2009年10月13日
下午 10:50

# Linear Models for Classification

$C_k$ : $K$ discrete class $\quad k = 1, \dots, K$

usually $C_k$ are disjoint

decision boundaries — $(D-1)$-dim hyperplanes within $D$-dim input space

input space is divided into decision regions

linearly separable
- number of classes
- dimensions

binary-class $\quad t \in \{0, 1\}$

multi-class $\quad t = (0, 1, 0, 0, 0)^T$ | 1-of-K scheme

different approaches

① discriminant functions

② generative $\quad p(C_k|\mathbf{x}) = \dfrac{p(\mathbf{x}|C_k)\, p(C_k)}{p(\mathbf{x})}$

③ directly model the conditional $\quad p(C_k|\mathbf{x})$



# Discriminant Functions

2009年10月20日
下午 03:22

## linear discriminant

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$$ ← bias | negative bias ≡ threshold

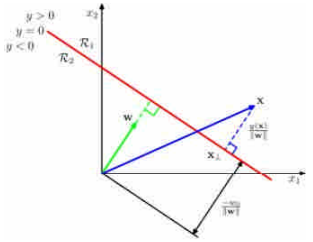
$$\begin{cases} C_1, & \text{if } y(\mathbf{x}) > 0, \\ C_2, & \text{otherwise.} \end{cases}$$

decision plane : $\quad y(\mathbf{x}) = 0$

$\mathbf{w}$ is orthogonal to every vector lying within the decision surface

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = 0$$

$$\hookrightarrow \quad \frac{\mathbf{w}^T\mathbf{x}}{\|\mathbf{w}\|} = -\frac{w_0}{\|\mathbf{w}\|}$$




$$\mathbf{x} = \mathbf{x}_\perp + r\frac{\mathbf{w}}{\|\mathbf{w}\|}$$

$$\begin{cases} y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 \\ y(\mathbf{x}_\perp) = \mathbf{w}^T\mathbf{x}_\perp + w_0 = 0 \end{cases}$$

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = \mathbf{w}^T\mathbf{x}_\perp + w_0 + r\frac{\mathbf{w}^T\mathbf{w}}{\|\mathbf{w}\|}$$

$$y(\mathbf{x}) = r\|\mathbf{w}\|$$

$$r = \frac{y(\mathbf{x})}{\|\mathbf{w}\|}$$

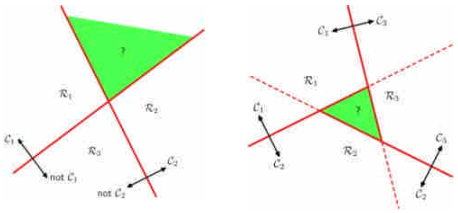
2009年10月20日
下午 03:33

# Multiple Classes

one-versus-the-rest classifier

one-versus-one classifier

ambiguous region







alternative way

$y_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x} + w_{k0}$

$\mathbf{x} \in C_k$ if $y_k(\mathbf{x}) > y_j(\mathbf{x})$ for all $j \neq k$


convex region

$\hat{\mathbf{x}} = \lambda \mathbf{x}_A + (1-\lambda)\mathbf{x}_B \quad 0 \leq \lambda \leq 1$

$\Rightarrow y_k(\hat{\mathbf{x}}) = \lambda y_k(\mathbf{x}_A) + (1-\lambda) y_k(\mathbf{x}_B)$

$y_k(\mathbf{x}_A) > y_j(\mathbf{x}_A), \; y_k(\mathbf{x}_B) > y_j(\mathbf{x}_B)$

for all $j \neq k$

$\Rightarrow \quad y_k(\hat{\mathbf{x}}) > y_j(\hat{\mathbf{x}})$



2009年10月20日
下午 03:40

# Least Squares for Classification

K binary coding

$y_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x} + w_{k0}$

$\mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}^T \widetilde{\mathbf{x}}$

$y_k = \widetilde{\mathbf{w}}_k^T \widetilde{\mathbf{x}}$

dummy input $x_0 = 1$

$\widetilde{\mathbf{w}} = (w_0, \mathbf{w})$

$\widetilde{\mathbf{x}} = (x_0, \mathbf{x})$

$\widetilde{\mathbf{W}} = [\widetilde{\mathbf{w}}_1, \widetilde{\mathbf{w}}_2, \ldots, \widetilde{\mathbf{w}}_K]$

$\{\mathbf{x}_n, \mathbf{t}_n\}$ training data

$$\widetilde{\mathbf{X}} = \begin{bmatrix} \widetilde{\mathbf{x}}_1^T \\ \vdots \\ \widetilde{\mathbf{x}}_N^T \end{bmatrix} \qquad \mathbf{T} = \begin{bmatrix} \mathbf{t}_1^T \\ \vdots \\ \mathbf{t}_N^T \end{bmatrix}$$

sum-of-squares error function

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2} \mathrm{Tr}\left\{ (\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})^T (\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T}) \right\}$$

set the derivatives w.r.t. $\widetilde{\mathbf{W}}$ to zero

$$\widetilde{\mathbf{W}}^* = (\widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^T \mathbf{T} = \widetilde{\mathbf{X}}^\dagger \mathbf{T}$$

$$\mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}^T \widetilde{\mathbf{x}} = \mathbf{T}^T (\widetilde{\mathbf{X}}^\dagger)^T \widetilde{\mathbf{x}}$$

closed-form solution

Problems: ① outliers Fig 4.4 ② assume a Gaussian conditional distribution — binary target vector not Gaussian

2009年10月20日
下午 08:11

# Fisher's linear discriminant or (LDA)

$N_1$ points in $C_1$ $N_2$ points in $C_2$

$$m_1 = \frac{1}{N_1}\sum_{n\in C_1} x_n \quad , \quad m_2 = \frac{1}{N_2}\sum_{n\in C_2} x_n$$

projected onto $w$

$$m_2 - m_1 = w^T(m_2 - m_1) \qquad \left| \; m_k = w^T m_k \right.$$

MAXIMIZE

$$\sum_i w_i^2 = 1$$

Lagrange multiplier

maximize $w^T(m_2 - m_1)$
subject to $w^T w = 1$

$$w \propto (m_2 - m_1) \qquad \left| \; (m_2 - m_1) + 2\lambda w = 0 \right.$$

NOT GOOD ENOUGH — MINIMIZE VARIANCE

$$s_k^2 = \sum_{n\in C_k} (y_n - m_k)^2 \qquad \left| \; y_n = w^T x_n \right.$$

total within-class variance $s_1^2 + s_2^2$



2009年10月20日
下午 08:52

$$J(w) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$ Fisher criterion

$$J(w) = \frac{w^T S_B w}{w^T S_w w}$$

between-class scatter matrix: $S_B = (m_2 - m_1)(m_2 - m_1)^T$

(total) within-class scatter matrix:

$$S_w = \sum_{n\in C_1} (x_n - m_1)(x_n - m_1)^T + \sum_{n\in C_2} (x_n - m_2)(x_n - m_2)^T$$

$$\frac{\partial}{\partial w} J(w) = 0$$

$$\frac{2 S_B w}{w^T S_w w} + \frac{w^T S_B w}{(w^T S_w w)^2}\left(-2 S_w w\right) = 0$$

$$\underbrace{(w^T S_B w)}_{\text{scalar}} S_w w = \underbrace{(w^T S_w w)}_{\text{scalar}} S_B w$$

Since $S_B w = (m_2 - m_1)(m_2 - m_1)^T w$

$S_B w$ is in the direction of $(m_2 - m_1)$

we have

$$w \propto S_w^{-1}(m_2 - m_1)$$

# RELATION TO LEAST SQUARES

2009年10月20日
下午 09:06

least-squares approach: making the model predictions as close as possible to a set of target values

Fisher criterion: maximum class separation in the output space

## Fisher criterion as a special case of least squares

Let the target values be $\frac{N}{N_1}$ and $-\frac{N}{N_2}$

($\frac{N}{N_1}$: $|C_1|$, $-\frac{N}{N_2}$: $|C_2|$)

sum-of-squares error function

$$E = \frac{1}{2}\sum_{n=1}^{N}\left(w^T x_n + w_0 - t_n\right)^2$$

$$\frac{\partial E}{\partial w_0} = 0 \qquad \sum_{n=1}^{N}\left(w^T x_n + w_0 - t_n\right) = 0$$

$$w_0 = -w^T m$$

$$\sum t_n = N_1\frac{N}{N_1} - N_2\frac{N}{N_2} = 0$$

$$\frac{\partial E}{\partial w} = 0 \qquad \sum_{n=1}^{N}\left(w^T x_n + w_0 - t_n\right)x_n = 0$$

$$\sum_{n=1}^{N} w^T x_n x_n + N m w_0 - \sum_{n=1}^{N} t_n x_n = 0$$

after one-page derivation

$$\left(S_W + \frac{N_1 N_2}{N} S_B\right) w = N(m_1 - m_2)$$

$$m = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{1}{N}\left(N_1 m_1 + N_2 m_2\right)$$

$$S_B w \propto (m_2 - m_1) \Rightarrow w \propto S_W^{-1}(m_2 - m_1)$$

same as Fisher



# FISHER'S DISCRIMINANT FOR MULTIPLE CLASSES

2009年10月20日
下午 09:24

$K > 2$

assume that the dimensionality $D$ of the input space is greater than the number $K$ of classes

we want to get $D'$ linear features

$$y_k = w_k^T x \qquad k = 1, \ldots, D'$$

weight vectors $\{w_k\}$

$$y = W^T x$$

within-class $S_W = \sum_{k=1}^{K} S_k$

$$S_k = \sum_{n \in C_k}(x_n - m_k)(x_n - m_k)^T$$

$$m_k = \frac{1}{N_k}\sum_{n \in C_k} x_n$$

total between-class $S_T = \sum_{n=1}^{N}(x_n - m)(x_n - m)^T$

$$m = \frac{1}{N}\sum_{k=1}^{K} N_k m_k \qquad N = \sum_{k=1}^{K} N_k$$

$$S_T = S_W + S_B$$

$$S_B = \sum_{k=1}^{K} N_k (m_k - m)(m_k - m)^T$$

2009年10月20日
下午 09:32

in the projected $D'$-dimensional $y$-space

$$S_W = \sum_{k=1}^{K} \sum_{n \in C_k} (y_n - \mu_k)(y_n - \mu_k)^T$$

$$S_B = \sum_{k=1}^{K} N_k (\mu_k - \mu)(\mu_k - \mu)^T$$

$$\mu_k = \frac{1}{N_k} \sum_{n \in C_k} y_n \quad , \quad \mu = \frac{1}{N} \sum_{k=1}^{K} N_k \mu_k$$

Criterion ?

$$J(W) = Tr\{S_W^{-1} S_B\}$$

$$\Downarrow$$

$$J(W) = Tr\{(W S_W W^T)^{-1} (W S_B W^T)\}$$

other criteria

$$J_2 = \ln |S_W^{-1} S_B| = \ln |S_B| - \ln |S_W|$$

$$J_3 = Tr\{S_B\} - \mu (Tr\{S_W\} - C)$$

$$J_4 = \frac{Tr\{S_B\}}{Tr\{S_W\}}$$

How to find $W$ that maximizes $J(W)$ ?



2009年10月20日
下午 09:44

$$J(W) = Tr\{S_W^{-1} S_B\} = Tr\{(W S_W W^T)^{-1} (W S_B W^T)\}$$

$$\frac{\partial J}{\partial W} = 0 \Rightarrow -2 S_W W S_W^{-1} S_B S_W^{-1} + 2 S_B W S_W^{-1} = 0$$

$$(S_W^{-1} S_B) W = W (S_W^{-1} S_B)$$

$S_W$, $S_B$ simultaneously diagonalized by $B$

$$B^T S_B B = \Lambda \quad , \quad B^T S_W B = I$$

If we apply the transformation $B$ to $y$
the value of the criterion will not change

$$Tr\{\tilde{S}_W^{-1} \tilde{S}_B\} = Tr\{(B^T S_W B)^{-1} (B^T S_B B)\}$$
$$= Tr\{B^{-1} S_W^{-1} S_B B\}$$
$$= Tr\{S_W^{-1} S_B B B^{-1}\}$$
$$= Tr\{S_W^{-1} S_B\}$$

$$S_B = B^{-T} \Lambda B^{-1} \quad , \quad S_W = B^{-T} B^{-1}$$

$$(S_W^{-1} S_B) W = W (B^{-T} B^{-1})^{-1} (B^{-T} \Lambda B^{-1}) = W B \Lambda B^{-1}$$

$$(S_W^{-1} S_B)(WB) = (WB) \Lambda$$

$\Lambda$: eigenvalues of $(S_W^{-1} S_B)$

Solution: Find the eigenvectors corresponding to the $D'$ largest eigenvalues of $S_W^{-1} S_B$

# THE PERCEPTRON ALGORITHM

2009年10月20日
下午 10:08

$$y(x) = f(w^T \phi(x))$$

$f(\cdot)$ nonlinear activation function

$$f(a) = \begin{cases} +1, & a \geq 0, \\ -1, & a < 0. \end{cases}$$

$\phi_0(x) = 1$ bias component

$t = +1$ for class $C_1$
$t = -1$ for class $C_2$

Goal: find $w$ such that $w^T \phi(x_n) t_n > 0$ for $n = 1, \ldots, N$

perceptron criterion

$$E_p(w) = -\sum_{n \in N} w^T \phi(x_n) t_n$$

minimize the criterion using stochastic gradient descent

one input at a time

if $w^{(\tau)T} \phi(x_n) t_n < 0$

$x_n$ is misclassified by $w^{(\tau)}$

update $w^{(\tau+1)} = w^{(\tau)} - \eta \nabla E_p(w) = w^{(\tau)} + \eta \phi(x_n) t_n$

$\eta$ is the learning rate parameter



2009年10月20日
下午 10:17

$\phi_n \equiv \phi(x_n)$

$\| \phi_n t_n \|^2 > 0$

$$E_p(w^{(\tau+1)}) = -w^{(\tau+1)T} \phi_n t_n$$
$$= -w^{(\tau)T} \phi_n t_n - (\phi_n t_n)^T (\phi_n t_n)$$
$$< -w^{(\tau)T} \phi_n t_n = E_p(w^{(\tau)})$$

$E_p(w)$ decreases for $(x_n, t_n)$

(The algorithm converges in a finite number of steps if the data are linearly separable)

## PERCEPTRON CONVERGENCE THEOREM

Let $R = \max \| \phi(x_n) \|$
Suppose that there exists a vector $w^*$ such that $\| w^* \| = 1$ and $t_n (w^{*T} \phi(x_n)) \geq \gamma$ for $n = 1, \ldots, N$
then the number of mistakes made by the perceptron algorithm on the training data is at most $\left(\frac{R}{\gamma}\right)^2$

Proof.

$$w^{(\tau+1)T} w^* = w^{(\tau)T} w^* + \eta (\phi(x_n) t_n)^T w^* \geq w^{(\tau)T} w^* + \eta \gamma$$

apply the inequality $\tau$ times $\Rightarrow$ $w^{(\tau+1)T} w^* \geq \tau \eta \gamma$ (1)

2009年10月20日
下午 10:36

$$\|w^{(\tau+1)}\|^2 = \|w^{(\tau)}\|^2 + \underbrace{2\eta t_n w^\tau \phi(x_n)}_{<0} + \eta^2 t_n^2 \|\phi(x_n)\|^2$$

$$\|w^{(\tau+1)}\|^2 \leq \|w^{(\tau)}\|^2 + \eta^2 \|\phi(x_n)\|^2 \qquad \Big| \; t_n^2 = 1$$

$$\leq \|w^{(\tau)}\|^2 + \eta^2 R^2$$

apply the inequality $\tau$ times $\Rightarrow \|w^{(\tau+1)}\|^2 \leq \tau \eta^2 R^2$ (2)

Combine (1) and (2)

$$\|w^*\| \sqrt{\tau}\, \eta R \geq \|w^*\| \underbrace{\|w^{(\tau+1)}\|}_{\text{Cauchy–Schwarz inequality}} \geq w^{(\tau+1)^T} w^* \geq \tau \eta \gamma$$

$$\|w^*\|^2 \tau \eta^2 R^2 \geq \tau^2 \eta^2 \gamma^2$$

$$\tau \leq \left(\frac{R}{\gamma}\right)^2$$

---

Problems of the perceptron algorithm

① cannot distinguish between a non separable problem and one that is simply slow to converge
② solution is not unique
③ not linearly separable → never converges
④ no probabilistic output
⑤ binary classification
⑥ linear combination of fixed basis functions