2009年12月5日
下午 03:49

RELEVANCE VECTOR MACHINES RVMs

Matrix Identities

(C.7) $(A+BD^{-1}C)^{-1} = A^{-1} - A^{-1}B(D+CA^{-1}B)^{-1}CA^{-1}$

(C.14) $|I_N + AB^T| = |I_M + A^TB|$

(C.15) $|I_N + ab^T| = 1 + a^Tb$

Gaussian Identities

(2.113) $p(x) = \mathcal{N}(x \mid \mu, \Lambda^{-1})$

(2.114) $p(y|x) = \mathcal{N}(y \mid Ax+b, L^{-1})$

(2.115) $p(y) = \mathcal{N}(y \mid A\mu + b, L^{-1} + A\Lambda^{-1}A^T)$

(2.116) $p(x|y) = \mathcal{N}(x \mid \Sigma\{A^TL(y-b) + \Lambda\mu\}, \Sigma)$

$$\Sigma = (\Lambda + A^TLA)^{-1}$$

Limitations of SVMs

1. posterior probilities ?
2. two-class $\Rightarrow$ multi-class
3. C, ν, ϵ parameter selection by cross-validation
4. positive definite kernels

RVM for regression

t: real-valued target value

x: input vector

Gaussian Noise: $p(t|x, w, \beta) = \mathcal{N}(t \mid \underbrace{y(x)}_{\text{mean}}, \underbrace{\beta^{-1}}_{\text{noise precision}})$

$$y(x) = \sum_{i=1}^{M} w_i \phi_i(x) = w^T\phi(x)$$

2009年12月5日
下午 04:02

for RVMs, we assume the following model

$$y(x) = \sum_{n=1}^{N} w_n \underbrace{k(x, x_n)}_{} + b$$

$\Downarrow$

as $\phi(x)$ in a linear model

no restriction to p.d kernels

Use the matrix notation and assume i.i.d.

$$X = \begin{bmatrix} \vdots \\ x_n^T \\ \vdots \end{bmatrix} \Big\} N \qquad t = \begin{bmatrix} \vdots \\ t_n \\ \vdots \end{bmatrix} \Big\} N$$

likelihood:

$$p(t \mid X, w, \beta) = \prod_{n=1}^{N} p(t_n \mid x_n, w, \beta)$$

Now introduce the prior on w

Key: each weight parameter w_i has a separate hyperparameter α_i

$$p(w|\alpha) = \prod_{i=1}^{M} \mathcal{N}(w_i \mid 0, \alpha_i^{-1})$$

if $\alpha_i \to \infty$ $\Rightarrow$ high precision, zero variance

$\Rightarrow$ w_i centered at the mean $= 0$

$\Rightarrow$ sparse model

2009年12月5日
下午 04:12

from the likelihood and the prior we can write the posterior as

$$p(\mathbf{w} \mid \mathbf{t}, \mathbf{X}, \alpha, \beta) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}, \Sigma)$$

(2.116)

where
$$\mathbf{m} = \Sigma \{ \Phi^T \beta \mathbf{t} \} \qquad \Phi_{ni} = \phi_i(x_n)$$
$$\Sigma = (A + \beta \Phi^T \Phi)^{-1} \qquad A = \mathrm{diag}(\alpha_i)$$

this is obtained by applying (2.113)+(2.114) → (2.116) to

as (2.113) ⇒ $p(\mathbf{w} \mid \alpha) = \prod_{i=1}^{M} \mathcal{N}(w_i \mid 0, \alpha_i^{-1})$

as (2.114) ⇒ $p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(\underbrace{\mathbf{w}^T \phi(x_n)}_{= y(x_n)}, \beta^{-1})$

(Note that $\Phi = K$ for RVM)

The optimal weights $\mathbf{w}^*$ is given by the mean of the posterior

$$\mathbf{w}^* = \mathbf{m} = \beta \Sigma \Phi^T \mathbf{t}$$

The next step is to decide the hyperparameters α, β.

We use "evidence approximation".

EA ① Compute the marginal likelihood

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

$p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \beta)$ and $p(\mathbf{w}, \alpha)$ are Gaussians

2009年12月5日
下午 04:23

Again, we don't need to do the integration explicitly for marginalization.
Just by observing and applying (2.113)+(2.114)→(2.115) we get

$$p(\mathbf{t} \mid \mathbf{X}, \alpha, \beta) = \int p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \beta)\, p(\mathbf{w} \mid \alpha)\, d\mathbf{w}$$
$$= \mathcal{N}(\mathbf{t} \mid 0, C)$$
$$C = \beta^{-1} I + \Phi A^{-1} \Phi^T \qquad A = \mathrm{diag}(\alpha_i)$$

EA ② maximize the logarithm of the marginal likelihood w.r.t. α and β

$$\ln p(\mathbf{t} \mid \mathbf{X}, \alpha, \beta) = \ln \mathcal{N}(\mathbf{t} \mid 0, C)$$
$$= -\frac{1}{2} \{ N \ln(2\pi) + \ln |C| + \mathbf{t}^T C^{-1} \mathbf{t} \}$$

It takes a couple of pages to write down the derivations. For now we write the results only, and go into the details later.

take the derivatives and set them to zero, we get

$$\alpha_i^{new} = \frac{\gamma_i}{m_i^2}$$

$$(\beta^{new})^{-1} = \frac{\| \mathbf{t} - \Phi \mathbf{m} \|^2}{N - \sum_i \gamma_i} \qquad \gamma_i = 1 - \alpha_i \Sigma_{ii}$$

$$\ln p(\mathbf{t}|X,\alpha,\beta) = -\frac{1}{2}\left\{N\ln(2\pi) + \ln|C| + \mathbf{t}^T C^{-1}\mathbf{t}\right\}$$

$$\frac{\partial}{\partial\alpha_i}\left\{\ln p(\mathbf{t}|X,\alpha,\beta) = \frac{\partial \ln|C|}{\partial\alpha_i} + \mathbf{t}^T\frac{\partial C^{-1}}{\partial\alpha_i}\mathbf{t}\right.$$

①
$$\begin{aligned}
\ln|C| &= \ln|\beta^{-1}I + \Phi A^{-1}\Phi^T| \\
&= \ln\beta^{-N}|I + \beta\Phi A^{-1}\Phi^T| \\
&= -N\ln\beta + \ln|I + \beta\Phi A^{-1}\Phi^T| \\
&= -N\ln\beta + \ln|I + \beta A^{-1}\Phi^T\Phi| \qquad (C.14) \\
&= -N\ln\beta - \ln|A| + \ln|A + \beta\Phi^T\Phi| \\
&= -N\ln\beta - \ln|A| + \ln|\Sigma^{-1}|
\end{aligned}$$

recall
w posterior
$\Sigma = (A + \beta\Phi^T\Phi)^{-1}$

$$\frac{\partial\ln|C|}{\partial\alpha_i} = -\mathrm{Tr}\left(A^{-1}\frac{\partial A}{\partial\alpha_i}\right) + \mathrm{Tr}\left(\Sigma\frac{\partial\Sigma^{-1}}{\partial\alpha_i}\right) \qquad (C.22)$$

$$= -\frac{1}{\alpha_i} + \Sigma_{ii}$$

$$\frac{\partial A}{\partial\alpha_i} = \begin{bmatrix} 0 & & 0 \\ & 1 & \\ 0 & & 0 \end{bmatrix}$$

$\frac{\partial\Sigma^{-1}}{\partial\alpha_i}$ is similar

②
$$\begin{aligned}
C^{-1} &= (\beta^{-1}I + \Phi A^{-1}\Phi^T)^{-1} \\
&= (\beta^{-1}(I + \beta\Phi A^{-1}\Phi^T))^{-1}
\end{aligned}$$

(C.7) $$= \beta(I + \beta\Phi A^{-1}\Phi^T)^{-1} = \beta\left\{I - \Phi(\beta^{-1}A + \Phi^T\Phi)^{-1}\Phi^T\right\}$$

$$= \beta\left\{I - \beta\Phi(A + \beta\Phi^T\Phi)^{-1}\Phi^T\right\} = \beta\left\{I - \beta\Phi\Sigma\Phi^T\right\}$$

$$\frac{\partial\, \mathbf{t}^T C^{-1}\mathbf{t}}{\partial\alpha_i} = \mathbf{t}^T\frac{\partial C^{-1}}{\partial\alpha_i}\mathbf{t}$$

$$\frac{\partial C^{-1}}{\partial\alpha_i} = \frac{\partial}{\partial\alpha_i}\left\{\beta(I - \beta\Phi\Sigma\Phi^T)\right\}$$

$$\begin{aligned}
&= -\beta^2\Phi\frac{\partial\Sigma}{\partial\alpha_i}\Phi^T \\
&= -\beta^2\Phi\left(-\Sigma\begin{bmatrix} 0 & & \\ & 1 & \\ & & 0\end{bmatrix}\Sigma\right)\Phi^T
\end{aligned}$$

by (C.21)
$$\frac{\partial}{\partial\alpha_i}(\Sigma) = -\Sigma\frac{\partial\Sigma^{-1}}{\partial\alpha_i}\Sigma$$
$$\Downarrow$$
$$\begin{bmatrix} 0 & & \\ & 1 & \\ & & 0\end{bmatrix}$$

$$\Rightarrow \frac{\partial\mathbf{t}^T C^{-1}\mathbf{t}}{\partial\alpha_i}$$

$$= \mathbf{t}^T\beta\Phi\Sigma\begin{bmatrix} 0 & & \\ & 1 & \\ & & 0\end{bmatrix}\Sigma\Phi^T\beta\mathbf{t}$$

$$= \mathbf{m}^T\begin{bmatrix} 0 & & \\ & 1 & \\ & & 0\end{bmatrix}\mathbf{m} = m_i^2$$

recall w posterior
$\mathbf{m} = \Sigma\Phi^T\beta\mathbf{t}$

combine the results of ① and ②

$$\frac{\partial\ln P(\mathbf{t}|X,\alpha,\beta)}{\partial\alpha_i} = \frac{1}{2\alpha_i} - \frac{1}{2}\Sigma_{ii} - \frac{1}{2}m_i^2$$

Set the derivative to zero

$$\alpha_i = \frac{1 - \alpha_i\Sigma_{ii}}{m_i^2}$$

α^* The update rule $\alpha_i^{new} = \frac{\gamma_i}{m_i^2}$ | $\gamma_i = 1 - \alpha_i^{old}\Sigma_{ii}$

2009年12月10日
下午 05:06

We also need to compute the derivatives w.r.t. β

Again $\ln p(\mathbf{t}|X,\alpha,\beta) = -\frac{1}{2}\{N\ln(2\pi)+\ln|C| + \mathbf{t}^T C^{-1}\mathbf{t}\}$

$$\frac{\partial}{\partial\beta}\ln p(\mathbf{t}|X,\alpha,\beta) = -\frac{1}{2}\left\{\frac{\partial}{\partial\beta}\ln|C| + \frac{\partial}{\partial\beta}\mathbf{t}^T C^{-1}\mathbf{t}\right\}$$

① recall $\ln|C| = -N\ln\beta - \ln|A| + \ln|\Sigma^{-1}|$

$$\frac{\partial \ln|C|}{\partial\beta} = \frac{-N}{\beta} + \frac{\partial}{\partial\beta}\ln|\Sigma^{-1}|$$

| A is irrelevant to β

Write $\frac{\partial}{\partial\beta}\ln|\Sigma^{-1}| = \mathrm{Tr}\left(\Sigma\frac{\partial\Sigma^{-1}}{\partial\beta}\right)$

| recall in posterior $\Sigma^{-1} = (A+\beta\Phi^T\Phi)$

(C.22)
$$= \mathrm{Tr}(\Sigma\Phi^T\Phi)$$
$$= \mathrm{Tr}(\Sigma\Phi^T\Phi + \beta^{-1}\Sigma A - \beta^{-1}\Sigma A)$$
$$= \mathrm{Tr}\{\Sigma(\Phi^T\Phi\beta + A)\beta^{-1} - \beta^{-1}\Sigma A\}$$
$$= \mathrm{Tr}\{(I - A\Sigma)\beta^{-1}\}$$

$$\frac{\partial\ln|C|}{\partial\beta} = -\frac{N}{\beta} + \frac{1}{\beta}\mathrm{Tr}\{I - A\Sigma\}$$

| let $\gamma_i = 1-\alpha_i\Sigma_{ii}$

$$= -\frac{N}{\beta} + \frac{1}{\beta}\sum_{i=1}^{N}\gamma_i$$

2009年12月10日
下午 07:13

② $\frac{\partial\,\mathbf{t}^T C^{-1}\mathbf{t}}{\partial\beta}$ | using previous results

$$= \frac{\partial}{\partial\beta}\{\mathbf{t}^T\beta(I - \beta\Phi\Sigma\Phi^T)\mathbf{t}\}$$
$$= \frac{\partial}{\partial\beta}\{\beta\mathbf{t}^T\mathbf{t} - \beta^2\mathbf{t}\Phi\Sigma\Phi^T\mathbf{t}\}$$
$$= \mathbf{t}^T\mathbf{t} - 2\beta\mathbf{t}\Phi\Sigma\Phi^T\mathbf{t} + \beta^2\mathbf{t}\Phi\Sigma\Phi^T\Phi\Sigma\Phi^T\mathbf{t}$$
$$= \|\mathbf{t} - \Phi\mathbf{m}\|^2$$

| $\mathbf{m} = \beta\Sigma\Phi^T\mathbf{t}$

| $\frac{\partial\Sigma}{\partial\beta} = -\Sigma\frac{\partial\Sigma^{-1}}{\partial\beta}\Sigma = -\Sigma\Phi^T\Phi\Sigma$

$$\frac{\partial}{\partial\beta}\ln p(\mathbf{t}|X,\alpha,\beta) = \frac{1}{2}\frac{N}{\beta} - \frac{1}{2}\frac{1}{\beta}\sum_{i=1}^{N}\gamma_i - \frac{1}{2}\|\mathbf{t} - \Phi\mathbf{m}\|^2 = 0$$

The update rule

$$\beta^* \qquad \frac{1}{\beta^{new}} = \frac{\|\mathbf{t} - \Phi\mathbf{m}\|^2}{N - \sum_{i=1}^{N}\gamma_i} \qquad \gamma_i = 1 - \alpha_i\Sigma_{ii}$$

2009年12月15日
上午 10:44

Predictive Distribution

$$p(t \mid \mathbf{x}, \mathbb{X}, \mathbf{t}, \alpha^*, \beta^*)$$

$$= \int \underbrace{p(t \mid \mathbf{x}, \mathbf{w}, \beta^*)}_{\mathcal{N}(t \mid y(\mathbf{x}), \beta^{-1})} \underbrace{p(\mathbf{w} \mid \mathbb{X}, \mathbf{t}, \alpha^*, \beta^*)}_{\mathcal{N}(\mathbf{w} \mid \mathbf{m}, \Sigma)} \, d\mathbf{w}$$

$$= \mathcal{N}\left(t \mid \underset{\uparrow \atop (\alpha^*, \beta^*)}{\mathbf{m}^T \phi(\mathbf{x})}, \sigma^2(\mathbf{x})\right)$$

$$\sigma^2(\mathbf{x}) = (\beta^*)^{-1} + \phi(\mathbf{x})^T \Sigma \phi(\mathbf{x})$$

$$\mathbf{m} = \beta^* \Sigma \Phi^T \mathbf{t}$$