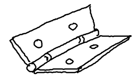
Relation to Logistic Regression

2009年12月4日
下午 08:07

The objective function of SVM can be written in the form

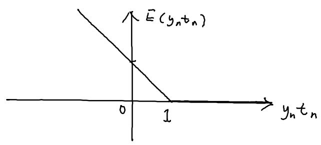
$$\sum_{n=1}^{N} E_{SV}(y_n t_n) + \lambda \| w \|^2$$

$E_{SV}(\cdot)$ is the hinge error function



$$E_{SV}(y_n t_n) = [1 - y_n t_n]_+$$

or equivalently, $E_{SV}(y_n t_n) = \begin{cases} 0, & \text{if } y_n t_n \geq 1, \\ 1 - y_n t_n, & \text{otherwise}. \end{cases}$



hinge loss

if $y_n t_n \geq 1$
the is no penalty
otherwise the penalty
increases linearly

consider the sigmoid function $\sigma(y) = \frac{1}{1+e^{-y}}$ for logistic regression
For two-class classification, we have $p(t=1 \mid y) = \sigma(y) = \frac{1}{1+e^{-y}}$,
and $p(t=-1 \mid y) = 1 - \sigma(y) = 1 - \frac{1}{1+e^{-y}} = \frac{e^{-y}}{1+e^{-y}} = \frac{1}{1+e^{y}} = \sigma(-y)$.

Therefore, we can write $p(t \mid y) = \sigma(yt)$

The error function consists of the negative logarithm of the likehood function with a quadratic regularizer

$$\sum_{n=1}^{N} E_{LR}(y_n t_n) + \lambda \| w \|^2 ,$$

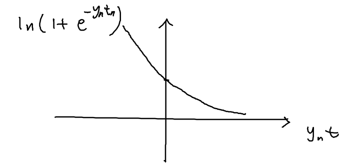
where $E_{LR}(yt) = \ln(1 + \exp(-yt))$,

2009年12月4日
下午 08:39

Given the iid data $D = \{(t_1, x_1), \ldots, (t_N, x_N)\}$, the likelihood function is defined by

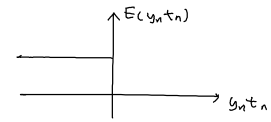
$$p(D) = \prod_{n=1}^{N} \sigma(y_n t_n)$$

$$-\ln P(D) = -\sum_{n=1}^{N} \ln \frac{1}{1+e^{-y_n t_n}} = \sum_{n=1}^{N} \ln(1 + e^{-y_n t_n})$$



logistic error

Both the logistic error and the hinge loss can be viewed as continuous approximations to the misclassification error.



misclassification error
the error function that we ideally we would like to minimize

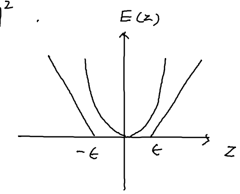
The approximation made by the hinge loss leads to sparse solutions.

SVM for Regression

2009年12月4日
下午 09:01

In linear regression, we minimize a regularized error function given by

$$\frac{1}{2}\sum_{n=1}^{N}\{y_n - t_n\}^2 + \frac{\lambda}{2}\|w\|^2 .$$



To obtain sparse solutions, we replace the quadratic error function by an ϵ-insensitive error function

$$E_\epsilon(y(x)-t) = \begin{cases} 0, & \text{if } |y(x)-t| < \epsilon, \\ |y(x)-t| - \epsilon, & \text{otherwise}. \end{cases}$$

We therefore minimize the following regularized error function

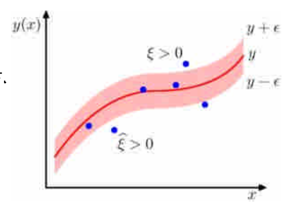
$$C\sum_{n=1}^{N} E_\epsilon(y(x_n)-t_n) + \frac{1}{2}\|w\|^2$$

$y(x_n) = w^T\phi(x_n) + b$

We can re-express the optimization problem by introducing slack variables.

$$t_n \le y(x_n) + \epsilon + \xi_n \qquad \xi_n \ge 0$$
$$t_n \ge y(x_n) - \epsilon - \hat{\xi}_n \qquad \hat{\xi}_n \ge 0$$

if the predicted value y_n lies inside the ϵ-tube, then we do not penalize it.



The optimization problem we want to solve is

minimize $C\sum_{n=1}^{N}(\xi_n + \hat{\xi}_n) + \frac{1}{2}\|w\|^2$

subject to $y_n + \epsilon - t_n + \xi_n \ge 0, \quad \xi_n \ge 0$

$\epsilon + t_n - y_n + \hat{\xi}_n \ge 0, \quad \hat{\xi}_n \ge 0 \qquad n = 1, \dots, N.$

2009年12月4日
下午 09:45

Introducing Lagrange multipliers $a_n \ge 0$, $\hat{a}_n \ge 0$, $\mu_n \ge 0$, $\hat{\mu}_n \ge 0$

The Lagrangian function

$$L = C\sum_{n=1}^{N}(\xi_n + \hat{\xi}_n) + \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N}(\mu_n\xi_n + \hat{\mu}_n\hat{\xi}_n)$$
$$- \sum_{n=1}^{N} a_n(\epsilon + \xi_n + y_n - t_n) - \sum_{n=1}^{N}\hat{a}_n(\epsilon + \hat{\xi}_n - y_n + t_n)$$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{n=1}^{N}(a_n - \hat{a}_n)\phi(x_n)$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{n=1}^{N}(a_n - \hat{a}_n) = 0$$

$$\frac{\partial L}{\partial \xi_n} = 0 \Rightarrow a_n + \mu_n = C$$

$$\frac{\partial L}{\partial \hat{\xi}_n} = 0 \Rightarrow \hat{a}_n + \hat{\mu}_n = C$$

$\mu_n \ge 0, \ \hat{\mu}_n \ge 0$

$\Rightarrow 0 \le a_n, \hat{a}_n \le C$

$$\tilde{L}(a, \hat{a}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \hat{a}_n)(a_m - \hat{a}_m)k(x_n, x_m)$$
$$- \epsilon\sum_{n=1}^{N}(a_n + \hat{a}_n) + \sum_{n=1}^{N}(a_n - \hat{a}_n)t_n$$

$$k(x_n, x_m) = \phi(x_n)^T\phi(x_m)$$

predictions for new inputs can be made by

$$y(x) = \sum_{n=1}^{N}(a_n - \hat{a}_n)k(x, x_n) + b$$

KKT conditions

2009年12月5日
下午 03:27

$$a_n(\epsilon + \xi_n + y_n - t_n) = 0$$
$$\hat{a}_n(\epsilon + \hat{\xi}_n - y_n + t_n) = 0$$
$$(C - a_n)\xi_n = 0$$
$$(C - \hat{a}_n)\hat{\xi}_n = 0$$

① a_n can only be nonzero if $\epsilon + \xi_n + y_n - t_n = 0$, which implies that the data point either lies on the upper boundary of the ϵ-tube ($\xi_n = 0$) or lies above the upper boundary ($\xi_n > 0$, $a_n = C$)

② adding $\epsilon + \xi_n + y_n - t_n = 0$ and $\epsilon + \hat{\xi}_n - y_n + t_n = 0$ but $2\epsilon + \xi_n + \hat{\xi}_n = 0$ is impossible, means a_n and $\hat{a}_n$ cannot be both nonzero.

③ all points within the ϵ-tube have $a_n = \hat{a}_n = 0$
$\Rightarrow$ sparse solution

to decide b ; consider a data point for which $0 < a_n < C$

$$\Rightarrow \xi_n = 0$$
$$\Rightarrow \epsilon + y_n - t_n = 0$$

$$b = t_n - \epsilon - \mathbf{w}^T \phi(\mathbf{x}_n)$$
$$= t_n - \epsilon - \sum_{m=1}^{N} (a_m - \hat{a}_m) k(\mathbf{x}_n, \mathbf{x}_m)$$

RELEVANCE VECTOR MACHINES RVMs

2009年12月5日
下午 03:49

Matrix Identities

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

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

(C.15) $|I_N + \mathbf{a}\mathbf{b}^T| = 1 + \mathbf{a}^T\mathbf{b}$

Gaussian Identities

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

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

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

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

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

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
$\mathbf{x}$: input vector

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

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

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.

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

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

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

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

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

(2.116) where

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

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

$\Phi_{ni} = \phi_i(x_n)$

$A = \mathrm{diag}(\alpha_i)$

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

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

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

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

The optimal weights w^* is given by the mean of the posterior

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

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

We use "evidence approximation".

EA ① Compute the marginal likelihood

$$p(t \mid X, \alpha, \beta) = \int p(t \mid X, w, \beta)\, p(w \mid \alpha)\, dw$$

$$= \int p(t, w \mid X, \alpha, \beta)\, dw$$

$p(\mathbb{t}|\mathbb{X}, \mathbb{w}, \beta)$ and $p(\mathbb{w}, \alpha)$ 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

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

$$= \mathcal{N}(\mathbb{t}|0, C)$$

$$C = \beta^{-1} I + \Phi A^{-1} \Phi^T \qquad A = \mathrm{diag}(\alpha_i)$$

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

$$\ln p(\mathbb{t}|\mathbb{X}, \alpha, \beta) = \ln \mathcal{N}(\mathbb{t}|0, C)$$

$$= -\frac{1}{2}\left\{N \ln(2\pi) + \ln|C| + \mathbb{t}^T C^{-1} \mathbb{t}\right\}$$

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

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

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