2009年11月10日
上午 10:42

$y(x, w)$ parametric model

learned

training data are thrown away after training

another class of PR techniques : nonparametric

e.g. "nearest neighbor"

fast to train , slow at making prediction

Dual Representation

combinations of a kernel function evaluated at the training data points

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

1964 Aizerman et al.

'kernel trick' 1992 Boser et al.

extensions of existing methods scalar product $\longrightarrow$ kernel

'kernelize'

e.g. Kernel PCA (1998) , Kernel Fisher Discriminant (1999)

different forms of kernel functions :

stationary kernels $k(x, x') = k(x - x')$

invariant to translation in input space

homogeneous kernels $k(x, x') = k(\|x - x'\|)$

radial basis functions (RBF)

depend only on the magnitude of the distance

DUAL REPRESENTATIONS

2009年11月10日
上午 10:57

Example : Linear Regression , regularized sum-of-squares error

$$J(w) = \frac{1}{2} \sum_{n=1}^{N} \left\{ w^T \phi(x_n) - t_n \right\}^2 + \frac{\lambda}{2} w^T w$$

$\lambda \geq 0$

Solution for w takes the form

$$w = -\frac{1}{\lambda} \sum_{n=1}^{N} \left\{ w^T \phi(x_n) - t_n \right\} \phi(x_n)$$

$$= \sum_{n=1}^{N} a_n \phi(x_n) = \Phi^T a$$

$a_n = -\frac{1}{\lambda} \left\{ w^T \phi(x_n) - t_n \right\}$ $\Phi : \begin{bmatrix} \vdots \\ \phi(x_n)^T \\ \vdots \end{bmatrix}$ $a = (a_1, \ldots, a_N)^T$

Dual Representation

$$J(a) = \frac{1}{2} a^T \Phi \Phi^T \Phi \Phi^T a - a^T \Phi \Phi^T t + \frac{1}{2} t^T t + \frac{\lambda}{2} a \Phi \Phi^T a$$

$t = (t_1, \ldots, t_N)^T$

Gram matrix $K = \Phi \Phi^T$ symmetric

$$K_{nm} = \phi(x_n)^T \phi(x_m) = k(x_n, x_m)$$

$$J(a) = \frac{1}{2} a^T K K a - a K t + \frac{1}{2} t^T t + \frac{\lambda}{2} a K a$$

$\nabla J(a) = 0$

$a = (K + \lambda I_N)^{-1} t$ coefficients

$$y(x) = w^T \phi(x) = a^T \Phi \phi(x) = \mathbf{k}(x)^T (K + \lambda I_N)^{-1} t$$

* avoid the explicit computation of feature mapping $\phi(x)$

$\mathbf{k}(x) = \begin{pmatrix} \vdots \\ k(x_n, x) \\ \vdots \end{pmatrix}$ vector

CONSTRUCTING KERNELS

2009年11月25日
下午 08:16

valid kernel $\quad k(\mathbf{x},\mathbf{x}') = \phi(\mathbf{x})^T\phi(\mathbf{x}')$
$= \sum_{i=1}^{M} \varphi_i(\mathbf{x})\, \varphi_i(\mathbf{x}')$

$k(\mathbf{x},\mathbf{z}) = (\mathbf{x}^T\mathbf{z})^2$ a valid kernel?

2-dimensional case $\mathbf{x} = (x_1, x_2)^T$

$$k(\mathbf{x},\mathbf{z}) = (\mathbf{x}^T\mathbf{z})^2 = (x_1 z_1 + x_2 z_2)^2$$
$$= x_1^2 z_1^2 + 2x_1 z_1 x_2 z_2 + x_2^2 z_2^2$$
$$= (x_1^2, \sqrt{2}x_1x_2, x_2^2)(z_1^2, \sqrt{2}z_1z_2, z_2^2)^T$$
$$= \phi(\mathbf{x})^T\phi(\mathbf{z})$$

so $\phi(\mathbf{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)^T$

The necessary and sufficient condition for $k(\mathbf{x},\mathbf{x}')$ to be a valid kernel : Gram matrix $K = [k(\mathbf{x}_n,\mathbf{x}_m)]_{nm}$ should be positive semidefinite for all possible choices of the set $\{\mathbf{x}_n\}$

$K_{ij} = k(\mathbf{x}_i,\mathbf{x}_j)$ $\quad \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$

K is symmetric

Let $K = V\Lambda V^T$ $\quad$ V is an orthogonal matrix $\quad \mathbf{v}_t$

Λ contains the eigenvalues $\quad \lambda_t$

assume λ_t are nonnegative.

consider the feature mapping $\phi: \mathbf{x}_i \mapsto (\sqrt{\lambda_t}\, v_{ti})_{t=1}^{n} \in \mathbb{R}^n$
$i = 1, \dots, n$

$\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j) = \sum_{t=1}^{n} \lambda_t v_{ti} v_{tj} = (V\Lambda V^T)_{ij} = K_{ij} = k(\mathbf{x}_i,\mathbf{x}_j)$

therefore we find a feature mapping ϕ for the kernel

---

requirement of λ being nonnegative is necessary: if $\lambda_s < 0$, eigenvector $\mathbf{v}_s$

we have $\mathbf{z} = \sum_{i=1}^{n} v_{si}\, \phi(\mathbf{x}_i) = \sqrt{\Lambda}\, V^T \mathbf{v}_s \Rightarrow \|\mathbf{z}\|^2 = \mathbf{v}_s^T V\Lambda V^T \mathbf{v}_s = \mathbf{v}_s^T K \mathbf{v}_s = \lambda_s < 0$

2009年11月26日
上午 02:11

P. 296 $\quad$ assume k_1 is a valid kernel function

$$k(\mathbf{x},\mathbf{x}') = f(\mathbf{x})\, k_1(\mathbf{x},\mathbf{x}')\, f(\mathbf{x}') \quad (6.14)$$
$$k(\mathbf{x},\mathbf{x}') = \exp(k_1(\mathbf{x},\mathbf{x}')) \quad (6.16)$$

Proof of (6.14): for all $\alpha_i, \alpha_j \in \mathbb{R}$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j\, k(\mathbf{x}_i,\mathbf{x}_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j f(\mathbf{x}_i)\, k_1(\mathbf{x}_i,\mathbf{x}_j)\, f(\mathbf{x}_j)$$
$$= \sum_{i=1}^{n} \alpha_i f(\mathbf{x}_i)\phi^T(\mathbf{x}_i) \sum_{j=1}^{n} \alpha_j \phi(\mathbf{x}_j) f(\mathbf{x}_j)$$
$$= \Psi^T\Psi \geq 0 \quad \text{positive semidefinite}$$
$$\Psi = \sum_{i=1}^{n} \alpha_i f(\mathbf{x}_i)\,\phi(\mathbf{x}_i)$$

other popular kernels :

polynomial kernels $\quad k(\mathbf{x},\mathbf{x}') = (\mathbf{x}^T\mathbf{x}' + c)^M$

Gaussian kernels $\quad k(\mathbf{x},\mathbf{x}') = \exp(-\|\mathbf{x}-\mathbf{x}'\|^2/2\sigma^2)$

$$k(\mathbf{x},\mathbf{x}') = \exp(-\|\mathbf{x}-\mathbf{x}'\|^2/2\sigma^2)$$
$$= \exp(-\mathbf{x}^T\mathbf{x}/2\sigma^2)\exp(\mathbf{x}^T\mathbf{x}'/\sigma^2)\exp(-(\mathbf{x}')^T\mathbf{x}'/2\sigma^2)$$

so by (6.14) (6.16) $\Rightarrow$ valid kernels

another property of kernels

$$k(\mathbf{x},\mathbf{x}')^2 = (\phi(\mathbf{x})^T\phi(\mathbf{x}'))^2 \leq \|\phi(\mathbf{x})\|^2\|\phi(\mathbf{x}')\|^2$$
$$= k(\mathbf{x},\mathbf{x})\, k(\mathbf{x}',\mathbf{x}')$$

Other types of kernels

generalized Gaussian kernels:

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

$K(x,x')$ a nonlinear kernel

set kernels:

A_1, A_2 are two subsets $A_1, A_2 \subseteq A$

$$k(A_1,A_2) = 2^{|A_1 \cap A_2|}$$ is a valid kernel

encoding: $\phi(A_1) = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix}$ $2^{|A|}$ bits that index the subsets of A

$|A_1|$ is the number of elements

$2^{|A_1|}$ is the number of subsets

Generative + Discriminative

① probability kernels: $k(x,x') = p(x)\,p(x')$

or $k(x,x') = \sum_i p(x|i)\,p(x'|i)\,p(i)$

or $k(x,x') = \int p(x|Z)\,p(x'|Z)\,p(Z)\,dZ$

↑ latent variable

② Fisher kernel

Fisher score $g(\theta,x) = \nabla_\theta \ln p(x|\theta)$

Fisher kernel $k(x,x') = g(\theta,x)^T F^{-1} g(\theta,x')$

$$F = E_x\left[g(\theta,x)\,g(\theta,x)^T\right]$$ is the Frsher information

continued. Fisher kernel

consider a Gaussian distribution $\mathcal{N}(x|\mu,S)$

$$g(\mu,x) = \nabla_\mu \ln \mathcal{N}(x|\mu S) = S^{-1}(x-\mu)$$

$$F = E_x\left[g(\mu,x)\,g(\mu,x)^T\right] = S^{-1}E_x\left[(x-\mu)(x-\mu)^T\right]S^{-T} = S^{-1}SS^{-T} = S^{-1}$$

$$k(x,x') = g(\theta,x)^T F^{-1} g(\theta,x) = (x-\mu)^T S^{-T}(S^{-1})^{-1}S^{-1}(x'-\mu) = (x-\mu)^T S^{-1}(x'-\mu)$$ (Mahalanobis)

Radial Basis Function

$$f(x) = \sum_{n=1}^{N} w_n\, h(\|x-x_n\|)$$ centered on each data point

Derivation:

consider noisy input, the error function is defined by

$$E = \frac{1}{2}\sum_{n=1}^{N}\int \{y(x_n+\xi)-t_n\}^2\,\nu(\xi)\,d\xi$$

ν is isotropic

to find $y(x)$ that minimizes E

we need to apply Calculus of variations

perturb $y(x) \Rightarrow y(x)+\epsilon\eta(x)$

$$E(y+\epsilon\eta) = \frac{1}{2}\sum_{n=1}^{N}\int\{y(x_n+\xi)+\epsilon\eta(x+\xi)-t_n\}^2\,\nu(\xi)\,d\xi$$

$$= E(y) + \underset{\epsilon\to 0}{\epsilon}\underbrace{\sum_{n=1}^{N}\int\{y(x_n+\xi)-t_n\}\,\nu(\xi)\,\eta(x+\xi)\,d\xi}_{=0} + O(\epsilon^2)$$

At the optimum, E should be stationary for arbitrary η

2009年11月26日
上午 04:28

choose $\eta(x) = \delta(x - z)$

$$\sum_{n=1}^{N} \int \{ y(x_n + \xi) - t_n \} \delta(x_n + \xi - z) \nu(\xi) \, d\xi$$

$$= \sum_{n=1}^{N} \{ y(z) - t_n \} \nu(z - x_n) = 0$$

$x_n + \xi - z = 0$

$\xi = z - x_n$

$$y(z) = \frac{t_n \nu(z - x_n)}{\sum_{n=1}^{N} \nu(z - x_n)}$$

$\Rightarrow$

$$y(x) = \sum_{n=1}^{N} t_n h(x - x_n)$$

$$h(x - x_n) = \frac{\nu(x - x_n)}{\sum_{n=1}^{N} \nu(x - x_n)}$$

Nadaraya – Watson model

$\{x_n, t_n\}$ $\quad p(x, t) = \frac{1}{N} \sum_{n=1}^{N} f(x - x_n, t - t_n)$

f: component density function

$$y(x) = E[t \mid x] = \int_{-\infty}^{\infty} t \, p(t \mid x) \, dt$$

$$= \frac{\int t \, p(x, t) \, dt}{\int p(x, t) \, dt}$$

from conditional to joint

$$= \frac{\sum_n \int t f(x - x_n, t - t_n) \, dt}{\sum_m f(x - x_m, t - t_m) \, dt}$$

2009年11月26日
上午 04:43

assume $\int_{-\infty}^{\infty} f(x, t) \, t \, dt = 0$ zero mean

$$\int t f(x - x_n, t - t_n) \, dt$$

$$= \underbrace{\int (t - t_n) f(x - x_n, t - t_n) \, dt}_{= 0} + \int t_n f(x - x_n, t - t_n) \, dt$$

so after "change of variable"

$$y(x) = \frac{\sum_n g(x - x_n) t_n}{\sum_m g(x - x_m)}$$

$$= \sum_n k(x, x_n) t_n$$

$$g(x) = \int_{-\infty}^{\infty} f(x, t) \, dt$$

$$p(t \mid x) = \frac{p(t, x)}{\int p(t, x) \, dt} = \frac{\sum_n f(x - x_n, t - t_n)}{\sum_m \int f(x - x_m, t - t_m) \, dt}$$

$f(x, t)$ zero-mean isotropic Gaussian