Chapter 5

Linear Discriminant Functions

Discriminant Function

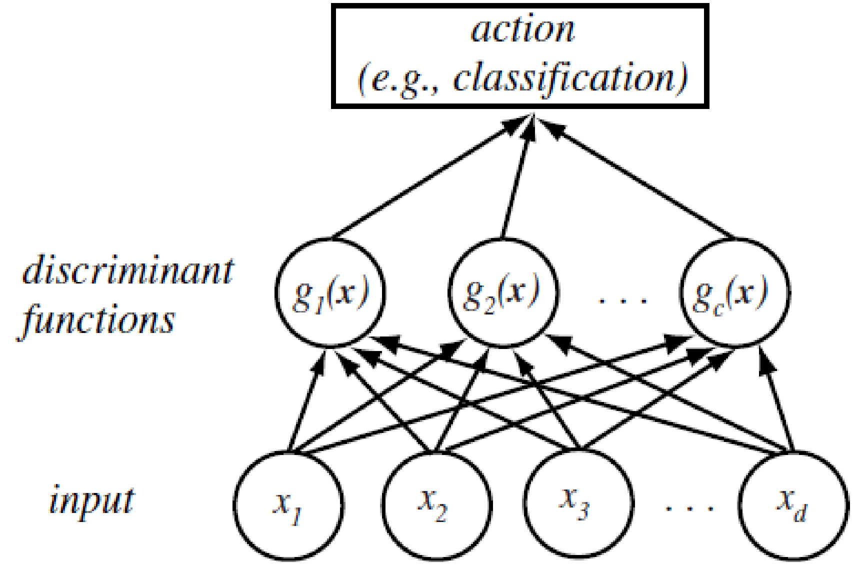
Discriminant functions

$$g_i : \mathbf{R}^d \to \mathbf{R} \quad (1 \leq i \leq c)$$

- *Useful way to represent classifiers*
- *One function per category*

Decide ω_i

if $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for all $j \neq i$



Minimum risk: $g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x}) \quad (1 \leq i \leq c)$

Minimum-error-rate: $g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) \quad (1 \leq i \leq c)$

Discriminant Function (Cont.)

Decision region

c discriminant functions → c decision regions

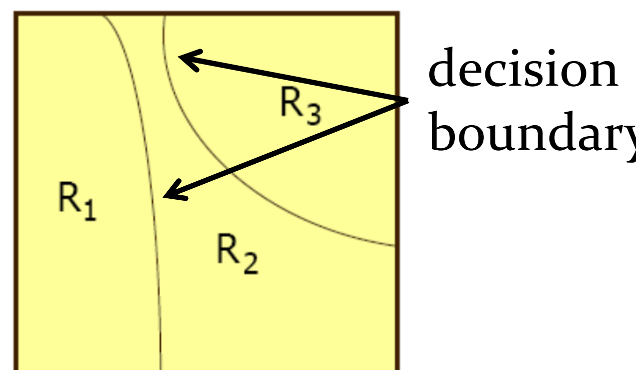
$g_i(\cdot)\ (1 \leq i \leq c)$ → $\mathcal{R}_i \subset \mathbf{R}^d\ (1 \leq i \leq c)$

$$\mathcal{R}_i = \{\mathbf{x} \mid \mathbf{x} \in \mathbf{R}^d : g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \neq i\}$$

where $\mathcal{R}_i \cap \mathcal{R}_j = \emptyset\ (i \neq j)$ and $\bigcup_{i=1}^{c} \mathcal{R}_i = \mathbf{R}^d$

Decision boundary

surface in feature space where ties occur among several largest discriminant functions



Linear Discriminant Functions

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + w_{i0} \qquad (i = 1, 2, \ldots, c)$$

$\mathbf{w}_i$: weight vector (权值向量, *d*-dimensional)

w_{i0} : bias/threshold (偏置/阈值, scalar)

$\mathbf{x} = (x_1, x_2, x_3)^t$ $\quad d = 3,\ c = 3$

$g_1(\mathbf{x}) = x_1 - 2x_2 + 4x_3$ $\quad \mathbf{w}_1 = (1, -2, 4)^t,\ w_{10} = 0$

$g_2(\mathbf{x}) = x_1 + 3x_3 + 4$ $\quad \mathbf{w}_2 = (1, 0, 3)^t,\ w_{20} = 4$

$g_3(\mathbf{x}) = -2$ $\quad \mathbf{w}_3 = (0, 0, 0)^t,\ w_{30} = -2$

Linear Discriminant Functions (Cont.)

The two-category case

$$g_1(\mathbf{x}) = \mathbf{w}_1^t\,\mathbf{x} + w_{10}$$

$$g_2(\mathbf{x}) = \mathbf{w}_2^t\,\mathbf{x} + w_{20}$$

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$$

Decide ω_1 if $g(\mathbf{x}) > 0$

Decide ω_2 otherwise

$$\begin{aligned} g(\mathbf{x}) &= g_1(\mathbf{x}) - g_2(\mathbf{x}) \\ &= (\mathbf{w}_1^t\,\mathbf{x} + w_{10}) - (\mathbf{w}_2^t\,\mathbf{x} + w_{20}) \\ &= (\mathbf{w}_1^t - \mathbf{w}_2^t)\,\mathbf{x} + (w_{10} - w_{20}) \\ &= (\mathbf{w}_1 - \mathbf{w}_2)^t\,\mathbf{x} + (w_{10} - w_{20}) \end{aligned}$$

Let $\mathbf{w} = \mathbf{w}_1 - \mathbf{w}_2$

$b = w_{10} - w_{20}$

$$g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + b$$

It suffices to consider only d+1 parameters ($\mathbf{w}$ and b) instead of 2(d+1) parameters under two-category case

Two-Category Case

Training set

$$\mathcal{D}^* = \{(\mathbf{x}_i, \omega_i) \mid i = 1, 2, \ldots, n\} \quad (\mathbf{x}_i \in \mathbf{R}^d, \omega_i \in \{-1, +1\})$$

The task

Determine $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + b$ which can classify all training examples in $\mathcal{D}^*$ correctly

$$g(\mathbf{x}_i) = \mathbf{w}^t\mathbf{x}_i + b > 0 \text{ if } \omega_i = +1$$

$$g(\mathbf{x}_i) = \mathbf{w}^t\mathbf{x}_i + b < 0 \text{ if } \omega_i = -1$$

$$\omega_i \cdot (\mathbf{w}^t\mathbf{x}_i + b) > 0 \;\; (i = 1, 2, \ldots, n)$$

Two-Category Case (Cont.)

Solution to (w, *b*) $(g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + b)$

Minimize a criterion/objective function (准则函数) $J(\mathbf{w},b)$ based on the training examples $\{(\mathbf{x}_i, \omega_i) \mid i = 1, 2, \ldots, n\}$

$$J(\mathbf{w}, b) = -\sum_{i=1}^{n} \mathtt{sign}[\omega_i \cdot g(\mathbf{x}_i)]$$

$$J(\mathbf{w}, b) = -\sum_{i=1}^{n} \omega_i \cdot g(\mathbf{x}_i)$$

$$J(\mathbf{w}, b) = \sum_{i=1}^{n} (g(\mathbf{x}_i) - \omega_i)^2$$

……

How to minimize the criterion function $J(\mathbf{w},b)$?

Gradient Descent

(梯度下降)

Gradient Descent

Taylor Expansion (泰勒展式)

$$f(\mathbf{x}+\Delta\mathbf{x}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^t \cdot \Delta\mathbf{x} + O(\Delta\mathbf{x}^t \cdot \Delta\mathbf{x})$$

$f : \mathbf{R}^d \rightarrow \mathbf{R}$: a real-valued *d*-variate **function**

$\mathbf{x} \in \mathbf{R}^d$: **a point** in the *d*-dimensional Euclidean space

$\Delta\mathbf{x} \in \mathbf{R}^d$: a **small shift** in the *d*-dimensional Euclidean space

$\nabla f(\mathbf{x})$: **gradient** of $f(\cdot)$ at $\mathbf{x}$

$O(\Delta\mathbf{x}^t \cdot \Delta\mathbf{x})$: the **big oh order** of $\Delta\mathbf{x}^t \cdot \Delta\mathbf{x}$ [appendix A.8]

Gradient Descent (Cont.)

Taylor Expansion (泰勒展式)

$$f(\mathbf{x}+\Delta\mathbf{x}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^t \cdot \Delta\mathbf{x} + O(\Delta\mathbf{x}^t \cdot \Delta\mathbf{x})$$

What happens if we set $\Delta\mathbf{x}$ to be *negatively proportional* to the gradient at x, i.e.:

$$\Delta\mathbf{x} = -\eta \cdot \nabla f(\mathbf{x})$$ (η being a *small* positive scalar)

$$f(\mathbf{x}+\Delta\mathbf{x}) = f(\mathbf{x}) - \eta \cdot \underline{\underline{\nabla f(\mathbf{x})^t \cdot \nabla f(\mathbf{x})}} + \underline{\underline{O(\Delta\mathbf{x}^t \cdot \Delta\mathbf{x})}}$$

being *non-negative*

ignored when $O(\Delta\mathbf{x}^t \cdot \Delta\mathbf{x})$ is small

Therefore, we have $f(\mathbf{x}+\Delta\mathbf{x}) \leq f(\mathbf{x})$!

Gradient Descent (Cont.)

Basic strategy

To minimize some d-variate function $f(\cdot)$, the general gradient descent techniques work in the following *iterative way*:

1. Set learning rate $\eta > 0$ and a small threshold $\epsilon > 0$
2. Randomly initialize $\mathbf{x}_0 \in \mathbf{R}^d$ as the starting point; Set k=0
3. **do** k=k+1
4. $\mathbf{x}_k = \mathbf{x}_{k-1} - \eta \cdot \nabla f(\mathbf{x}_{k-1})$ (*gradient descent step*)
5. **until** $|f(\mathbf{x}_k) - f(\mathbf{x}_{k-1})| < \epsilon$
6. Return $\mathbf{x}_k$ and $f(\mathbf{x}_k)$

Gradient Descent for Two-Category Linear Discriminant Functions

Task revisited

Determine (**w**,b) such that $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + b$ can classify all examples in $\mathcal{D}^*$ correctly, where $\mathcal{D}^* = \{(\mathbf{x}_i, \omega_i) \mid 1 \leq i \leq n\}$

The solution

Choose certain criterion function J(**w**,b) defined over $\mathcal{D}^*$ [ref: slide 8]

→

Invoke the standard gradient descent procedure on the (d+1)-variate function $J(\cdot,\cdot)$ to determine (**w**,b)

Gradient Descent for Two-Category Linear Discriminant Functions (Cont.)

Two examples

$$J(\mathbf{w}, b) = -\sum_{i=1}^{n} \omega_i \cdot g(\mathbf{x}_i)$$

$$\nabla J(\mathbf{w}, b) = -\sum_{i=1}^{n} \omega_i \cdot \begin{bmatrix} \mathbf{x}_i \\ 1 \end{bmatrix}$$

$$J(\mathbf{w}, b) = \sum_{i=1}^{n} (g(\mathbf{x}_i) - \omega_i)^2$$

$$\nabla J(\mathbf{w}, b) = 2 \cdot \sum_{i=1}^{n} (\mathbf{w}^t \mathbf{x}_i + b - \omega_i) \cdot \begin{bmatrix} \mathbf{x}_i \\ 1 \end{bmatrix}$$

Summary

- Discriminant functions
- Linear discriminant functions
  - The general setting: $g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0} \quad (i = 1, 2, \ldots, c)$
  - The two-category case: $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + b$
  - Minimization of criterion/objection function

$$J(\mathbf{w}, b) = -\sum_{i=1}^{n} \mathtt{sign}[\omega_i \cdot g(\mathbf{x}_i)]$$

$$J(\mathbf{w}, b) = -\sum_{i=1}^{n} \omega_i \cdot g(\mathbf{x}_i)$$

$$J(\mathbf{w}, b) = \sum_{i=1}^{n} (g(\mathbf{x}_i) - \omega_i)^2 \qquad \ldots\ldots$$

Summary (Cont.)

- Gradient descent
  - Taylor expansion

$$f(\mathbf{x}+\Delta\mathbf{x}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^t \cdot \Delta\mathbf{x} + O(\Delta\mathbf{x}^t \cdot \Delta\mathbf{x})$$

  - Key iterative gradient descent step

$$\mathbf{x}_k = \mathbf{x}_{k-1} - \eta \cdot \nabla f(\mathbf{x}_{k-1})$$

  - For two-category linear discriminant functions

$$J(\mathbf{w}, b) = -\sum_{i=1}^{n} \omega_i \cdot g(\mathbf{x}_i) \Rightarrow \nabla J(\mathbf{w}, b) = -\sum_{i=1}^{n} \omega_i \cdot \begin{bmatrix} \mathbf{x}_i \\ 1 \end{bmatrix}$$

$$J(\mathbf{w}, b) = \sum_{i=1}^{n} (g(\mathbf{x}_i) - \omega_i)^2 \Rightarrow \nabla J(\mathbf{w}, b) = 2 \cdot \sum_{i=1}^{n} (\mathbf{w}^t\mathbf{x}_i + b - \omega_i) \cdot \begin{bmatrix} \mathbf{x}_i \\ 1 \end{bmatrix}$$