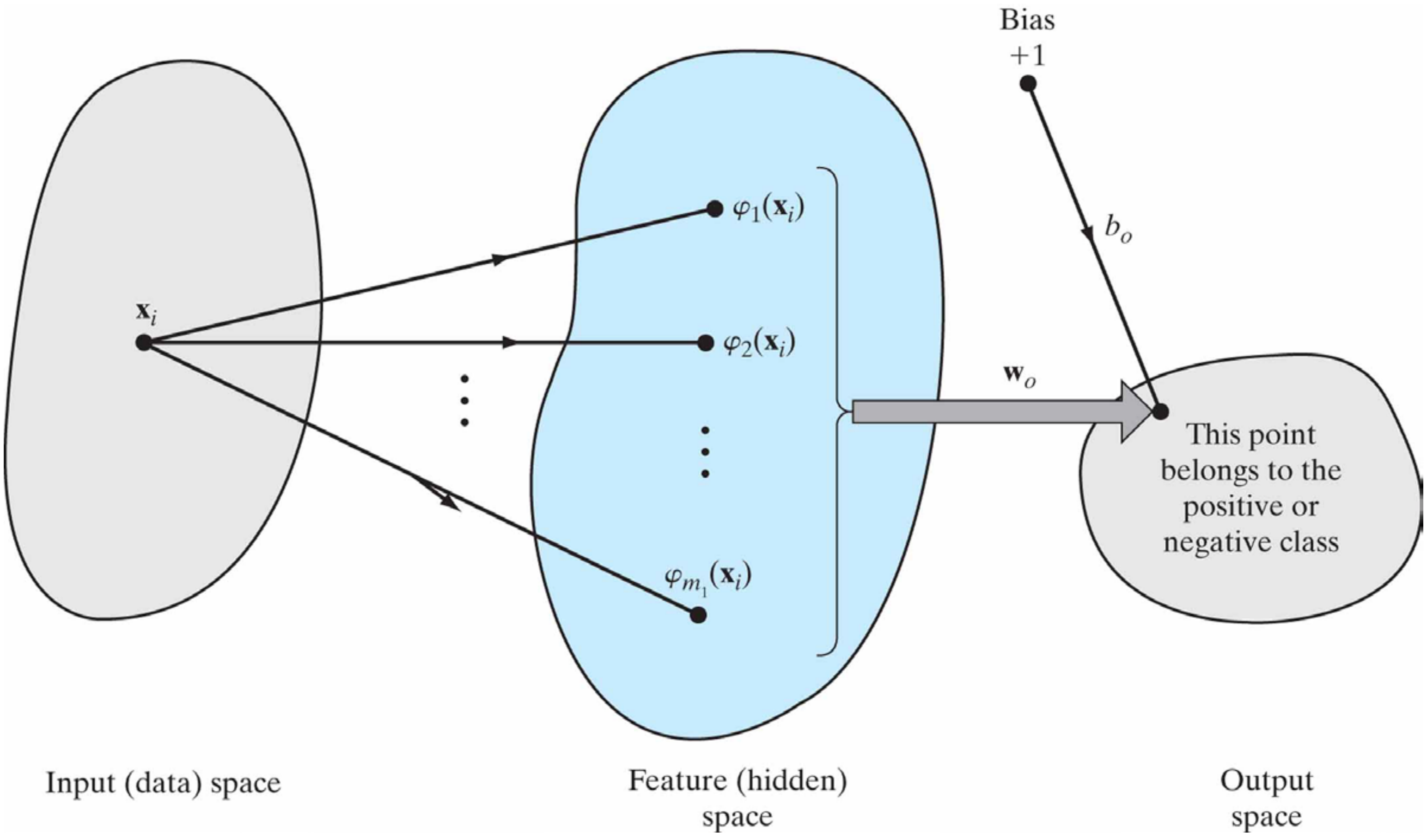
- A kernel is a function of two data points such that
$$k(x, x') = \phi^T(x)\phi(x')$$
For some function $\phi(x)$
- It is therefore symmetric: $k(x, x') = k(x', x)$
- Can compute $k(x, x')$ from an explicit $\phi(x)$
- Or prove that $k(x, x')$ corresponds to some $\phi(x)$
  - Never need to actually compute $\phi(x)$

SVM as a kernel machine

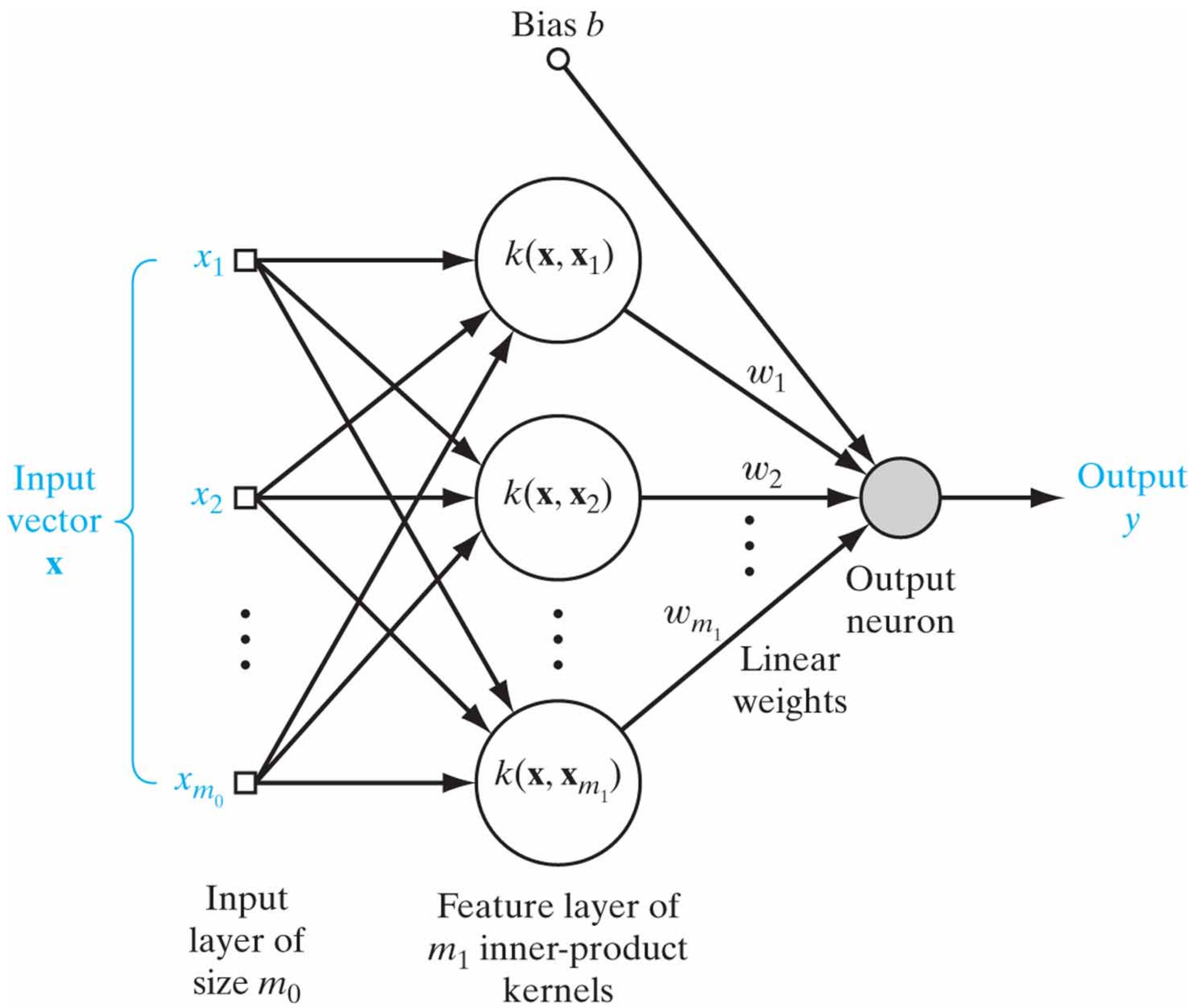
- **Cover's theorem**: A complex classification problem, cast in a high-dimensional space nonlinearly, is more likely to be linearly separable than in the low-dimensional input space

- SVM for pattern classification
  1. Nonlinear mapping of the input space onto a high-dimensional feature space
  2. Constructing the optimal hyperplane for the feature space

Kernel machine illustration



Kernelized SVM looks a lot like an RBF net



Kernel matrix

- The matrix

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_N) \\ & \vdots & \\ \ldots & k(\mathbf{x}_i, \mathbf{x}_j) & \ldots \\ & \vdots & \\ k(\mathbf{x}_N, \mathbf{x}_1) & \ldots & k(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}$$

is called the kernel matrix, or the Gram matrix.

- **K** is positive semidefinite

Mercer's theorem relates kernel functions and inner product spaces

- Suppose that for all finite sets of points $\{\boldsymbol{x}_p\}_{p=1}^{N}$ and real numbers $\{\boldsymbol{a}\}_{p=1}^{\infty}$

$$\sum_{i,j} a_j a_i k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) \geq 0$$

- Then $\boldsymbol{K}$ is called a positive semidefinite kernel
- And can be written as

$$k(\boldsymbol{x}, \boldsymbol{x}') = \phi^T(\boldsymbol{x})\phi(\boldsymbol{x}')$$

- For some vector-valued function $\phi(\boldsymbol{x})$

Kernels can be applied in many situations

- Kernel trick: when predictions are based on inner products of data points, replace with kernel function
- Some algorithms where this is possible
  - Linear / ridge regression
  - Principal components analysis
  - Canonical correlation analysis
  - Perceptron classifier

Some popular kernels

- Polynomial kernel, parameters c and p

$$k(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^T \boldsymbol{x}' + c)^p$$

  - Finite-dimensional $\phi(\boldsymbol{x})$ can be explicitly computed

- Gaussian or RBF kernel, parameter σ

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{1}{2\sigma} \|\boldsymbol{x} - \boldsymbol{x}'\|^2\right)$$

  - Infinite-dimensional $\phi(\boldsymbol{x})$
  - Equivalent to RBF network, but more principled way of finding centers

Some popular kernels

- Hypebolic tangent kernel, parameters β_1 and β_2

$$k(\boldsymbol{x}, \boldsymbol{x}') = \tanh(\beta_1 \boldsymbol{x}^T \boldsymbol{x}' + \beta_2)$$

  - Only positive semidefinite for some values of β_1 and β_2
  - Inspired by neural networks, but more principled way of selecting number of hidden units

- String kernels or other structure kernels
  - Can prove that they are positive definite
  - Computed between non-numeric items
  - Avoid converting to fixed-length feature vectors

Example: polynomial kernel

- Polynomial kernel in 2D, $c = 1, p = 2$

$$k(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^T\boldsymbol{x}' + 1)^2 = (x_1x_1' + x_2x_2' + 1)^2$$

$$= x_1^2{x_1'}^2 + x_2^2{x_2'}^2 + 2x_1x_1'x_2x_2' + 2x_1x_1' + 2x_2x_2' + 1$$

- If we define

$$\phi(\boldsymbol{x}) = \left[x_1^2, x_2^2, \sqrt{2}\mathrm{x}_1\mathrm{x}_2, \sqrt{2}\mathrm{x}_1, \sqrt{2}\mathrm{x}_2, 1\right]^T$$

- Then $k(\boldsymbol{x}, \boldsymbol{x}') = \phi^T(\boldsymbol{x})\phi(\boldsymbol{x}')$

Example: XOR problem again

- Consider (once again) the XOR problem
- The SVM can solve it using a polynomial kernel
  - With $p = 2$ and $c = 1$

TABLE 6.2 XOR Problem

Input vector $\mathbf{x}$	Desired response d
$(-1, -1)$	-1
$(-1, +1)$	$+1$
$(+1, -1)$	$+1$
$(+1, +1)$	-1

XOR: first compute the kernel matrix

- In general, $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(1 + \boldsymbol{x}_i^T \boldsymbol{x}_j\right)^2$
- For example,

$$K_{11} = k\left(\begin{bmatrix}-1\\-1\end{bmatrix}, \begin{bmatrix}-1\\-1\end{bmatrix}\right) = (1+2)^2 = 9$$

$$K_{12} = k\left(\begin{bmatrix}-1\\-1\end{bmatrix}, \begin{bmatrix}-1\\+1\end{bmatrix}\right) = (1+0)^2 = 1$$

- So

$$K = \begin{bmatrix} 9 & 1 & 1 & 1 \\ 1 & 9 & 1 & 1 \\ 1 & 1 & 9 & 1 \\ 1 & 1 & 1 & 9 \end{bmatrix}$$

XOR: first compute the kernel matrix

- Or compute $\phi(\boldsymbol{x}_i)$ and their inner products, e.g.,
  - Remember, $\phi(\boldsymbol{x}) = \left[x_1^2, x_2^2, \sqrt{2}\mathrm{x}_1\mathrm{x}_2, \sqrt{2}\mathrm{x}_1, \sqrt{2}\mathrm{x}_2, 1\right]^T$ For b
  - Since $\phi(\boldsymbol{x})$ includes 1, no need for separate b later

$$\phi(\boldsymbol{x}_1) = \phi\left(\begin{bmatrix}-1\\-1\end{bmatrix}\right) = \left[1,1,\sqrt{2},-\sqrt{2},-\sqrt{2},1\right]^T$$

$$\phi(\boldsymbol{x}_2) = \phi\left(\begin{bmatrix}-1\\+1\end{bmatrix}\right) = \left[1,1,-\sqrt{2},-\sqrt{2},\sqrt{2},1\right]^T$$

- Then

$$K_{11} = \phi^T(\boldsymbol{x}_1)\phi(\boldsymbol{x}_1) = 1+1+2+2+2+1 = 9$$
$$K_{12} = \phi^T(\boldsymbol{x}_1)\phi(\boldsymbol{x}_2) = 1+1-2+2-2+1 = 1$$

- Results in same K matrix, but more computation

XOR: Combine class labels into K

- Define matrix $\widetilde{K}$ such that $\widetilde{K}_{ij} = K_{ij} d_i d_j$
- Recall $\boldsymbol{d} = [-1, +1, +1, -1]^T$

$$\widetilde{K} = \begin{bmatrix} +9 & -1 & -1 & +1 \\ -1 & +9 & +1 & -1 \\ -1 & +1 & +9 & -1 \\ +1 & -1 & -1 & +9 \end{bmatrix}$$

XOR: Solve dual Lagrangian for $\boldsymbol{a}$

- Find fixed points of
$$\tilde{L}(\boldsymbol{a}) = \mathbf{1}^T \boldsymbol{a} - \frac{1}{2} \boldsymbol{a}^T \tilde{K} \boldsymbol{a}$$
- Set matrix gradient to 0
$$\nabla \tilde{L} = \mathbf{1} - \tilde{K} \boldsymbol{a} = \mathbf{0}$$
$$\Rightarrow \boldsymbol{a} = \tilde{K}^{-1} \mathbf{1} = \left[\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right]^T$$
- Satisfies all conditions: $a_p \geq 0 \; \forall p \quad \sum_p a_p d_p = 0$
  - So this is the solution
- All points are support vectors

XOR: Compute $\boldsymbol{w}$ (including b) from $\boldsymbol{a}$

$$\boldsymbol{w} = \sum_p a_p d_p \boldsymbol{x}_p$$

$$= -\frac{1}{8}\phi(\boldsymbol{x}_1) + \frac{1}{8}\phi(\boldsymbol{x}_2) + \frac{1}{8}\phi(\boldsymbol{x}_3) - \frac{1}{8}\phi(\boldsymbol{x}_4)$$

$$= \frac{1}{8}\left(-\begin{bmatrix} 1 \\ 1 \\ \sqrt{2} \\ -\sqrt{2} \\ -\sqrt{2} \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ -\sqrt{2} \\ -\sqrt{2} \\ \sqrt{2} \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ -\sqrt{2} \\ \sqrt{2} \\ -\sqrt{2} \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ \sqrt{2} \\ \sqrt{2} \\ \sqrt{2} \\ 1 \end{bmatrix} \right) = \begin{bmatrix} 0 \\ 0 \\ -\frac{1}{\sqrt{2}} \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{array}{l} \\ \\ \\ \\ \\ b \end{array}$$

XOR: Examine prediction function

- Prediction function

$$y(\boldsymbol{x}) = \boldsymbol{w}^T \phi(\boldsymbol{x})$$

$$= \left[0,0,-\frac{1}{\sqrt{2}},0,0,0\right]^T \left[x_1^2, x_2^2, \sqrt{2}\mathrm{x}_1\mathrm{x}_2, \sqrt{2}\mathrm{x}_1, \sqrt{2}\mathrm{x}_2, 1\right]$$

$$= -x_1 x_2$$

- Predictions are based on product of the dimensions

$$y(\boldsymbol{x}_1) = -(-1)(-1) = -1$$
$$y(\boldsymbol{x}_2) = -(-1)(+1) = +1$$
$$y(\boldsymbol{x}_3) = -(+1)(-1) = +1$$
$$y(\boldsymbol{x}_4) = -(+1)(+1) = -1$$

XOR: Decision boundaries

- Decision boundary at $y(\boldsymbol{x}) = -x_1 x_2 = 0$
- Support vectors at $y(\boldsymbol{x}) = -x_1 x_2 = 1$