Assignment

2009年10月7日
下午 07:50

Show that

① $\int (x-\mu)^T \Sigma^{-1} (x-\mu)\, \mathcal{N}(x; m, S)\, dx$

$= (\mu-m)^T \Sigma^{-1} (\mu-m) + Tr\{\Sigma^{-1} S\}$

② $\int (\overline{W}x-\mu)^T \Sigma^{-1} (\overline{W}x-\mu)\, \mathcal{N}(x; m, S)\, dx$

$= (\mu-\overline{W}m)^T \Sigma^{-1} (\mu-\overline{W}m) + Tr\{\overline{W}^T \Sigma^{-1} \overline{W} S\}$

Expected Loss = $(bias)^2$ + variance + noise

$(bias)^2 = \int \{E_D[y(x;D)] - h(x)\}^2 p(x)\, dx$

$variance = \int E_D[\{y(x;D) - E_D[y(x;D)]\}^2] p(x)\, dx$

$noise = \iint \{h(x)-t\}^2 p(x,t)\, dx\, dt$

flexible models : low bias, high variance

rigid models : high bias, low variance

Run Matlab

Fig 3.5 large λ, large bias

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

2009年10月7日
下午 08:00

In mathematics you don't understand things. You just get used to them.

— John von Neumann

Bayesian Linear Regression

model complexity :
number of basis functions
maximizing the likelihood function ⟶ overfitting
add regularization terms

Bayesian :
prior probability distribution over w (model params.)

If the likelihood $p(t|w)$ is exponential of a quadratic function of w, we may choose a conjugate prior in the form of Gaussian distribution

$$p(w) = \mathcal{N}(w | m_0, S_0)$$

In Bayesian probability theory, a class of prior probability distribution $p(\theta)$ is said to be conjugate to a class of likelihood functions $p(x|\theta)$ if the resulting posterior distribution $p(\theta|x)$ are in the same family as $p(\theta)$

Why conjugate prior ?

① For mathematical convenience.
Closed-form expression for the posterior

② Intuitive meaning : the likelihood updates the prior

likelihood : $p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} \mathcal{N}(t_n | \mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1})$

(3.10) $= \mathcal{N}(\mathbf{t} | \mathbf{\Phi}\mathbf{w}, \beta^{-1}\mathbf{I})$

According to the marginal and conditional Gaussians listed on PRML p.93

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, \Lambda^{-1}) \qquad p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b}, \mathbf{L}^{-1})$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x} | \Sigma\{\mathbf{A}^T\mathbf{L}(\mathbf{y}-\mathbf{b}) + \Lambda\mu\}, \Sigma)$$

$$\Sigma = (\Lambda + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}$$

the posterior is also a Gaussian

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \mathbf{S}_N)$$

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^T\mathbf{t})$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^T\mathbf{\Phi}$$

Some properties

① if $\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$, with $\alpha \to 0$, $\mathbf{m}_N \to \mathbf{w}_{ML}$

$\alpha \to 0$ means low precision, $\mathbf{S}_0^{-1} \to 0$

$$\mathbf{S}_N^{-1} = \beta\mathbf{\Phi}^T\mathbf{\Phi}$$

$$\mathbf{m}_N = \beta^{-1}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\beta\mathbf{\Phi}^T\mathbf{t} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{t}$$

maximum likelihood solution

② if $N = 0$ posterior → prior

③ sequential update
If the data points arrive sequentially, then the posterior distribution at any stage act as the prior distribution for the subsequent data point.

consider a simpler form of prior

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

i.e. $\mathbf{m}_0 = \mathbf{0}$, $\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$
zero mean diagonal covariance (isotropic)

the posterior Gaussian has

$$\mathbf{m}_N = \beta\mathbf{S}_N\mathbf{\Phi}^T\mathbf{t}$$

$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi}$$

2009年10月7日
下午 08:41

Therefore, the log posterior is

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\phi(x_n)\}^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + \text{const.}$$

Maximizing Gaussian posterior is equivalent to "minimizing sum-of-squares error function with a quadratic regularization term."

$$\lambda = \frac{\alpha}{\beta}$$

Fig 3.7 run Matlab

$y(x, \mathbf{w}) = w_0 + w_1 x$

synthetic data $f(x, a) = a_0 + a_1 x$ $\quad a_0 = -0.3,\ a_1 = 0.5$

$x \sim U(x|-1, 1)$

adding noise $\mathcal{N}(0, 0.2^2)$

goal: recover a_0, a_1

assume the noise variance (0.2^2) is known

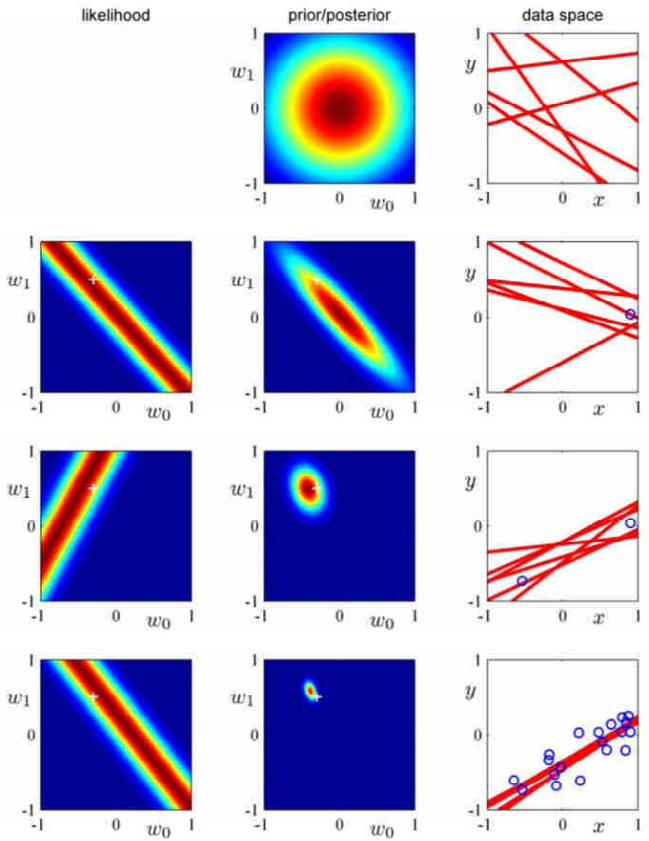
set the hyperparameter α of the prior as 2.0

Apply Bayesian inference

In the limit of an infinite number of data points the posterior distribution would become a delta function centered on the true parameter values

Fig 3.7

2009年10月6日
上午 10:28



Predictive Distribution

2009年10月7日
下午 08:53

Usually we are more interested in predicting t for new input x.

Instead of considering w, we may evaluate the predictive distribution by marginalizing out w through integration

$$p(t \mid \mathbf{t}, \alpha, \beta) = \int p(t \mid w, \beta)\, p(w \mid \mathbf{t}, \alpha, \beta)\, dw$$

$$\mathcal{N}(t \mid y(x, w), \beta^{-1}) \qquad \mathcal{N}(w \mid m_N, S_N)$$

$$\mathcal{N}(t \mid w^T \phi(x), \beta^{-1})$$

Use the equations on PRML p. 93 again, we get

$$p(t \mid \mathbf{X}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t \mid m_N^T \phi(x), \sigma_N^2(x))$$

$$\sigma_N^2 = \frac{1}{\beta} + \phi(x)^T S_N \phi(x)$$

noise — uncertainty associated with w

$\hat{t} = m_N^T \phi(x)$ can be our prediction of t (predictive mean)

Fig 3.8 (Run Matlab)

2009年10月7日
下午 09:10

When the number N of data points goes to ∞,

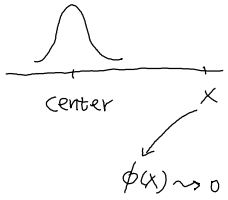
$$\sigma_N^2 \to \frac{1}{\beta}$$

we become very certain about the prediction

Problem:

If we use localized basis functions such as Gaussians, then in regions away from the basis function centers, the contribution from the second term in the predictive variance σ_N^2 will go to zero.

x is a new input point



The model becomes very confident in its predictions when exprapolating outside the region occupied by the basis functions, which is generally an undesirable behavior.

The problem is due to the pre-defined basis functions in the model.

We will see later how to avoid the problem by using a different Bayesian approach known as a Gaussian Process.

Equivalent Kernel

2009年10月7日
下午 09:27

The posterior mean solution m_N for the linear basis function model has an interpretation related to kernel methods.

the predictive mean can be written as

$$y(x, m_N) = m_N^T \phi(x) = \beta \phi(x)^T S_N \Phi^T t$$
$$= \sum_{n=1}^{N} \beta \phi(x)^T S_N \phi(x_n) t_n$$

so

$$y(x, m_N) = \sum_{n=1}^{N} k(x, x_n) t_n$$

where

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

is known as the equivalent kernel.

The mean of the predictive distribution at x, given by $y(x, m_N)$, is obtained by forming a weighted combination of the target values in which data points close to x are given higher weight than points further removed from x.

2009年10月7日
下午 09:58

Further insight:

$$\text{cov}[y(x), y(x')] = \text{cov}[\phi(x)^T w, w^T \phi(x')]$$
$$= \phi(x)^T \text{cov}[w, w^T] \phi(x')$$
$$= \phi(x)^T S_N \phi(x')$$
$$= \beta^{-1} k(x, x')$$

$\text{cov}[w, w^T]$
$= E[(w - m_N)(w - m_N)^T]$
$= S_N$
since $p(w|t) = N(w|m_N, S_N)$

Hint of the idea of Gaussian Processes:
Instead of using a set of basis functions, we may a localized kernel directly on the data points.

Other properties of equivalent kernels

$$\sum_{n=1}^{N} k(x, x_n) = 1$$

and

$$k(x, z) = \psi(x)^T \psi(z) \qquad \text{(inner product)}$$

where

$$\psi(x) = \beta^{1/2} S_N^{1/2} \phi(x)$$

(A draft of proof on the next page.)

2009年10月7日
下午 10:17

PRML Exercise 3.14

Suppose that $\phi_0(x) = 1$ and $\phi_j(x)$ linearly independent

$S_N^{-1} = \alpha I + \beta \Phi^T \Phi$ for $\alpha = 0$

we may find a new basis set $\psi_j(x)$
such that $k(x, x') = \psi(x)^T \psi(x')$
and

$$\sum_{n=1}^{N} \psi_j(x_n)\, \psi_k(x_n) = I_{jk} \qquad I_{jk} = \begin{cases} 1, & \text{if } j=k \\ 0, & \text{otherwise} \end{cases}$$

$$\sum_{n=1}^{N} \psi_j(x_n)\, \psi_0(x_n) = \sum_{n=1}^{N} \psi_j(x_n) = I_{j0}$$

$$\Rightarrow \quad \sum_{n=1}^{N} k(x, x_n) = \sum_{n=1}^{N} \psi(x)^T \psi(x_n) = \sum_{n=1}^{N} \sum_{i=0}^{M} \psi_i(x)\, \psi_i(x_n)$$

$$= \sum_{i=0}^{M} \psi_i(x) \sum_{n=1}^{N} \psi_i(x_n)$$

$$= \sum_{i=0}^{M} \psi_i(x)\, I_{i0} = \psi_0(x) = 1$$

The Evidence Approximation

2009年10月7日
下午 10:27

hyperparameters α, β

α: for w prior; β: noise, for likelihood

We may also introduce prior distributions on α, β, and marginalize them.
Although we can integrate analytically over either w or over the hyperparameters, the complete marginalization over all of these variables is analytically intractable.

So, how to analyze α and β ?

Evidence Approximation: we set the hyperparameters to specific values determined by maximizing the marginal likelihood function obtained by first integrating over the parameter w.
(aka Emprical Bayes)

$$p(t \mid \mathbf{t}) = \iiint p(t \mid w, \beta)\, p(w \mid \mathbf{t}, \alpha, \beta)\, p(\alpha, \beta \mid \mathbf{t})\, dw\, d\alpha\, d\beta$$

$$p(t \mid w, \beta) = \mathcal{N}\left(t \mid y(x, w), \beta^{-1}\right) \qquad (3.8)$$

$$p(w \mid \mathbf{t}, \alpha, \beta) = \mathcal{N}(w \mid m_N, S_N) \qquad (3.49)$$

$$m_N = \beta S_N \Phi^T \mathbf{t} \qquad (3.53)$$

$$S_N^{-1} = \alpha I + \beta \Phi^T \Phi \qquad (3.54)$$

2009年10月8日
上午 09:30

If the posterior distribution $p(\alpha, \beta \mid \mathbf{t})$ is sharply peaked around values $\hat{\alpha}$, $\hat{\beta}$, then

$$p(t \mid \mathbf{t}) \simeq p(t \mid \mathbf{t}, \hat{\alpha}, \hat{\beta})$$

(delta function)

$$= \int p(t \mid \mathbf{w}, \hat{\beta})\, p(\mathbf{w} \mid \mathbf{t}, \hat{\alpha}, \hat{\beta})\, d\mathbf{w}$$

From Bayes' theorem, the posterior for α, β is

$$p(\alpha, \beta \mid \mathbf{t}) \propto p(\mathbf{t} \mid \alpha, \beta)\, p(\alpha, \beta)$$

If the prior $p(\alpha, \beta)$ is relatively flat, then $\hat{\alpha}$ and $\hat{\beta}$ are obtained by maximizing the marginal likelihood function $p(\mathbf{t} \mid \alpha, \beta)$

Evaluate the evidence function

marginal likelihood

$$p(\mathbf{t} \mid \alpha, \beta) = \int p(\mathbf{t} \mid \mathbf{w}, \beta)\, p(\mathbf{w} \mid \alpha)\, d\mathbf{w}$$

(3.11) $$\ln p(\mathbf{t} \mid \mathbf{w}, \beta) = \sum_{n=1}^{N} \ln \mathcal{N}(t_n \mid \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

$$= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w})$$

(3.12) $$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{ t_n - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n) \}^2$$

(3.52) $$p(\mathbf{w} \mid \alpha) = \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha^{-1} \mathbf{I})$$

2009年10月8日
上午 09:48

$$p(\mathbf{t} \mid \alpha, \beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \int \exp\{-E(\mathbf{w})\}\, d\mathbf{w}$$

(M is the dim. of $\mathbf{w}$)

$$E(\mathbf{w}) = \beta E_D(\mathbf{w}) + \alpha E_W(\mathbf{w})$$

$$= \frac{\beta}{2} \| \mathbf{t} - \boldsymbol{\Phi} \mathbf{w} \|^2 + \frac{\alpha}{2} \mathbf{w}^T \mathbf{w}$$

(regularized sum-of-squares error)

See next page for the derivation

$$E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2} (\mathbf{w} - \mathbf{m}_N)^T \mathbf{A} (\mathbf{w} - \mathbf{m}_N)$$

$$\begin{cases} \mathbf{A} = \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi} \\ E(\mathbf{m}_N) = \frac{\beta}{2} \| \mathbf{t} - \boldsymbol{\Phi} \mathbf{m}_N \|^2 + \frac{\alpha}{2} \mathbf{m}_N^T \mathbf{m}_N \end{cases}$$

Observation:

$$\mathbf{A} = \nabla \nabla E(\mathbf{w})$$

$$\mathbf{m}_N = \beta \mathbf{A}^{-1} \boldsymbol{\Phi}^T \mathbf{t}$$

$$\mathbf{A} = \mathbf{S}_N^{-1}$$

$\mathbf{m}_N = \beta \mathbf{S}_N \boldsymbol{\Phi}^T \mathbf{t}$ (3.53)

$\mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi}$ (3.54)

Back to $$p(\mathbf{t} \mid \alpha, \beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \int \exp\{-E(\mathbf{w})\}\, d\mathbf{w}$$

$$\int \exp\{-E(\mathbf{w})\}\, d\mathbf{w}$$

$$= \exp\{-E(\mathbf{m}_N)\} \int \exp\left\{-\frac{1}{2} (\mathbf{w} - \mathbf{m}_N)^T \mathbf{A} (\mathbf{w} - \mathbf{m}_N)\right\} d\mathbf{w}$$

$$= \exp\{-E(\mathbf{m}_N)\} (2\pi)^{M/2} |\mathbf{A}|^{-1/2}$$

Gaussian normalization

previous page in detail

2009年10月8日
上午 10:11

$$E(w) = \frac{\beta}{2}\|t - \Phi w\|^2 + \frac{\alpha}{2} w^T w$$

$$= \frac{\beta}{2} t^T t + \frac{\beta}{2} w \Phi^T \Phi w - \frac{\beta}{2} t^T \Phi w - \frac{\beta}{2} w^T \Phi^T t + \frac{\alpha}{2} w^T w$$

$$= \frac{\beta}{2} t^T t - \beta t^T \Phi w + \frac{1}{2} w^T A w$$

$$= \frac{\beta}{2} t^T t - \beta t^T \Phi A^{-1} A w + \frac{1}{2} w^T A w + \frac{1}{2} m_N^T A m_N - \frac{1}{2} m_N^T A m_N$$

$$= \frac{\beta}{2} t^T t - \frac{1}{2} m_N^T A m_N + \frac{1}{2} (w - m_N)^T A (w - m_N)$$

$$= \frac{\beta}{2}\|t - \Phi m_N\|^2 + \frac{\alpha}{2} m_N^T m_N + \frac{1}{2} (w - m_N)^T A (w - m_N)$$

‖ use $m_N = \beta A^{-1} \Phi^T t$

$$\nabla\nabla E(w) = \nabla\nabla \left(\frac{\beta}{2} (t^T t + w^T \Phi^T \Phi w - t^T \Phi w - (\Phi w)^T t) + \frac{\alpha}{2} w^T w \right)$$

$$= \nabla \left(\frac{\beta}{2} (2\Phi^T \Phi w - t^T \Phi - t^T \Phi) + \alpha I w \right)$$

$$= \beta \Phi^T \Phi + \alpha I \equiv A = S_N^{-1}$$

$$m_N = \beta A^{-1} \Phi^T t$$

2009年10月8日
上午 09:59

log marginal likelihood

$$\ln p(t|\alpha, \beta) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - E(m_N) - \frac{1}{2} \ln |A| - \frac{N}{2} \ln (2\pi)$$

↳ evidence function

given α and β, the evidence function can show that how well the model explains the observed data t

Maximizing the evidence function

① maximizing $p(t|\alpha, \beta)$ w.r.t. α

defining the eigenvector equation $(\beta \Phi^T \Phi) u_i = \lambda_i u_i$

since $A = \alpha I + \beta \Phi^T \Phi$

A has eigenvalues $\alpha + \lambda_i$

$$\frac{d}{d\alpha} \ln |A| = \frac{d}{d\alpha} \ln \prod_i (\lambda_i + \alpha) = \frac{d}{d\alpha} \sum_i \ln (\lambda_i + \alpha) = \sum_i \frac{1}{\lambda_i + \alpha}$$

$$0 = \frac{M}{2\alpha} - \frac{1}{2} m_N^T m_N - \frac{1}{2} \sum_i \frac{1}{\lambda_i + \alpha}$$

$$\alpha m_N^T m_N = M - \alpha \sum_i \frac{1}{\lambda_i + \alpha} = \gamma$$

$$\gamma = \sum_i \frac{\lambda_i}{\alpha + \lambda_i}$$

2009年10月8日
上午 10:33

iterative

$$\alpha \Rightarrow m_N$$

$$\gamma = \sum_i \frac{\lambda_i}{\alpha + \lambda_i}$$

$$\alpha = \frac{\gamma}{m_N^T m_N}$$

② maximizing $p(t|\alpha,\beta)$ w.r.t. β

$(\beta \Phi^T \Phi) u_i = \lambda_i u_i$ the eigenvalues λ_i are proportional to β, i.e. $\lambda_i = \beta \eta_i$

$$\frac{d\lambda_i}{d\beta} = \frac{d\beta\eta_i}{d\beta} = \eta_i = \frac{\lambda_i}{\beta}$$

$$\frac{d}{d\beta} \ln|A| = \frac{d}{d\beta} \sum_i \ln(\lambda_i + \alpha)$$

$$= \frac{1}{\beta} \sum_i \frac{\lambda_i}{\lambda_i + \alpha} = \frac{\gamma}{\beta}$$

$$0 = \frac{N}{2\beta} - \frac{1}{2} \sum_{n=1}^{N} \{t_n - m_N^T \phi(x_n)\}^2 - \frac{\gamma}{2\beta}$$

$$\beta^{-1} = \frac{1}{N-\gamma} \sum_{n=1}^{N} \{t_n - m_N^T \phi(x_n)\}^2$$

iteratively update m_N, γ, β

2009年10月8日
上午 10:50

Effective number of parameters (γ)

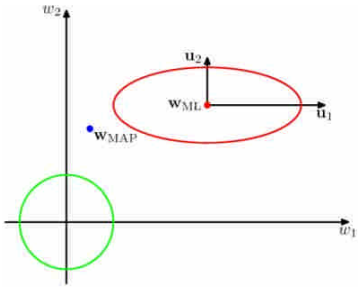
$$0 < \frac{\lambda_i}{\lambda_i + \alpha} < 1 \qquad 0 \le \gamma \le M$$

$\lambda_i \gg \alpha$ $w_i \to$ maximum likelihood

$$\gamma = \frac{\lambda_i}{\lambda_i + \alpha} \to 1$$ well determined

$\lambda_i \ll \alpha$ $w_i \to 0$

$$\gamma = \frac{\lambda_i}{\lambda_i + \alpha} \to 0$$



If $N \gg M$, all the parameters will be well determined.
In this case $\gamma = M$, and

$$\alpha = \frac{M}{2E_W(m_N)}$$

$$\beta = \frac{M}{2E_D(m_N)}$$

} easy-to-compute approximation

2009年10月8日
上午 10:44

Limitations of Fixed Basis Functions

curse of dimensionality
the number of basis functions needs to grow rapidly, often exponentially, with the dimensionality D of the input space

Good news:

① Input: $\{x_n\}$ on a manifold

intrinsic dimensionality is small

we can use localized basis functions

② Target: high dependence on a small number of possible directions within the data manifold