2009年10月22日
上午 09:54

NEWTON-RAPHSON ITERATIVE OPTIMIZATION

$$\mathbb{w}^{(new)} = \mathbb{w}^{(old)} - \mathbb{H}^{-1} \nabla E(\mathbb{w}^{(old)})$$

$\mathbb{H}$ is the Hessian matrix whose elements comprise the second derivatives of $E(\mathbb{w})$ w.r.t. $\mathbb{w}$

$$\boxed{H_{ij} = \frac{\partial^2 E}{\partial w_i \partial w_j}}$$

1D Newton's method

$$x^{(new)} = x^{(old)} - g(x^{(old)}) / g'(x^{(old)}) \quad \text{finding } g=0$$

$$x^{(new)} = x^{(old)} - f'(x^{(old)}) / f''(x^{(old)})$$

finding $f' = 0$

Try to apply the Newton-Raphson method to linear regression as a practice

$$\nabla E(\mathbb{w}) = \sum_{n=1}^{N} (\mathbb{w}^T \phi_n - t_n)\phi_n = \Phi^T \Phi \mathbb{w} - \Phi^T \mathbb{t}$$

$$\mathbb{H} = \nabla\nabla E(\mathbb{w}) = \sum_{n=1}^{N} \phi_n \phi_n^T = \Phi^T \Phi$$

Newton's update

$$\mathbb{w}^{(new)} = \mathbb{w}^{(old)} - (\Phi^T\Phi)^{-1}\{\Phi^T\Phi\mathbb{w}^{(old)} - \Phi^T\mathbb{t}\}$$

$$= (\Phi^T\Phi)^{-1}\Phi^T\mathbb{t}$$

$$\Phi = \begin{bmatrix} \vdots \\ \phi_n^T \\ \vdots \end{bmatrix}$$

$N \times M$

we get the standard least-squares solution

2009年10月22日
上午 10:21

Now try to apply the Newton-Raphson method to

$$E(\mathbb{w}) = -\ln p(\mathbb{t}\mid\mathbb{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$$

$$\nabla E(\mathbb{w}) = \sum_{n=1}^{N}(y_n - t_n)\phi_n = \Phi^T(\mathbb{y} - \mathbb{t})$$

$$\mathbb{H} = \nabla\nabla E(\mathbb{w}) = \sum_{n=1}^{N} y_n(1-y_n)\phi_n\phi_n^T = \Phi^T \mathbb{R} \Phi$$

The Hessian is no longer constant; it depends on $\mathbb{w}$ through the weighting matrix $\mathbb{R}$.

$\mathbb{R}$ $N \times N$ diagonal

$R_{nn} = y_n(1-y_n)$

$0 < y_n < 1$.

$\Rightarrow \mathbb{v}^T \mathbb{H} \mathbb{v} > 0$ for an arbitrary $\mathbb{v}$

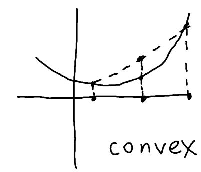
$\Rightarrow \mathbb{H}$ is positive definite

The error function is a convex function of $\mathbb{w}$ and hence has a unique minimum.

$$\mathbb{w}^{(new)} = \mathbb{w}^{(old)} - (\Phi^T\mathbb{R}\Phi)^{-1}\Phi^T(\mathbb{y}-\mathbb{t})$$

$$= (\Phi^T\mathbb{R}\Phi)^{-1}\{\Phi^T\mathbb{R}\Phi\mathbb{w}^{(old)} - \Phi^T(\mathbb{y}-\mathbb{t})\}$$

$$= (\Phi^T\mathbb{R}\Phi)^{-1}\Phi^T\mathbb{R}\mathbb{z}$$



$$\mathbb{z} = \Phi\mathbb{w}^{(old)} - \mathbb{R}^{-1}(\mathbb{y}-\mathbb{t})$$

meaning?

$\mathbb{R}$ depends on $\mathbb{w}$. we need to update $\mathbb{w}$ iteratively

(Iterative Reweighted Least Squares, IRLS)

2009年10月22日
上午 10:39

The weighting matrix $\mathbb{R}$ can be interpreted as variances

$$E[t] = \sigma(x) = y$$

Bournelli P.685
$p(t) = y^t (1-y)^{1-t}$

$$\mathrm{var}[t] = E[t^2] - E[t]^2 = E[t] - E[t]^2 = y - y^2 = y(1-y)$$

$t \in \{0,1\}$
$t^2 = 1$

linearized problem in the space of $a = w^T \phi$

local linear approximation to logistic function

$$\begin{cases} a_n(\mathbb{W}) \simeq a_n(\mathbb{W}^{(old)}) + \left.\dfrac{d a_n}{d y_n}\right|_{\mathbb{W}^{(old)}} (t_n - y_n) \\ \quad = \phi_n^T \mathbb{W}^{(old)} - \dfrac{(y_n - t_n)}{y_n(1-y_n)} = z_n \end{cases}$$

$\boxed{\text{n-th element of } \mathbb{Z}}$

Meaning ?

z_n : as an effective target value in the space obtained by making a local linear approximation to the logistic sigmoid function around the current operating point $\mathbb{W}^{(old)}$

$\sigma = \frac{1}{1+e^{-a}}$

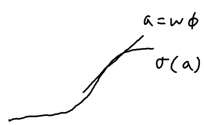
$a = \ln\left(\frac{\sigma}{1-\sigma}\right)$

$\frac{da}{d\sigma} = \frac{1}{\sigma(1-\sigma)}$

$\sigma = y$ see (4.61) (4.88)

(compared with least-squares solutions)

approximate y_n by a_n



MULTCLASS LOGISTIC REGRESSION

2009年10月27日
下午 10:09

K classes

$$p(C_k|\phi) = y_k(\phi) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

$$a_k = w_k^T \phi$$

(softmax posterior

discriminative approach

recall the 1-of-K coding scheme

the target value t_n for feature vector ϕ_n is in the form $\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$ only the kth component equals 1

$\phi_n \in C_k$

likelihood : $p(\mathbb{T}|w_1, \dots, w_K) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(C_k|\phi_n)^{t_{nk}} = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}}$

N×K matrix

$$E(w_1, \dots, w_K) = -\ln p(\mathbb{T}|w_1, \dots, w_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}$$

(cross entropy)

We want to find $w_1, \dots, w_K$ that maximize E

To apply Newton-Raphson Method (IRLS), we need to compute ∇E and the Hessian $\nabla\nabla E$

2009年10月27日
下午 10:25

To begin with

$$\frac{\partial y_k}{\partial a_j} = y_k (I_{kj} - y_j)$$

$$I_{kj} = \begin{cases} 1, & j=k, \\ 0, & \text{otherwise.} \end{cases}$$

$$\frac{\partial y_k}{\partial a_k} = \frac{e^{a_k}}{\sum_i e^{a_i}} - \left(\frac{e^{a_k}}{\sum_i e^{a_i}}\right)^2 = y_k(1-y_k)$$

$$\frac{\partial y_k}{\partial a_j} = \frac{-e^{a_k}e^{a_j}}{(\sum_i e^{a_i})^2} = -y_k y_j \qquad j \neq k$$

$$\frac{\partial E}{\partial y_{nk}} = -\frac{t_{nk}}{y_{nk}}$$

$$\frac{\partial E}{\partial a_{nj}} = \sum_{k=1}^{K} \frac{\partial E}{\partial y_{nk}} \frac{\partial y_{nk}}{\partial a_{nj}} = -\sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}} y_{nk} (I_{kj} - y_{nj})$$
$$= -\sum_{k=1}^{K} t_{nk} (I_{kj} - y_{nj})$$
$$= -t_{nj} + \sum_{k=1}^{K} t_{nk} y_{nj}$$
$$= y_{nj} - t_{nj}$$

1-of-K coding
$\forall n, \sum_k t_{nk} = 1$

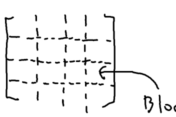
$$a_{nj} = w_j^T \phi_n$$
$$\nabla_{w_j} a_{nj} = \phi_n$$

Compute the gradient w.r.t w_j

$$\nabla_{w_j} E(w_1, \dots, w_K) = \sum_{n=1}^{N} \frac{\partial E}{\partial a_{nj}} \nabla_{w_j} a_{nj} = \sum_{n=1}^{N} (y_{nj} - t_{nj}) \phi_n$$

Similarly the Hessian can be obtained by

$$\nabla_{w_k} \nabla_{w_j} E(w_1, \dots, w_K) = \sum_{n=1}^{N} y_{nk} (I_{kj} - y_{nj}) \phi_n \phi_n^T$$



Blocks of size M×M

2009年10月27日
下午 10:52

PROBIT REGRESSION

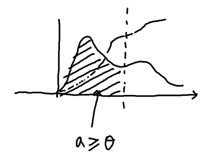
$$p(t=1 \mid a) = f(a) \qquad a = w^T \phi$$

for input ϕ_n, $a_n = w^T \phi_n$

Set the target by $\begin{cases} t_n = 1, & \text{if } a_n \geq \theta, \\ t_n = 0, & \text{otherwise.} \end{cases}$

If the value of θ is drawn from a density $p(\theta)$ the

$$f(a) = \int_{-\infty}^{a} p(\theta)\, d\theta$$



Suppose $p(\theta) \sim \mathcal{N}(\theta \mid 0, 1)$

inverse probit function

$$\Psi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta \mid 0, 1)\, d\theta$$

c.d.f. of a Gaussian

a related function:

$$\operatorname{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a \exp(-\theta^2)\, d\theta$$

$$\Psi(a) = \frac{1}{2} \left\{ 1 + \operatorname{erf}\left(\frac{a}{\sqrt{2}}\right) \right\}$$

in MATLAB
erf
erfinv

The inverse probit function has a similar shape as the logistic sigmoid function.
We will use the inverse probit function in Bayesian logistic regression later.

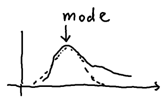
LAPLACE APPROXIMATION

2009年10月27日
下午 11:05

For prediction, we may integrate over the parameter W to get the predictive distribution. (Marginalization)

But the posterior distribution is no longer Gaussian in logistic regression.

How to resolve? Find a Gaussian centered at the mode of the posterior distribution as an approximation



Example:

$$p(z) = \frac{1}{Z} f(z) \qquad Z = \int f(z)\, dz$$

z is a single continuous variable

① mode of $p(z)$: find z_0 s.t. $p'(z_0) = 0$

$$\left.\frac{d}{dz} f(z)\right|_{z=z_0} = 0$$

② Taylor expansion:

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2} a (z - z_0)^2$$

first derivative = 0

$$a = -\left.\frac{d^2}{dz^2} \ln f(z)\right|_{z=z_0}$$

$$f(z) \simeq f(z_0) \exp\left\{-\frac{a}{2}(z - z_0)^2\right\}$$

normalized

$$q(z) = \left(\frac{a}{2\pi}\right)^{\frac{1}{2}} \exp\left\{-\frac{a}{2}(z - z_0)^2\right\}$$

z_0 must be a local maximum $\Rightarrow a > 0$

LAPLACE APPROXIMATION

2009年10月29日
上午 10:17

M-dimensional

$$p(\mathbf{z}) = \frac{1}{Z} f(\mathbf{z})$$

$$\ln f(\mathbf{z}) \simeq \ln f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^T A (\mathbf{z} - \mathbf{z}_0)$$

$$\boxed{A = -\nabla\nabla \ln f(\mathbf{z}) \,\Big|_{\mathbf{z} = \mathbf{z}_0}} \qquad \text{Hessian}$$

$$q(\mathbf{z}) = \frac{|A|^{\frac{1}{2}}}{(2\pi)^{\frac{M}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^T A (\mathbf{z} - \mathbf{z}_0)\right\}$$

$$= \mathcal{N}(\mathbf{z} \mid \mathbf{z}_0, A^{-1})$$

mode $\mathbf{z}_0$ ⇒ local maximum ⇒ A positive definite

So, the Laplace approximation takes two steps:

① find the mode $\mathbf{z}_0$;
② evaluate the Hessian at the mode.

More suitable for large data sets

Weakness: relies on a specific value of the variable, might fail to capture global properties.

BAYESIAN LOGISTIC REGRESSION

2009年10月29日
上午 10:40

Exact Bayesian inference for logistic regression is intractable.
(The likelihood function comprises a product of logistic sigmoid function.)

Use Laplace approximation
fitting a Gaussian centered at the mode of the posterior distribution

Begin with a Gaussian prior $p(w) = \mathcal{N}(w \mid m_0, S_0)$

fixed hyperparameter

posterior $p(w \mid \mathbf{t}) \propto p(w)\, p(\mathbf{t} \mid w)$

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

$$\ln p(w \mid \mathbf{t}) = -\frac{1}{2}(w - m_0)^T S_0^{-1} (w - m_0) + \sum_{n=1}^{N} \{ t_n \ln y_n + (1 - t_n) \ln(1 - y_n) \} + \text{const}$$

(PRML 4.89)

$y_n = \sigma(w^T \phi_n)$

approximated by a Gaussian
↓
what are the mean and the covariance?

2009年10月29日
上午 10:53

the mean : w_{MAP} Maximum a posteriori solution

We may use Newton-Raphson method to compute w_{MAP} as in maximum likelihood
(regularized logistic regress vs. logistic regression)

the covariance :

$$S_N^{-1} = -\nabla\nabla \ln p(w \mid \mathbf{t}) = S_0^{-1} + \sum_{n=1}^{N} y_n (1 - y_n) \phi_n \phi_n^T$$

so $q(w) = \mathcal{N}(w \mid w_{MAP}, S_N)$

is our approximation to the posterior distribution

Predictive Distribution (for class C_1)

Given a new feature vector $\phi(x)$
marginalizing w.r.t. the posterior distribution $p(w \mid \mathbf{t})$

$$p(C_1 \mid \phi, \mathbf{t}) = \int p(C_1 \mid \phi, w)\, p(w \mid \mathbf{t})\, dw \simeq \int \sigma(w^T \phi)\, q(w)\, dw$$

Still intractable since $\sigma(w^T\phi)$ is the logistic sigmoid

2009年10月29日
上午 11:15

An interesting technique

$a = w^T \phi$

we write $\sigma(w^T \phi) = \int \delta(a - w^T \phi)\, \sigma(a)\, da$

δ is the Dirac delta function

$\sigma(w^T\phi)$ is expressed in the form of integration

Therefore, $\int \sigma(w^T \phi)\, q(w)\, dw$

$$= \iint \delta(a - w^T\phi)\, \sigma(a)\, q(w)\, da\, dw$$

$$= \int \sigma(a) \int \delta(a - w^T \phi)\, q(w)\, dw\, da$$

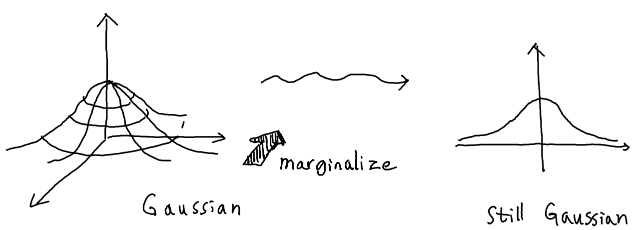
$$= \int \sigma(a)\, p(a)\, da$$

where $p(a) = \int \delta(a - w^T \phi)\, q(w)\, dw$

observe that $\delta(a - w^T\phi)$ imposes a linear constraint on w

$p(a)$ is actually a marginal distribution from the joint distribution $q(w)$ by integrating out all directions orthogonal to ϕ.

$q(w)$ is a Gaussian $\xrightarrow{\text{marginalization}}$ still a Gaussian



2009年10月29日
上午 11:37

what is the mean and variance of $p(a)$

$$\mu_a = E[a] = \int p(a)\, a\, da$$

$$= \iint \delta(a - w^T\phi)\, q(w)\, a\, dw\, da$$

$$= \int q(w) \int \delta(a - w^T\phi)\, a\, da\, dw$$

$$= \int q(w)\, w^T \phi\, dw$$

$$= w^T_{MAP}\, \phi = m_N^T \phi$$

$$s_a^2 = var[a] = \int p(a) \{a^2 - E[a]^2\}\, da$$

$$= \iint \delta(a - w^T \phi)\, q(w) \{a^2 - E[a]^2\}\, dw\, da$$

$$= \int q(w) \int \delta(a - w^T\phi) \{a^2 - E[a]^2\}\, da\, dw$$

$$= \int q(w) \{(w^T\phi)^2 - E[w^T\phi]^2\}\, dw$$

$$= \phi^T S_N \phi$$

the predictive distribution becomes

$$p(C_1 | t) \simeq \int \sigma^T(w^T\phi)\, q(w)\, dw$$

$$= \int \sigma(a)\, p(a)\, da$$

$$= \int \sigma(a)\, \mathcal{N}(a | \mu_a, s_a^2)\, da$$

$\boxed{\text{meaning ?}}$

2009年10月29日
下午 12:02

$$p(C_1|\mathbf{t}) \simeq \int \sigma(a)\, \mathcal{N}(a|\mu_a, s_a^2)\, da$$

convolution of a Gaussian with logistic sigmoid

cannot be evaluated analytically

By approximation again

What kind of function might look like a logistic sigmoid?

recall the inverse probit function (c.d.f of Gaussian)

$$\int \Psi(\lambda_a)\, \mathcal{N}(a|\mu, \sigma^2)\, da = \Psi\left(\frac{\mu}{(\lambda^{-2}+\sigma^2)^{\frac{1}{2}}}\right)$$

requires a lot of work but is doable $\qquad \lambda^2 = \frac{\pi}{8}$

So if we approximate $\sigma(a)$ by the inverse probit function $\Psi(\lambda a)$, we can write down the integration analytically.

$$p(C_1|\phi, \mathbf{t}) \simeq \int \sigma(a)\, \mathcal{N}(a|\mu_a, s_a^2)\, da \simeq \sigma(\kappa(s_a^2)\mu_a)$$

$$\kappa(s_a^2) = (1 + \pi s_a^2/8)^{-\frac{1}{2}}$$

observe that if we have $p(C_1|\phi, \mathbf{t}) = 0.5$ as the decision boundary due to $\mu_a = 0$, which means $w_{MAP}^T \phi = 0$, So the boundary will be the same as the one obtained by the MAP solution. $\Rightarrow$ equal prior

if $\mu_a \neq 0$, $p(C_1|\phi, \mathbf{t}) \neq 0.5$ more general