2009年10月22日
上午 11:09

PROBABILISTIC GENERATIVE MODELS

$p(\mathbf{x} \mid C_k)$ class-conditional densities

$p(C_k)$ class priors

posterior probability $p(C_k \mid \mathbf{x})$

for two classes

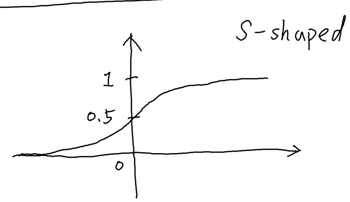
$$p(C_1 \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_1)\, p(C_1)}{p(\mathbf{x} \mid C_1)\, p(C_1) + p(\mathbf{x} \mid C_2)\, p(C_2)}$$

$$= \frac{1}{1 + \exp(-a)} = \sigma(a)$$

$$a = \ln \frac{p(\mathbf{x} \mid C_1)\, p(C_1)}{p(\mathbf{x} \mid C_2)\, p(C_2)}$$

$$\boxed{\sigma(a) = \frac{1}{1+\exp(-a)} \quad \text{logistic sigmoid function}}$$

$$\sigma(-a) = 1 - \sigma(a)$$



$$a = \ln\left(\frac{\sigma}{1-\sigma}\right) \quad \text{logit function}$$

2009年10月22日
上午 11:17

for $K > 2$

$$p(C_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_k)\, p(C_k)}{\sum_j p(\mathbf{x} \mid C_j)\, p(C_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

$$a_k = \ln p(\mathbf{x} \mid C_k)\, p(C_k)$$

$\boxed{\text{softmax function}}$

if $a_k \gg a_j$ for all $j \neq k$

then $p(C_k \mid \mathbf{x}) \simeq 1$, and $p(C_j \mid \mathbf{x}) \simeq 0$

Continuous Inputs (Gaussian)

class-conditional densities

$$p(\mathbf{x} \mid C_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \mu_k)^T \Sigma^{-1} (\mathbf{x} - \mu_k) \right\}$$

shared covariance

two-class case:

$$p(C_1 \mid \mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x} + w_0)$$

$$\mathbf{w} = \Sigma^{-1}(\mu_1 - \mu_2)$$

$$w_0 = -\frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_2^T \Sigma^{-1} \mu_2 + \ln \frac{p(C_1)}{p(C_2)}$$

2009年10月22日
上午 11:25

for K classes

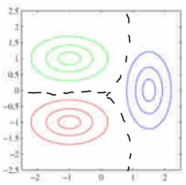
$$a_k(\mathbf{x}) = \mathbf{W}_k^T \mathbf{x} + W_{k0}$$

$$\mathbf{W}_k = \Sigma^{-1} \mu_k$$

$$W_{k0} = -\frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \ln p(C_k)$$

if $p(\mathbf{x}|C_k)$ has its own covariance matrix Σ_k

$\Rightarrow$ quadratic discriminant



How to choose μ_K for $p(\mathbf{x}|C_k)$?

↳ maximum likelihood solution

parameters : $\pi, \mu_1, \mu_2, \Sigma$

$t_n = 1$ for C_1 , $t_n = 0$ for C_2

prior class probality $p(C_1) = \pi$ $\quad p(C_2) = 1 - \pi$

$$p(\mathbf{x}_n, C_1) = p(C_1)\, p(\mathbf{x}_n | C_1) = \pi \mathcal{N}(\mathbf{x}_n | \mu_1, \Sigma)$$

$$p(\mathbf{x}_n, C_2) = p(C_2)\, p(\mathbf{x}_n | C_2) = (1-\pi) \mathcal{N}(\mathbf{x}_n | \mu_2, \Sigma)$$

$$p(\mathbf{t} | \pi, \mu_1, \mu_2, \Sigma)$$

$$= \prod_{n=1}^{N} \left[\pi \mathcal{N}(\mathbf{x}_n | \mu_1, \Sigma)\right]^{t_n} \left[(1-\pi) \mathcal{N}(\mathbf{x}_n | \mu_2, \Sigma)\right]^{1-t_n}$$

2009年10月22日
上午 11:36

maximization w.r.t. π

terms in the log likelihood function that depend on π

$$\sum_{n=1}^{N} \left\{ t_n \ln \pi + (1-t_n) \ln (1-\pi) \right\}$$

set the derivative to 0 $\Rightarrow$ $\pi = \frac{1}{N} \sum_{n=1}^{N} t_n = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}$

maximization w.r.t. μ_1

$$\sum_{n=1}^{N} t_n \ln \mathcal{N}(\mathbf{x}_n | \mu_1, \Sigma) = -\frac{1}{2} \sum_{n=1}^{N} t_n (\mathbf{x} - \mu_1)^T \Sigma^{-1} (\mathbf{x}_n - \mu_1) + \text{const.}$$

set the derivative to 0 $\Rightarrow$ $\mu_1 = \frac{1}{N_1} \sum_{n=1}^{N} t_n \mathbf{x}_n$

similarly $\mu_2 = \frac{1}{N_2} \sum_{n=1}^{N} (1-t_n) \mathbf{x}_n$

maximization w.r.t. Σ

related terms

$$-\frac{1}{2} \sum_{n=1}^{N} t_n \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^{N} t_n (\mathbf{x}_n - \mu_1)^T \Sigma^{-1} (\mathbf{x}_n - \mu_1)$$

$$-\frac{1}{2} \sum_{n=1}^{N} (1-t_n) \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^{N} (1-t_n)(\mathbf{x}_n - \mu_2)^T \Sigma^{-1} (\mathbf{x}_n - \mu_2)$$

$$= -\frac{N}{2} \ln |\Sigma| - \frac{N}{2} \mathrm{Tr}\{\Sigma^{-1} S\}$$

where

$$S = \frac{N_1}{N} S_1 + \frac{N_2}{N} S_2$$

$$S_1 = \frac{1}{N_1} \sum_{n \in C_1} (\mathbf{x}_n - \mu_1)(\mathbf{x}_n - \mu_1)^T$$

$$S_2 = \frac{1}{N_2} \sum_{n \in C_2} (\mathbf{x}_n - \mu_2)(\mathbf{x}_n - \mu_2)^T$$

2009年10月22日
上午 11:37

Set the derivative with respect to Σ to zero
we obtain $\Sigma = S$

$$\left.\begin{array}{l} \frac{\partial}{\partial \Sigma} \ln|\Sigma| = \left(\Sigma^{-1}\right)^T \\ \frac{\partial}{\partial \Sigma} \mathrm{Tr}\{\Sigma^{-1} S\} = (-\Sigma^{-2})^T S \end{array}\right\} \Rightarrow \quad (\Sigma^{-1})^T = (\Sigma^{-2})^T S$$

Discrete Features

$x_i \in \{0, 1\}$

Naive Bayes
feature values are treated as independent conditioned on the class C_k

$$p(\mathbf{x} | C_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i} (1 - \mu_{ki})^{1-x_i} \quad \rightarrow \quad a_k = \ln p(\mathbf{x}|C_k) p(C_k)$$

$$a_k(\mathbf{x}) = \sum_{i=1}^{D} \{ x_i \ln \mu_{ki} + (1-x_i) \ln (1-\mu_{ki}) \} + \ln p(C_k)$$

linear functions of the input values x_i

2009年10月21日
下午 10:45

PROBABILISTIC DISCRIMINATIVE MODELS

difference between probabilistic generative models ?
$\Rightarrow$ indirect approach to finding the parameters of a generalized linear model, by fitting class-conditional densities and class priors separately and then applying Bayes' theorem.

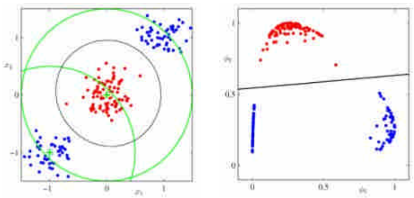
Probabilistic discriminative models :
direct approach : maximizing a likelihood function defined through the conditional distribution $p(C_k|\mathbf{x})$.

An advantage of the discriminative approach :
fewer parameters to be determined
$\Rightarrow$ improved predictive performance

Using fixed basis function again

$\mathbf{x} \rightarrow \phi(\mathbf{x})$

$\phi_0(\mathbf{x}) = 1$



LOGISTIC REGRESSION (for classification not regression)

the posterior probability of class C_1 can be written as a logistic sigmoid acting on a linear function of the feature vector ϕ :

$$p(C_1|\phi) = y(\phi) = \sigma(\mathbf{w}^T \phi)$$

$$p(C_2|\phi) = 1 - p(C_1|\phi)$$

2009年10月21日
下午 11:16

$\sigma(\cdot)$ is the logistic sigmoid function

$$\sigma(a) = \frac{1}{1+\exp(-a)}$$

fewer parameters
- M-dim $\phi \longrightarrow M$ parameters
- For generative models, we need $2M$ parameters for means and $M(M+1)/2$ parameters for covariance matrix

Use maximum likelihood to determine the parameters of the logistic regression model.

To begin with, we write $\frac{d\sigma}{da} = \sigma(1-\sigma)$

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

$$\frac{\partial \sigma}{\partial a} = \frac{1}{(1+e^{-a})^2} \cdot (e^{-a}) = \frac{1}{1+e^{-a}} \cdot \frac{e^{-a}}{1+e^{-a}} = \sigma(1-\sigma)$$

for a data set $\{\phi_n, t_n\}$, where $t_n \in \{0,1\}$, $\phi_n = \phi(x_n)$ with $n = 1, \ldots, N$

the likelihood can be written as

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} \{1-y_n\}^{1-t_n}$$

$$y_n = p(C_1|\phi_n)$$

2009年10月21日
下午 11:36

define an error function by taking the negative logarithm of the likelihood

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

where $y_n = \sigma(a_n)$, $a_n = \mathbf{w}^T \phi_n$.

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n) \phi_n$$

Using $\frac{\partial \sigma}{\partial a} = (1-\sigma)\sigma$

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^{N} \left\{ \frac{t_n}{y_n} (y_n(1-y_n)) \phi_n - \frac{1-t_n}{1-y_n} (1-y_n) y_n \phi_n \right\}$$
$$= -\sum_{n=1}^{N} \{ t_n(1-y_n)\phi_n - (1-t_n) y_n \phi_n \}$$

So the contribution to the gradient from data point n is given by the error $y_n - t_n$ times the basis function vector ϕ_n

(no closed-form solution)

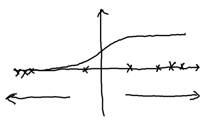
We may derive a sequential-update algorithm

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta (y_n - t_n) \phi_n$$

2009年10月22日
上午 09:27

maximum likelihood
overfitting for data sets that are linearly separable

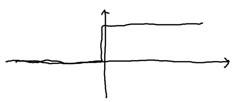
multiple solutions which one is the best ?
we need to include a prior for MAP , or add a regularization term



$w^T \phi = 0 \quad \sigma = 0.5$
decision plane

$w^T \phi_n \geq 0 \quad t_n = 1 \quad , \quad w^T \phi_n < 0 \quad t_n = 0$

⇓



Heaviside step function

ITERATIVE REWEIGHTED LEAST SQUARES

Recall : linear regression , maximum likelihood solution
assume a Gaussian noise model
→ closed-form solution

For logistic regression , no closed-form solution

$$\nabla E(w) = \sum_{n=1}^{N} (y_n - t_n)\phi_n \overset{?}{=} 0$$

$$y_n = \sigma(w^T \phi_n) \qquad \sum_n \sigma \quad \sigma^{-1} \quad ?$$

2009年10月22日
上午 09:54

NEWTON - RAPHSON ITERATIVE OPTIMIZATION

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

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

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

1D Newton's method

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

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

finding $f' = 0$

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

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

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

Newton's update

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

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

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

$N \times M$

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

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

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

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

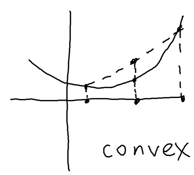
The Hessian is no longer constant; it depends on $\mathbf{w}$ through the weighting matrix $\mathbb{R}$.

$\mathbb{R}$ $N\times N$ diagonal
$R_{nn} = y_n(1-y_n)$

$0 < y_n < 1$.
$\Rightarrow$ $\mathbf{v}^T \mathbb{H}\,\mathbf{v} > 0$ for an arbitrary $\mathbf{v}$
$\Rightarrow$ $\mathbb{H}$ is positive definite

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

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



convex

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

meaning?

$\mathbb{R}$ depends on $\mathbf{w}$, we need to update $\mathbf{w}$ 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

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

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

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

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

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

local linear approximation to logistic function

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

$\boxed{\text{n-th element of } \mathbf{z}}$

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 $\mathbf{w}^{(old)}$

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

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

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

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

(compared with least-squares solutions)

approximate y_n by a_n