Radial Basis Function (RBF) Networks

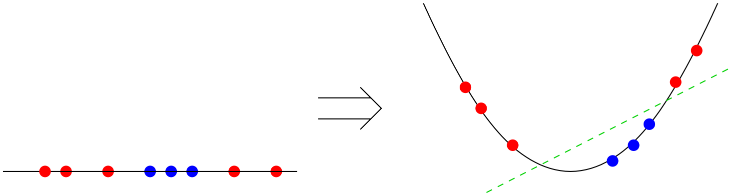
Seungjin Choi

Department of Computer Science and Engineering
Pohang University of Science and Technology
77 Cheongam-ro, Nam-gu, Pohang 37673, Korea
seungjin@postech.ac.kr

Cover's Theorem

Theorem

A complex pattern classification problem cast in a high-dimensional space nonlinearly, is more likely to be linearly separable than in a low-dimensional space.



Basis Functions and Feature Space

Let $\mathcal{X}$ be a set of N m-dimensional patterns, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, each of which is assigned to be one of two classes, C_1 and C_2.
Define a function $\varphi(\boldsymbol{x})$ as $\varphi(\boldsymbol{x}) = [\varphi_1(\boldsymbol{x}), \ldots, \varphi_r(\boldsymbol{x})]^\top$, $\varphi(\boldsymbol{x})$: $\mathbb{R}^m \mapsto \mathbb{R}^r$.

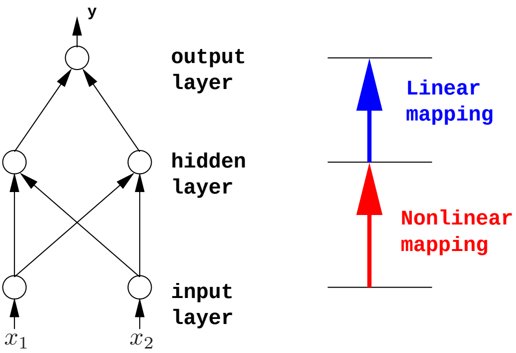
- Basis functions: $\{\varphi_i(\boldsymbol{x})\}$ (hidden functions)
- Feature space: The space spanned by a set of basis functions, $\{\varphi_i(\boldsymbol{x})\}_{i=1}^r$ (hidden space)

A dichotomy $\{C_1, C_2\}$ of $\mathcal{X}$ is said to be φ-separable if there exists $\boldsymbol{w} \in \mathbb{R}^r$ such that

$$\begin{aligned} \boldsymbol{w}^\top \varphi(\boldsymbol{x}) &> 0 \quad \text{if } \boldsymbol{x} \in C_1, \\ \boldsymbol{w}^\top \varphi(\boldsymbol{x}) &< 0 \quad \text{if } \boldsymbol{x} \in C_2. \end{aligned}$$

The separating surface (decision boundary) is given by $\boxed{\boldsymbol{w}^\top \varphi(\boldsymbol{x}) = 0}$.

RBF Net Structure



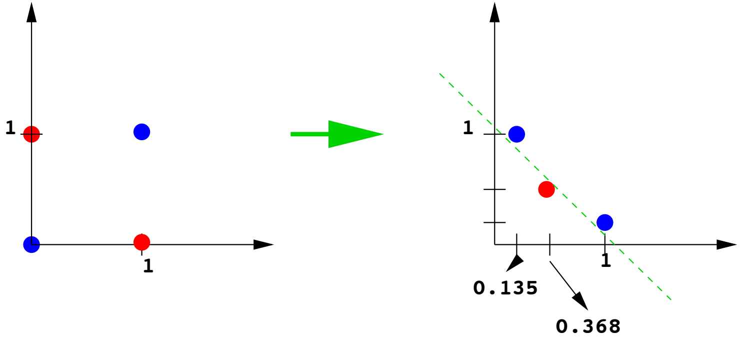
- Nonlinear mapping: Determine the parameters governing the basis functions and do nonlinear mapping $\Rightarrow$ $\varphi_i(\boldsymbol{x}) = \exp\left\{-\lambda_i \|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2\right\}$.
- Linear mapping: Output is a linear combination of basis functions. Determine weights by solving a linear problem (through LS) $\Rightarrow$ $y = w_1\varphi_1(\boldsymbol{x}) + w_2\varphi_2(\boldsymbol{x})$.

Example: XOR

Consider basis functions given by

$$\varphi_1(\boldsymbol{x}) = \exp\left\{-\|\boldsymbol{x} - \boldsymbol{\mu}_1\|^2\right\}, \quad \varphi_2(\boldsymbol{x}) = \exp\left\{-\|\boldsymbol{x} - \boldsymbol{\mu}_2\|^2\right\},$$

where $\boldsymbol{\mu}_1 = [1, 1]^\top$ and $\boldsymbol{\mu}_2 = [0, 0]^\top$.



Exact Interpolation: RBF Approach

- Exact interpolation: Every input vector is mapped exactly onto the corresponding target vector.
- Consider a mapping, $h(\boldsymbol{x}) : \mathbb{R}^m \mapsto \mathbb{R}$. Given a data set $\{(\boldsymbol{x}_1, t_1), \ldots, (\boldsymbol{x}_N, t_N)\}$, the goal of exact interpolation is to find a function $h(\boldsymbol{x})$ such that $h(\boldsymbol{x}_i) = t_i$ for $i = 1, \ldots, N$.
- RBF approach: $\boxed{h(\boldsymbol{x}) = \sum_{i=1}^N w_i \, \varphi(\|\boldsymbol{x} - \boldsymbol{x}_i\|)}$.
- Define $\boldsymbol{\Phi} = [\Phi_{ij}] = [\varphi(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|) \in \mathbb{R}^{N \times N}$, $\boldsymbol{w} = [w_1, \ldots, w_N]^\top \in \mathbb{R}^N$, and $\boldsymbol{t} = [t_1, \ldots, t_N]^\top \in \mathbb{R}^N$. With these definitions, the interpolation condition leads to

  $$\boxed{\boldsymbol{\Phi} \boldsymbol{w} = \boldsymbol{t}}.$$

  Thus,

  $$\boxed{\boldsymbol{w} = \boldsymbol{\Phi}^{-1} \boldsymbol{t}}.$$

Several Forms of Basis Functions

1. Gaussian function, $\varphi(x) = \exp\left\{-\frac{x^2}{2\sigma^2}\right\}$.
2. $\varphi(x) = \left(x^2 + \sigma^2\right)^{-\alpha}$, $\alpha > 0$.
3. $\varphi(x) = x^2 \log x$.
4. $\varphi(x) = \left(x^2 + \sigma^2\right)^{\beta}$, $0 < \beta < 1$.
5. $\varphi(x) = x^3$.
6. $\varphi(x) = x$.

From Exact Interpolation to RBF Net

- The exact interpolation requires that each RBF passes exactly through every data point.
  - Too complex.
  - Highly oscillatory for noisy data.
- The RBF net is based on some modifications:
  - The number of basis functions is much less than the number of data points.
  - The centers of the basis functions are no longer constrained to be given by input data vectors. Instead, the determination of suitable centers becomes a part of the training process. Typically we use a clustering method (for example, k-means) to determine centers of basis functions.
  - Instead of having a common width parameter σ, each basis function is given its own width σ_j whose value is also determined during training. (kernel width selection)

RBF Net Mapping

The output of a RBF net is given by

$$
\begin{aligned}
y_k(\boldsymbol{x}) &= \sum_{j=1}^{M} w_{kj}\varphi_j(\boldsymbol{x}) + w_{k0} \\
&= \sum_{j=0}^{M} w_{kj}\varphi_j(\boldsymbol{x}). \quad (\varphi_0 = 1)
\end{aligned}
$$

Define

$$
\begin{aligned}
\boldsymbol{y} &= [y_1, \ldots, y_c]^\top, \\
\boldsymbol{\varphi} &= [\varphi_1, \ldots, \varphi_M]^\top, \\
\boldsymbol{W} &= [w_{kj}].
\end{aligned}
$$

With these definitions, we have

$$
\boxed{\boldsymbol{y} = \boldsymbol{W}\boldsymbol{\varphi}}.
$$

RBF Net Training

The error function is

$$\mathcal{E} = \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{M} \left\{ y_k(\boldsymbol{x}_i) - t_k(i) \right\}^2 .$$

Solving $\frac{\partial \mathcal{E}}{\partial \boldsymbol{W}} = 0$, leads to

$$\boldsymbol{\Phi}^\top \boldsymbol{\Phi} \boldsymbol{W}^\top = \boldsymbol{\Phi}^\top \boldsymbol{T},$$

where $[\boldsymbol{T}]_{ik} = t_k(i)$ and $[\boldsymbol{\Phi}]_{ij} = \varphi_j(\boldsymbol{x}_i)$.
The weight matrix $\boldsymbol{W}$ is found by

$$\boxed{\boldsymbol{W}^\top = \left(\boldsymbol{\Phi}^\top \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{T}}.$$