

Feature Selection through Functional Approximation

- Suppose the features of a pattern class can be characterized by a function $f(x)$.
- Given the training data:

$$(x_1, f(x_1)), \dots, (x_N, f(x_N))$$

Find a functional approximation $\hat{f}(x)$ to $f(x)$ according to a certain criterion.

- Functional approximation:
 - functional expansion
 - stochastic approximation
 - kernel approximation

- Functional expansion:

Given M classes $f_i(x)$ — the feature function of the i -th class

training data for the i -th class:

$$\{(x_1^i, y_1^i), \dots, (x_{N_i}^i, y_{N_i}^i)\}$$

Error criterion:

$$e^i = \sum_{k=1}^{N_i} w_k^i [y_k^i - \hat{f}_i(x_k^i)]^2, \quad i = 1, \dots, M$$

- $\hat{f}_i(x)$ — a linear combination of basic functions

$$\begin{aligned} \hat{f}_i(x) &= \sum_{j=1}^{m_i} c_{ij} \varphi_{ij}(x) \\ &= c_i' \varphi_i(x) \end{aligned}$$

$$c_i = (c_{i1}, \dots, c_{im_i})'$$

$$\varphi_i = (\varphi_{i1}, \dots, \varphi_{im_i})'$$

•

$$[\varphi_i(x_1^i), \dots, \varphi_i(x_{N_i}^i)]_{m_i \times N_i}$$

has a rank m_i ($\ll N_i$).

•

$$\frac{\partial e^i}{\partial c_i} = 0 :$$

$$\begin{aligned} \frac{\partial e^i}{\partial c_i} &= \frac{\partial}{\partial c_i} \sum_{k=1}^{N_i} w_k^i [y_k^i - c_i' \varphi_i(x_k^i)]^2 \\ &= - \sum_k w_k^i [y_k^i - c_i' \varphi_i(x_k^i)] \varphi_i(x_k^i) \\ &= 0 \Rightarrow \end{aligned}$$

$$\underbrace{\left(\sum_k w_k^i \varphi_i(x_k^i) \varphi_i'(x_k^i) \right)}_{(B_i)_{m_i \times m_i}} c_i = \underbrace{\sum_k w_k^i y_k^i \varphi_i(x_k^i)}_{(v_i)_{m_i \times 1}}$$

-

$$c_i = B_i^{-1}v_i$$

- If $B_i = \text{diag}\{\lambda_1, \dots, \lambda_{m_i}\}$,

$$c_i = \text{diag}\{\lambda_1^{-1}, \dots, \lambda_{m_i}^{-1}\}v_i.$$

- Stochastic approximation:

- If observed values of $f_i(x)$ at the sample points x_k^i are random variables for $k = 1, \dots, N_i$, we consider

$$e^i = \mathbb{E}_i\{G_i(f_i(x) - \hat{f}_i(x))\}$$

$G_i(\cdot)$ — a convex function such as

$$|f_i(x) - \hat{f}_i(x)|, \quad [f_i(x) - \hat{f}_i(x)]^2$$

•

$$\hat{f}_i(x) = c_i' \varphi_i(x)$$

$$e^i = \mathbb{E}_i\{G_i(f_i(x) - c_i' \varphi_i(x))\}$$

$$\begin{aligned} \frac{\partial e^i}{\partial c_i} &= -\mathbb{E}_i\{g_i(f_i(x) - c_i' \varphi_i(x)) \varphi_i(x)\} \\ &= 0 \end{aligned}$$

where $g_i(y) = \frac{d}{dy} G_i(y)$

- To find the root of

$$-\mathbb{E}_i\{g_i(f_i(x) - c'_i\varphi_i(x))\varphi_i(x)\}$$

we use the R-M algorithm:

$$c_i(k+1) = c_i(k) + \alpha_k g_i(y_k^i - c'_i(k)\varphi_i(x_k^i))\varphi_i(x_k^i)$$

$$\alpha_k > 0, \quad \alpha_k \rightarrow 0$$

$$\sum_k \alpha_k = \infty, \quad \sum_k \alpha_k^2 < \infty$$

$$\lim \mathbb{E}_i\{[c_i(k) - c_i^*]^2\} = 0$$

$$\text{Prob}\{\lim c_i(k) = c_i^*\} = 1$$

- Kernel approximation
- Use of feature functions in classification:

$$\hat{f}_i(x), \quad i = 1, \dots, M$$

Given a new pattern x^* ,

$$x^* \rightarrow \omega_i \text{ if } \hat{f}_i(x^*) = \max$$