Lecture 25

Radial Basis Network (II)

Outline

- Regularization Network Formulation
- Radial Basis Network Type 2

Regularization Problem Formulation

$$E(F) = E_s(F) + \lambda E_r(F) = \frac{1}{2}\sum_{i=1}^{K}[d_i - F(\mathbf{x}_i)]^2 + \frac{1}{2}\lambda \| \mathbf{P} F \|^2 \qquad (4)$$

where **P** is a linear differential operator which imposes some smoothness constraints on the solution of F. The first term is *standard error term*, and the second term is the *regularization term*.

- The solution to above regularization problem is

$$F(\mathbf{x}) = \sum_{i=1}^{N} w_i \, G(\mathbf{x}; \mathbf{x}_i)) \qquad (5)$$

The Solution – Green's function

- $G(\mathbf{x}; \mathbf{x}_i)$ is called the Green's function corresponding to the self-adjoin differential operator $\mathbf{P}^*\mathbf{P}$ such that

$$\mathbf{P}^*\mathbf{P}\, G(\mathbf{x}; \mathbf{x}i) = \delta(\mathbf{x} - \mathbf{x}i) \qquad (6)$$

  where $\delta(\bullet)$ is the delta (distribution) function which vanishes everywhere except the origin.

- Let matrix $\mathbf{G} = [G(\mathbf{x}_i; \mathbf{x}_j)]$, then (5) leads to a solution to $\mathbf{w}$ as

$$\mathbf{w} = (\mathbf{G} + \lambda\mathbf{I})^{-1}\mathbf{d} \qquad (7)$$

  Note the effect of regularization factor λ now appear in the solution.

Gaussian Basis

- Under the special differential operator $\|\mathbf{P}F\|^2 = \sum_{m=0}^{M} a_m\|D^m F(\mathbf{x})\|2$, a solution to the Green's function is:

$$G(\mathbf{x}) = \exp(-\tfrac{1}{2}\|\mathbf{x}\|^2), \qquad (8)$$

and $$G(\mathbf{x};\mathbf{x}_i) = G(\|\mathbf{x}-\mathbf{x}_i\|) \qquad (9)$$

hence, $$F(\mathbf{x}) = \sum_{i=1}^{K} w_i \bullet \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{x}_i\|^2\right) \qquad (10)$$

Properties of Regularization network

- The regularization function in (10) is a *universal approximator* in that it can approximate arbitrarily well any multivariate continuous function on a compact subset of R^N where N is the dimension of feature vectors, given sufficient number of hidden neurons.

  It is also has the best approximation property, and is optimal in that it minimizes E(F).

- In (10), K is the number of available (training) data samples. When K is too large, often we want to use a smaller value J (< K) such that

$$F(\mathbf{x}) = \sum_{j=1}^{J} w_j \bullet \exp(-\|\mathbf{x} - \mathbf{t}_j\|^2/(2\sigma_j^2)) \qquad (11)$$

  There are J(2+N) parameters to solve if J is given.

Clustering Based Solution

- To cluster the K training samples into J clusters to yield a solution to $\{\mathbf{t}_j, \sigma_j^2;\ 1 \le j \le J\}$. In particular, $\mathbf{t}_j$ is the centroid of j^{th} cluster, and σ_j^2 is the within-cluster variance which is the sum of the diagonal of the within-cluster scattering matrix.
- The $\{w_j;\ 1 \le j \le J\}$ is the least square solution:

$$\mathbf{w} = \mathbf{G}^{\dagger}\mathbf{d} = (G^T G)^{-1} G^T d \qquad (12)$$

where $\mathbf{G}^{\dagger}$ is the pseudo-inverse of the matrix $\mathbf{G} = [G(i,j)]$ such that

$$G(i,j) = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}_i - \mathbf{t}_j\|^2\right) \qquad (13)$$

Tuning Clustering Parameters

- Only the standard error term is used. Define

$$e(i) = d_i - F(\mathbf{x}_i) = d_i - \sum_{j=1}^{J} w_j \bullet \exp\left(-\frac{1}{2\sigma_j^2}\|\mathbf{x}_i - \mathbf{t}_j\|^2\right), \qquad (15)$$

- Goal: to minimize $e = \frac{1}{2}\sum_{i=1}^{K} [e(i)]^2$.
- $$\nabla_{wj}\, e = \sum_{i=1}^{K} e(i)\bullet\exp\left(-\frac{1}{2\sigma_j^2}\|\mathbf{x}_i - \mathbf{t}_j\|^2\right) \qquad (16)$$

$$\nabla_{\mathbf{t}j}\, e = \sum_{i=1}^{K} e(i)\bullet w_j \exp\left(-\frac{1}{2\sigma_j^2}\|\mathbf{x}_i - \mathbf{t}_j\|^2\right) \frac{1}{2\sigma_j^2}[\mathbf{x}_i - \mathbf{t}_j] \quad (17)$$

$$\nabla_{\sigma_j^2}\, e = \sum_{i=1}^{K} e(i)\bullet w_j \exp\left(-\frac{1}{2\sigma_j^2}\|\mathbf{x}_i - \mathbf{t}_j\|^2\right) \frac{1}{2\sigma_j^4}\|\mathbf{x}_i - \mathbf{t}_j\|^2 \quad (18)$$

RBN2 Algorithm Summary

Given: $\{x_i;\ 1 \le i \le K\}$, d: desired output, and J: # of radial basis neurons

- Cluster $\{x_i\}$ into J clusters, find clustering centers $\{t_j; 1 \le j \le J\}$. Variance σ_j^2 or covariance matrix Σ_j^{-1} are also computed. It is better to force $\sigma_j^2 \approx 1$ or $\Sigma_j \approx I$ (identify matrix) to ensure better performance.
- Compute G matrix (K by c)

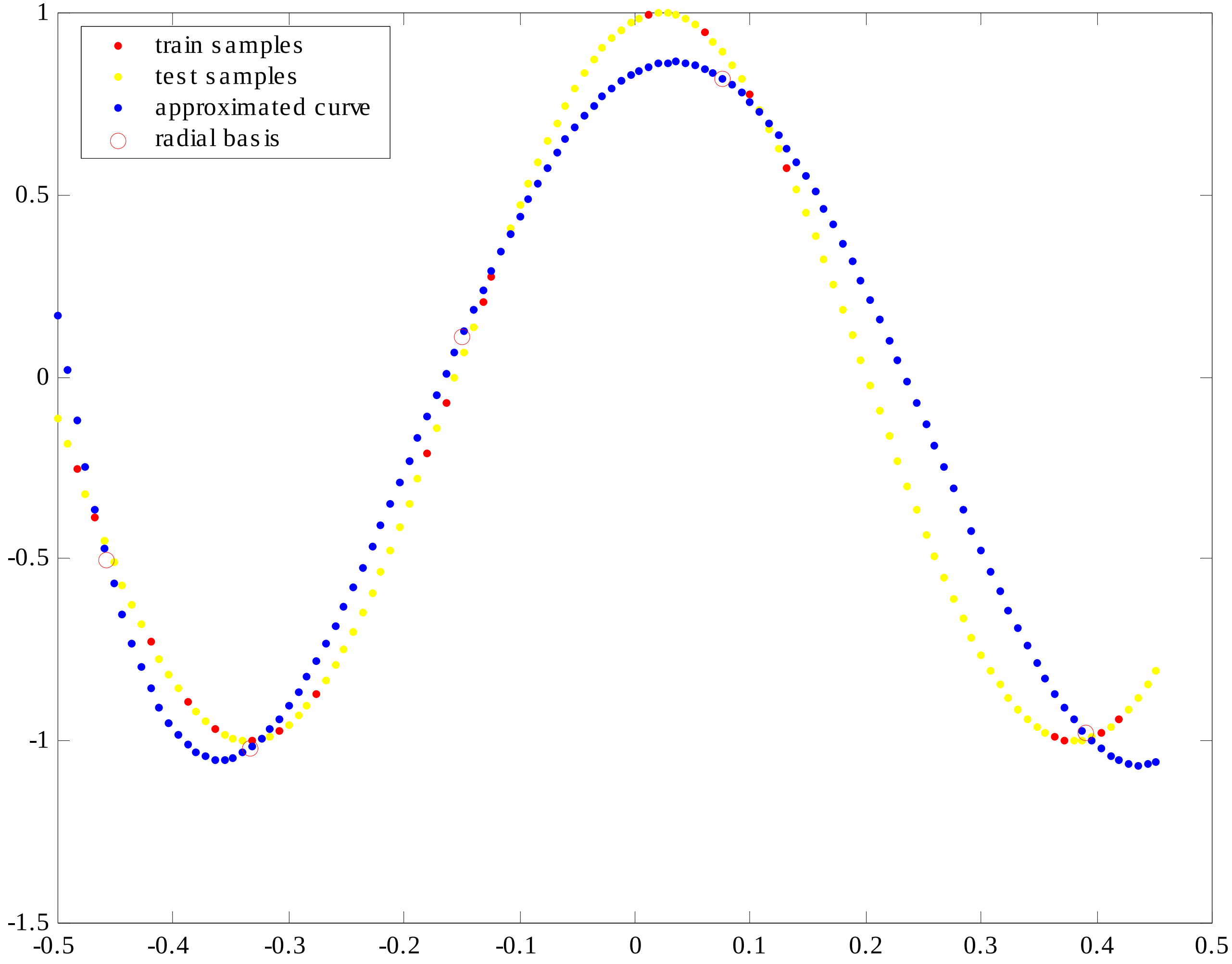
$G_{i,j+1} = \exp(-0.5\|x(i)-t_j\|^2/\sigma_j^2)$ or

$G_{i,j+1} = \exp(-0.5(x(i)-t_j)^T\Sigma_j^{-1}(x(i)-t_j))$

Solve $w = G^\dagger d$

- If desired, fine tune W, T, and Σ by eqs. (16)–(18)

Example


train samples
test samples
approximated curve
radial basis
1
0.5
0
-0.5
-1
-1.5
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5