



Lecture Slides for

INTRODUCTION TO

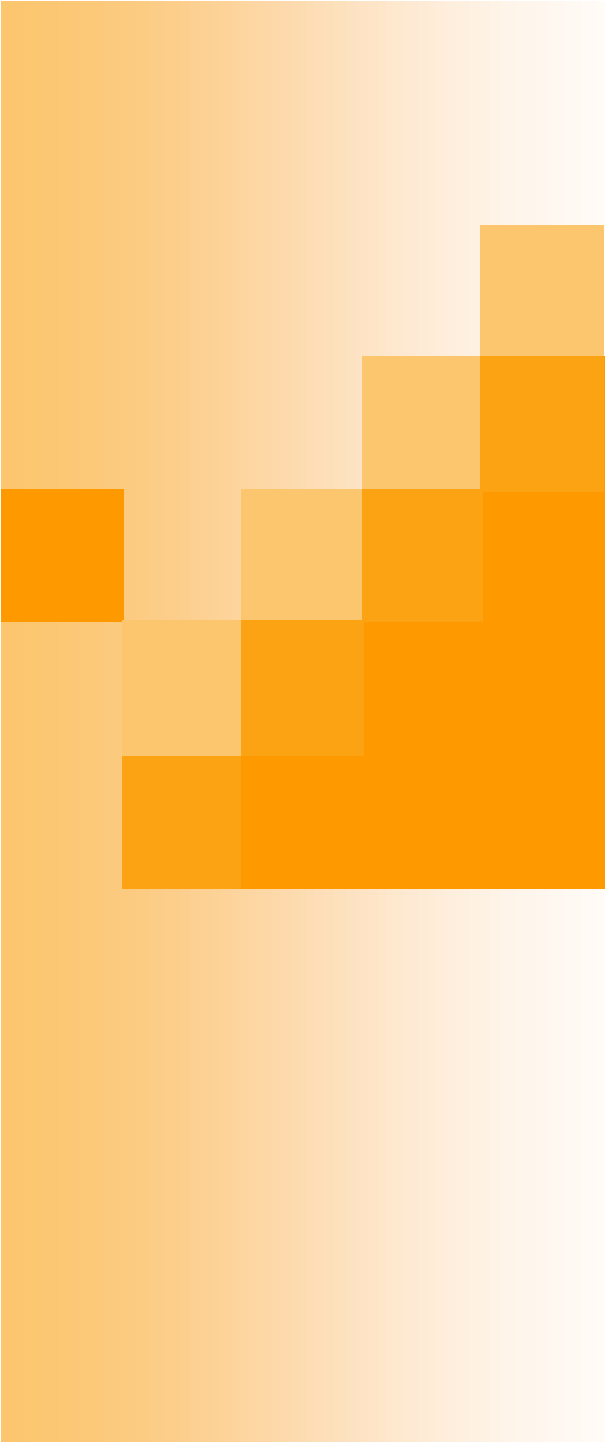
Machine Learning

ETHEM ALPAYDIN

© The MIT Press, 2004

alpaydin@boun.edu.tr

<http://www.cmpe.boun.edu.tr/~ethem/i2ml>

A decorative graphic on the left side of the slide consisting of a grid of orange squares of varying shades and sizes, arranged in a stepped pattern.

CHAPTER 10:

Linear Discrimination



Likelihood- vs. Discriminant-based Classification

- **Likelihood-based:** Assume a model for $p(\mathbf{x}|C_i)$, use Bayes' rule to calculate $P(C_i|\mathbf{x})$
$$g_i(\mathbf{x}) = \log P(C_i|\mathbf{x})$$
- **Discriminant-based:** Assume a model for $g_i(\mathbf{x}|\Phi_i)$; no density estimation
- Estimating the boundaries is enough; no need to accurately estimate the densities inside the boundaries



Linear Discriminant

- Linear discriminant:

$$g_i(\mathbf{x}|\mathbf{w}_i, w_{i0}) = \mathbf{w}_i^T \mathbf{x} + w_{i0} = \sum_{j=1}^d w_{ij}x_j + w_{i0}$$

- Advantages:

- Simple: $O(d)$ space/computation
- Knowledge extraction: Weighted sum of attributes; positive/negative weights, magnitudes (credit scoring)
- Optimal when $p(\mathbf{x}|C_i)$ are Gaussian with shared cov matrix; useful when classes are (almost) linearly separable



Generalized Linear Model

- Quadratic discriminant:

$$g_i(\mathbf{x}|\mathbf{W}_i, \mathbf{w}_i, w_{i0}) = \mathbf{x}^T \mathbf{W}_i \mathbf{x} + \mathbf{w}_i \mathbf{x} + w_{i0}$$

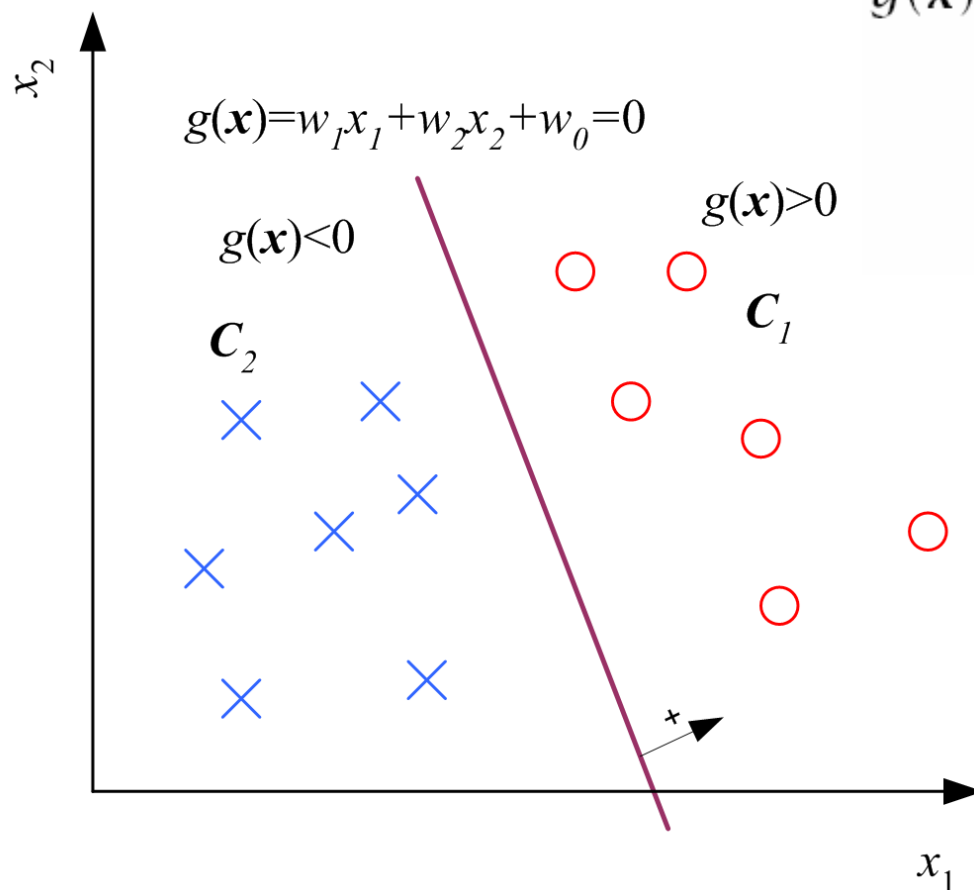
- Higher-order (product) terms:

$$Z_1 = x_1, Z_2 = x_2, Z_3 = x_1^2, Z_4 = x_2^2, Z_5 = x_1 x_2$$

Map from $\mathbf{x}$ to $\mathbf{z}$ using **nonlinear basis functions**
and use a linear discriminant in $\mathbf{z}$ -space

$$g_i(\mathbf{x}) = \sum_{j=1}^k w_j \phi_{ij}(\mathbf{x})$$

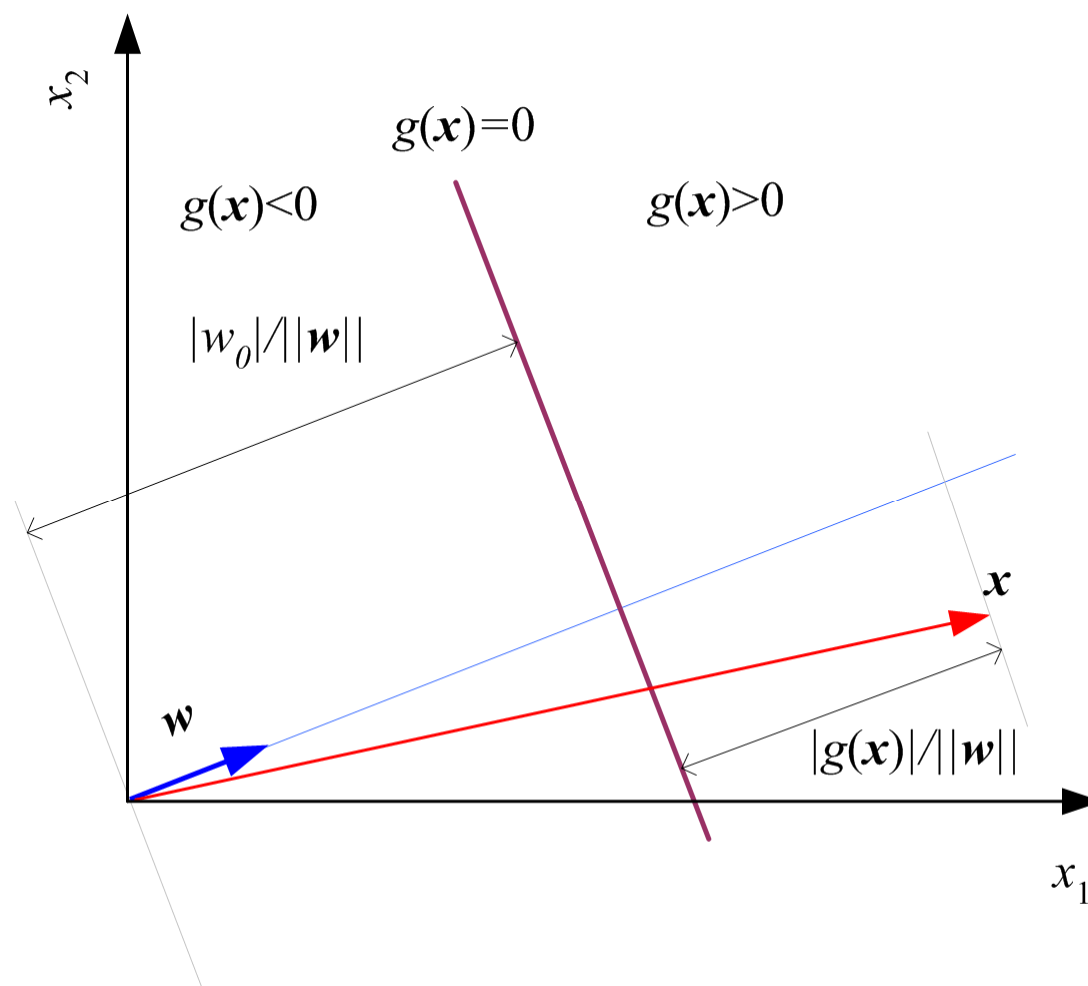
Two Classes



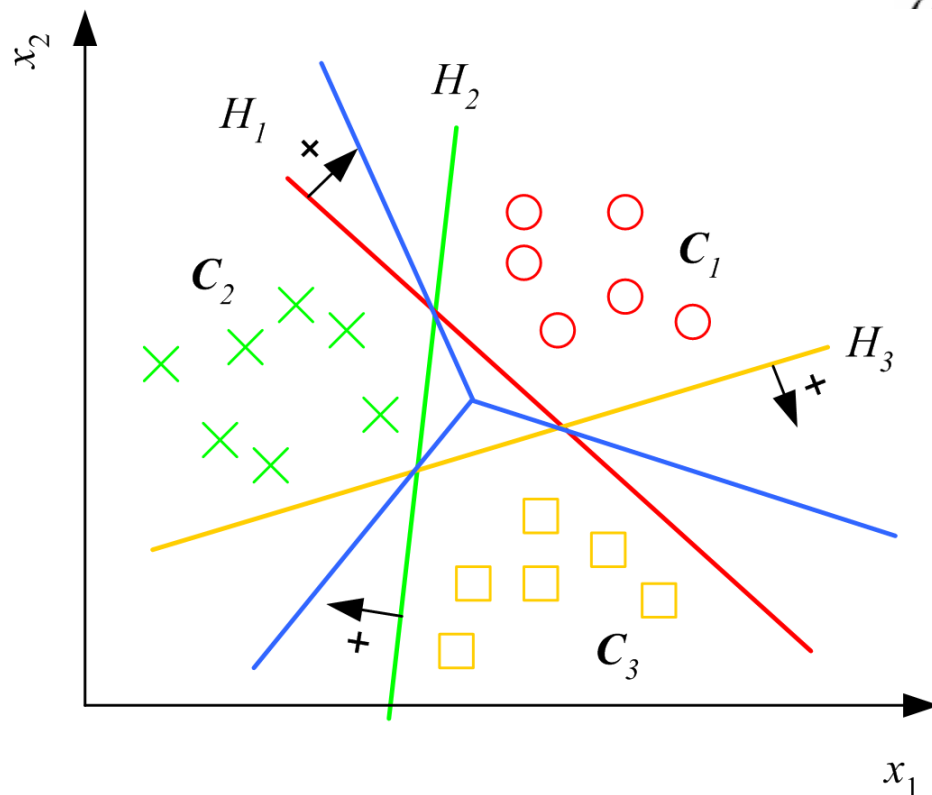
$$\begin{aligned} g(\mathbf{x}) &= g_1(\mathbf{x}) - g_2(\mathbf{x}) \\ &= (\mathbf{w}_1^T \mathbf{x} + w_{10}) - (\mathbf{w}_2^T \mathbf{x} + w_{20}) \\ &= (\mathbf{w}_1 - \mathbf{w}_2)^T \mathbf{x} + (w_{10} - w_{20}) \\ &= \mathbf{w}^T \mathbf{x} + w_0 \end{aligned}$$

$$\text{choose } \begin{cases} C_1 & \text{if } g(\mathbf{x}) > 0 \\ C_2 & \text{otherwise} \end{cases}$$

Geometry



Multiple Classes



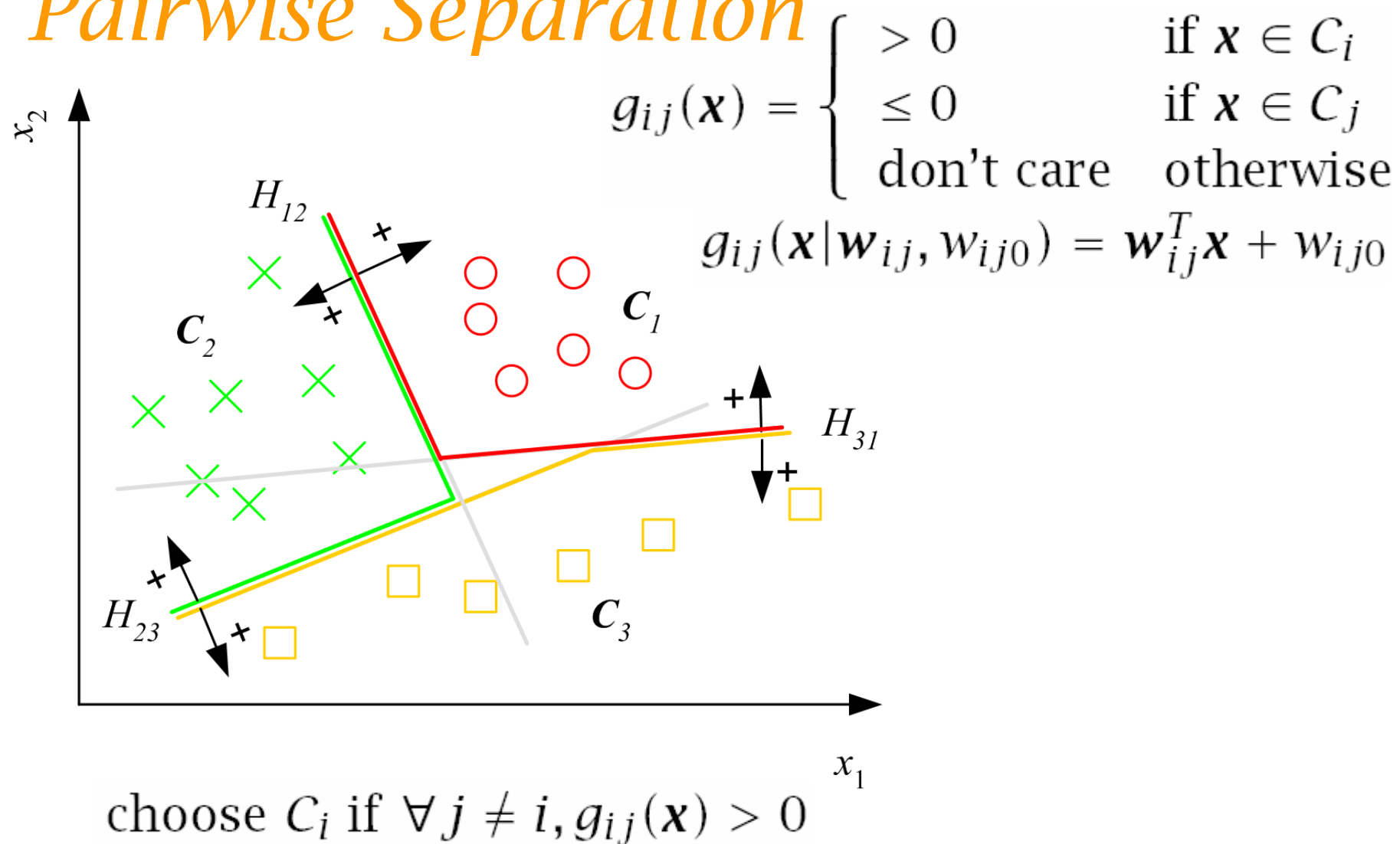
$$\pi_i(\mathbf{x} | \mathbf{w}_i, w_{i0}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

Choose C_i if

$$g_i(\mathbf{x}) = \max_{j=1}^K g_j(\mathbf{x})$$

Classes are
linearly separable

Pairwise Separation





From Discriminants to Posteriors

When $p(\mathbf{x} | C_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

$$\mathbf{w}_i = \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i$$

$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i + \log P(C_i)$$

$y \equiv P(C_1 | \mathbf{x})$ and $p(C_2 | \mathbf{x}) = 1 - y$

$$\text{choose } C_1 \text{ if } \begin{cases} y > 0.5 \\ \frac{y}{1-y} > 1 \\ \log \frac{y}{1-y} > 0 \end{cases} \quad \text{and } C_2 \text{ otherwise}$$



$$\begin{aligned}
 \text{logit}(P(C_1|\mathbf{x})) &= \log \frac{P(C_1|\mathbf{x})}{1 - P(C_1|\mathbf{x})} = \log \frac{P(C_1|\mathbf{x})}{P(C_2|\mathbf{x})} \\
 &= \log \frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)} + \log \frac{P(C_1)}{P(C_2)} \\
 &= \log \frac{(2\pi)^{-d/2} |\Sigma|^{-1/2} \exp[-(1/2)(\mathbf{x} - \boldsymbol{\mu}_1)^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_1)]}{(2\pi)^{-d/2} |\Sigma|^{-1/2} \exp[-(1/2)(\mathbf{x} - \boldsymbol{\mu}_2)^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_2)]} + \log \frac{P(C_1)}{P(C_2)} \\
 &= \mathbf{w}^T \mathbf{x} + w_0
 \end{aligned}$$

where

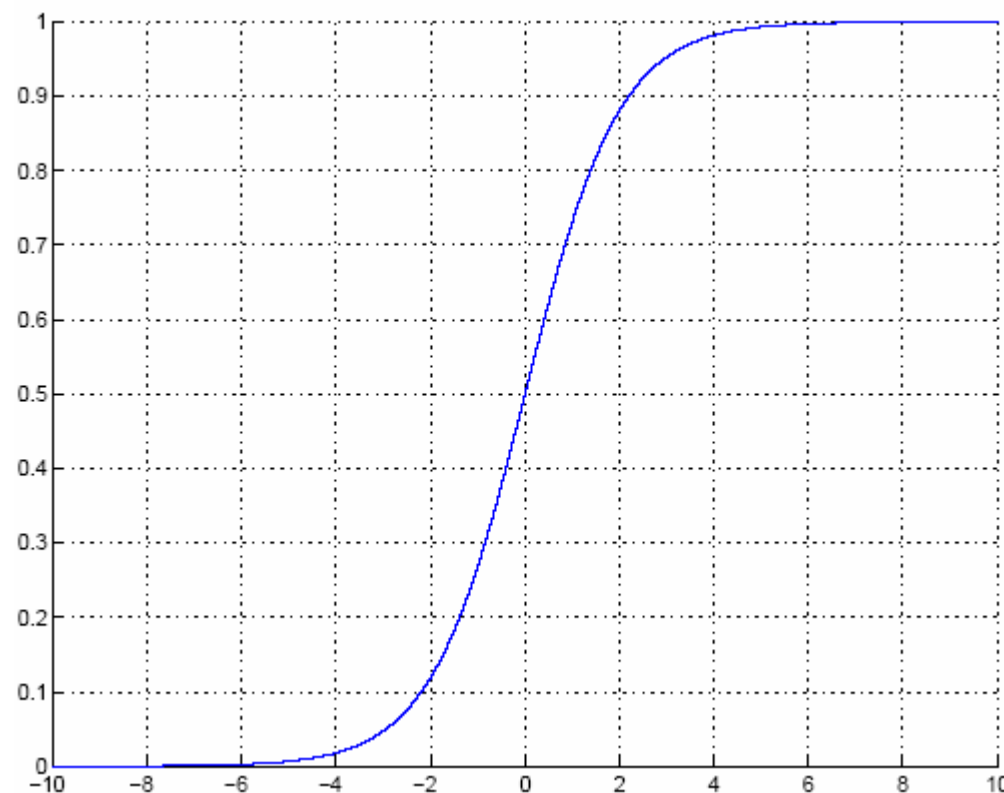
$$\begin{aligned}
 \mathbf{w} &= \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
 w_0 &= -\frac{1}{2} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \log \frac{P(C_1)}{P(C_2)}
 \end{aligned}$$

The inverse of logit

$$\log \frac{P(C_1|\mathbf{x})}{1 - P(C_1|\mathbf{x})} = \mathbf{w}^T \mathbf{x} + w_0$$

$$P(C_1|\mathbf{x}) = \text{sigmoid}(\mathbf{w}^T \mathbf{x} + w_0) = \frac{1}{1 + \exp[-(\mathbf{w}^T \mathbf{x} + w_0)]}$$

Sigmoid (Logistic) Function



1. calculate $g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0$ and choose C_1 if $g(\mathbf{x}) > 0$, or
2. calculate $y = \text{sigmoid}(\mathbf{w}^T \mathbf{x} + w_0)$ and choose C_1 if $y > 0.5$



Gradient-Descent

- $E(\mathbf{w}|\mathcal{X})$ is error with parameters $\mathbf{w}$ on sample $\mathcal{X}$

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} E(\mathbf{w}|\mathcal{X})$$

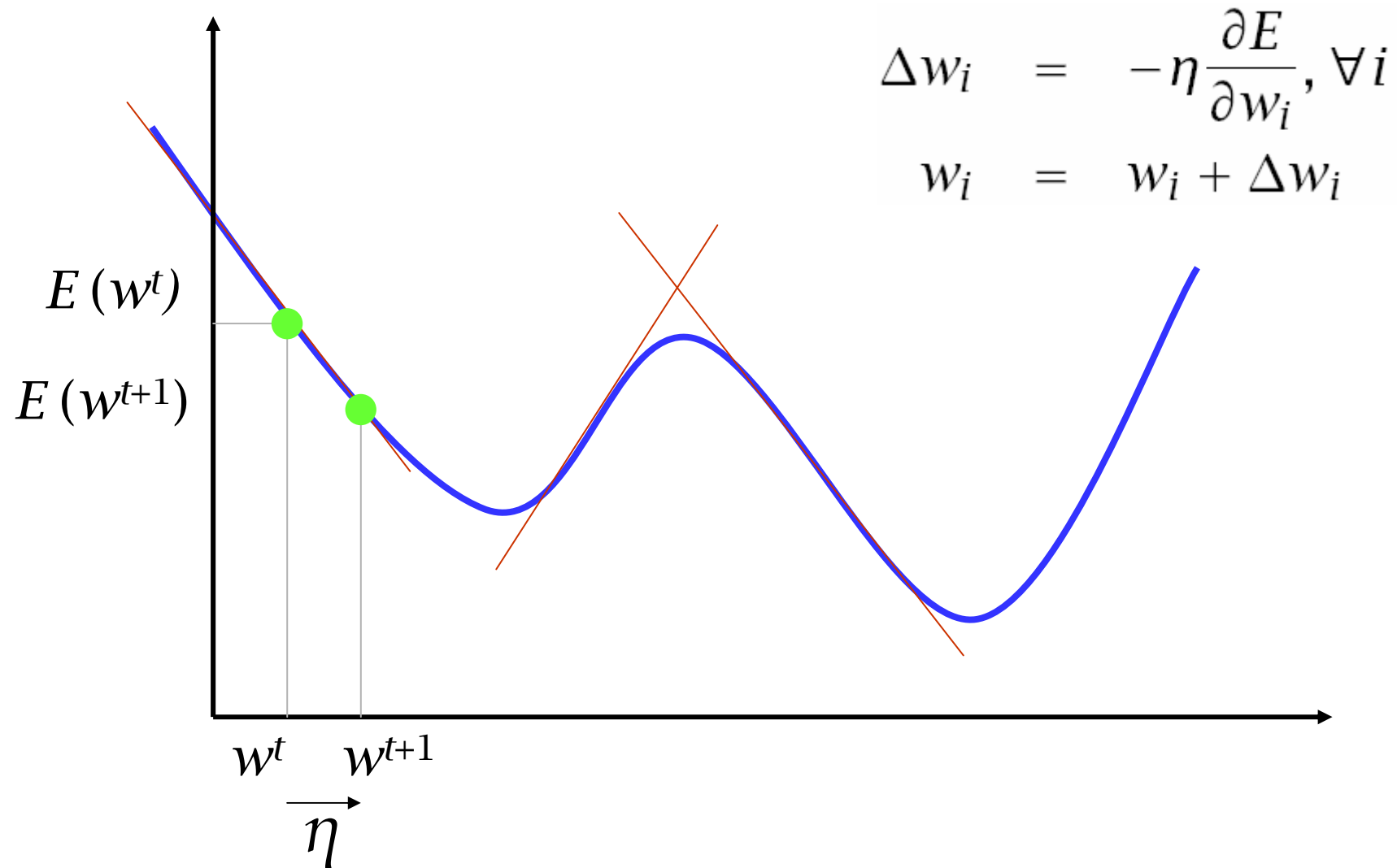
- Gradient

$$\nabla_{\mathbf{w}} E = \left[\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \dots, \frac{\partial E}{\partial w_d} \right]^T$$

- Gradient-descent:

Starts from random $\mathbf{w}$ and updates $\mathbf{w}$ iteratively in the negative direction of gradient

Gradient-Descent



Logistic Discrimination

- Two classes: Assume log likelihood ratio is linear

$$\log \frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)} = \mathbf{w}^T \mathbf{x} + w_0^o$$

$$\begin{aligned} \text{logit}(P(C_1|\mathbf{x})) &= \log \frac{P(C_1|\mathbf{x})}{1 - P(C_1|\mathbf{x})} \\ &= \log \frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)} + \log \frac{P(C_1)}{P(C_2)} \\ &= \mathbf{w}^T \mathbf{x} + w_0 \quad w_0 = w_0^o + \log \frac{P(C_1)}{P(C_2)} \end{aligned}$$

$$y = \hat{P}(C_1|\mathbf{x}) = \frac{1}{1 + \exp[-(\mathbf{w}^T \mathbf{x} + w_0)]}$$



Training: Two Classes

$$\mathcal{X} = \{\mathbf{x}^t, r^t\} \quad r^t | \mathbf{x}^t \sim \text{Bernoulli}(y^t)$$

$$y = \hat{P}(C_1 | \mathbf{x}) = \frac{1}{1 + \exp[-(\mathbf{w}^T \mathbf{x} + w_0)]}$$

$$l(\mathbf{w}, w_0 | \mathcal{X}) = \prod_t (y^t)^{(r^t)} (1 - y^t)^{(1-r^t)}$$

$$E = -\log l$$

$$E(\mathbf{w}, w_0 | \mathcal{X}) = - \sum_t r^t \log y^t + (1 - r^t) \log(1 - y^t)$$



Training: Gradient-Descent

$$E(\mathbf{w}, w_0 | \mathcal{X}) = - \sum_t r^t \log y^t + (1 - r^t) \log(1 - y^t)$$

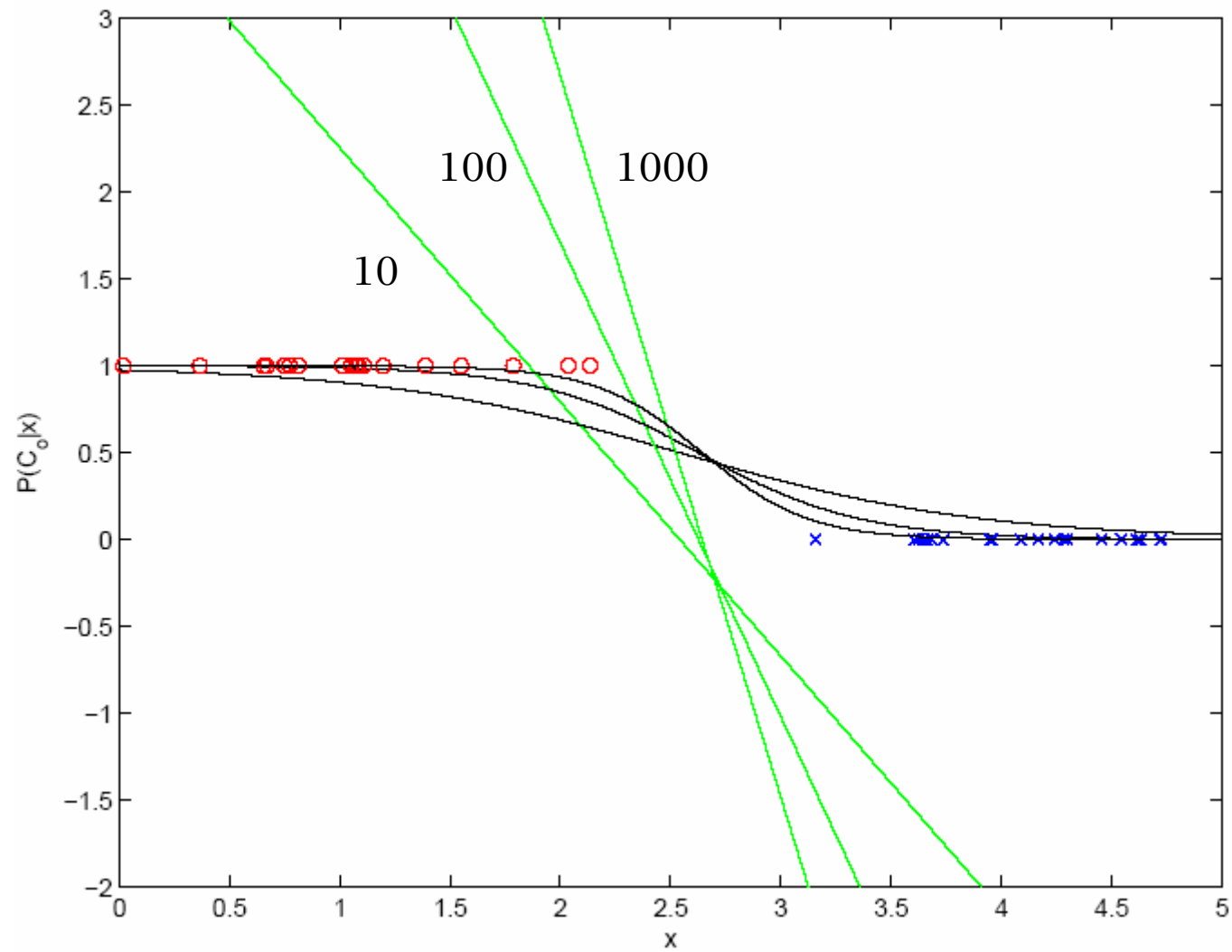
$$\text{If } y = \text{sigmoid}(a) \quad \frac{dy}{da} = y(1 - y)$$

$$\begin{aligned} \Delta w_j &= -\eta \frac{\partial E}{\partial w_j} = \eta \sum_t \left(\frac{r^t}{y^t} - \frac{1 - r^t}{1 - y^t} \right) y^t (1 - y^t) x_j^t \\ &= \eta \sum_t (r^t - y^t) x_j^t, j = 1, \dots, d \end{aligned}$$

$$\Delta w_0 = -\eta \frac{\partial E}{\partial w_0} = \eta \sum_t (r^t - y^t)$$



```
For  $j = 0, \dots, d$   
     $w_j \leftarrow \text{rand}(-0.01, 0.01)$   
Repeat  
    For  $j = 0, \dots, d$   
         $\Delta w_j \leftarrow 0$   
    For  $t = 1, \dots, N$   
         $o \leftarrow 0$   
        For  $j = 0, \dots, d$   
             $o \leftarrow o + w_j x_j^t$   
         $y \leftarrow \text{sigmoid}(o)$   
         $\Delta w_j \leftarrow \Delta w_j + (r^t - y)x_j^t$   
    For  $j = 0, \dots, d$   
         $w_j \leftarrow w_j + \eta \Delta w_j$   
Until convergence
```



K > 2 Classes

$$y_i = \hat{P}(C_i|\mathbf{x}) = \frac{\exp[\mathbf{w}_i^T \mathbf{x} + w_{i0}]}{\sum_{j=1}^K \exp[\mathbf{w}_j^T \mathbf{x} + w_{j0}]}, \quad i = 1, \dots, K$$

softmax

$$\mathbf{r}^t | \mathbf{x}^t \sim \text{Mult}_K(1, \mathbf{y}^t)$$

$$l(\{\mathbf{w}_i, w_{i0}\}_i | \mathcal{X}) = \prod_t \prod_i (y_i^t)^{r_i^t}$$

$$E(\{\mathbf{w}_i, w_{i0}\}_i | \mathcal{X}) = - \sum_t \sum_i r_i^t \log y_i^t$$

$$\Delta \mathbf{w}_j = \eta \sum_t (r_j^t - y_j^t) \mathbf{x}^t$$

$$\Delta w_{j0} = \eta \sum_t (r_j^t - y_j^t)$$

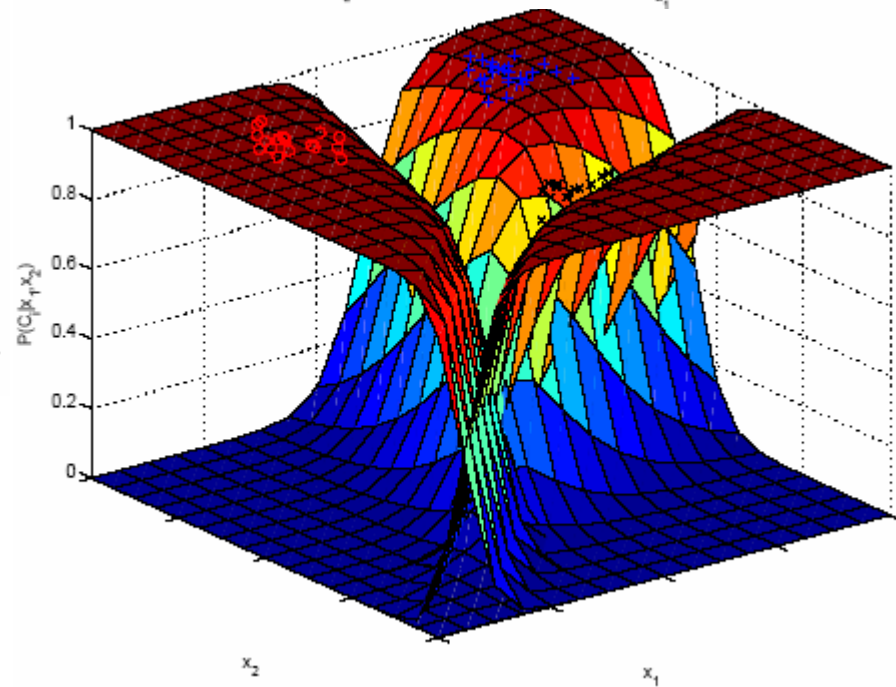
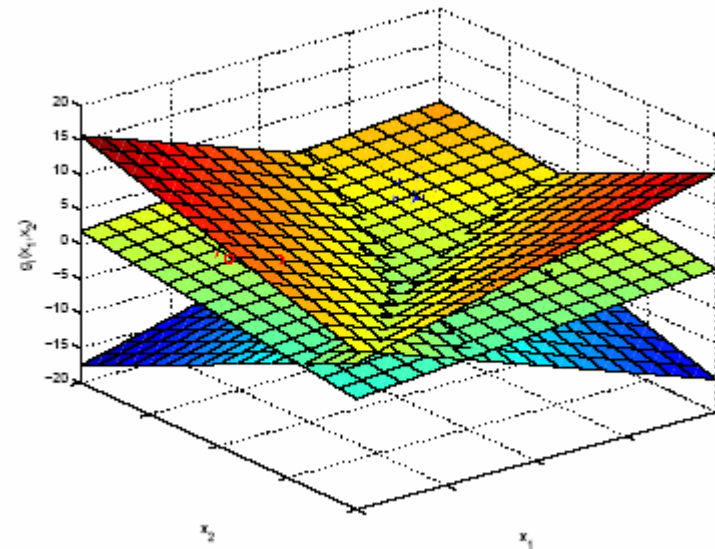
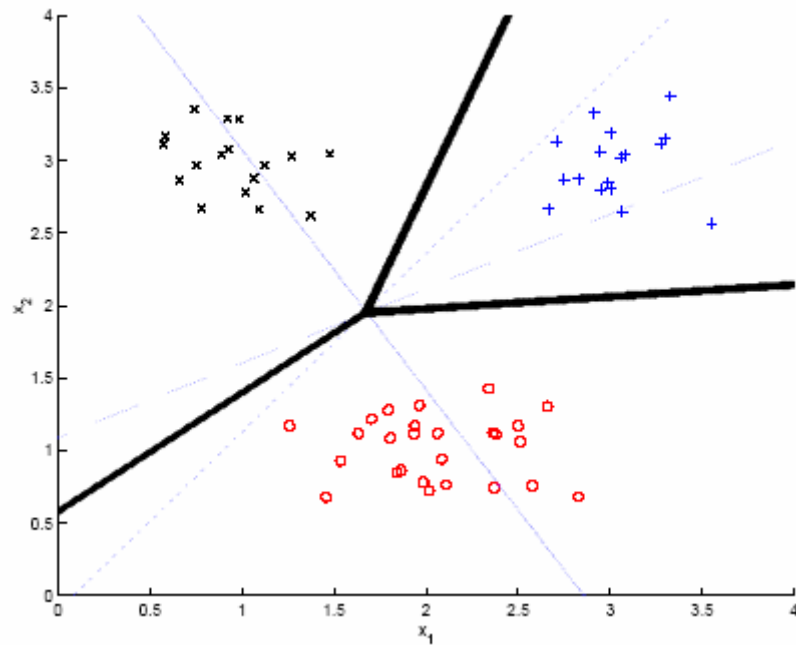


```

For  $i = 1, \dots, K$ , For  $j = 0, \dots, d$ ,  $w_{ij} \leftarrow \text{rand}(-0.01, 0.01)$ 
Repeat
  For  $i = 1, \dots, K$ , For  $j = 0, \dots, d$ ,  $\Delta w_{ij} \leftarrow 0$ 
  For  $t = 1, \dots, N$ 
    For  $i = 1, \dots, K$ 
       $o_i \leftarrow 0$ 
      For  $j = 0, \dots, d$ 
         $o_i \leftarrow o_i + w_{ij} x_j^t$ 
      For  $i = 1, \dots, K$ 
         $y_i \leftarrow \exp(o_i) / \sum_k \exp(o_k)$ 
      For  $i = 1, \dots, K$ 
        For  $j = 0, \dots, d$ 
           $\Delta w_{ij} \leftarrow \Delta w_{ij} + (r_i^t - y_i) x_j^t$ 
    For  $i = 1, \dots, K$ 
      For  $j = 0, \dots, d$ 
         $w_{ij} \leftarrow w_{ij} + \eta \Delta w_{ij}$ 
  Until convergence

```

Example





Generalizing the Linear Model

- Quadratic:

$$\log \frac{p(\mathbf{x}|C_i)}{p(\mathbf{x}|C_k)} = \mathbf{x}^T \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

- Sum of basis functions:

$$\log \frac{p(\mathbf{x}|C_i)}{p(\mathbf{x}|C_k)} = \mathbf{w}_i^T \boldsymbol{\phi}(\mathbf{x}) + w_{i0}$$

where $\boldsymbol{\phi}(\mathbf{x})$ are basis functions

- Kernels in SVM
- Hidden units in neural networks



Discrimination by Regression

- Classes are NOT mutually exclusive and exhaustive

$$r^t = y^t + \epsilon \quad \text{where } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$y^t = \text{sigmoid}(\mathbf{w}^T \mathbf{x}^t + w_0) = \frac{1}{1 + \exp[-(\mathbf{w}^T \mathbf{x}^t + w_0)]}$$

$$l(\mathbf{w}, w_0 | \mathcal{X}) = \prod_t \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(r^t - y^t)^2}{2\sigma^2}\right]$$

$$E(\mathbf{w}, w_0 | \mathcal{X}) = \frac{1}{2} \sum_t (r^t - y^t)^2$$

$$\Delta \mathbf{w} = \eta \sum_t (r^t - y^t) y^t (1 - y^t) \mathbf{x}^t$$



Optimal Separating Hyperplane

$\mathcal{X} = \{\mathbf{x}^t, r^t\}$ where $r^t = +1$ if $\mathbf{x}^t \in C_1$
 $r^t = -1$ if $\mathbf{x}^t \in C_2$

find $\mathbf{w}$ and w_0 such that

$$\mathbf{w}^T \mathbf{x}^t + w_0 \geq +1 \quad \text{for} \quad r^t = +1$$

$$\mathbf{w}^T \mathbf{x}^t + w_0 \leq -1 \quad \text{for} \quad r^t = -1$$

which can be rewritten as

$$r^t (\mathbf{w}^T \mathbf{x}^t + w_0) \geq +1$$

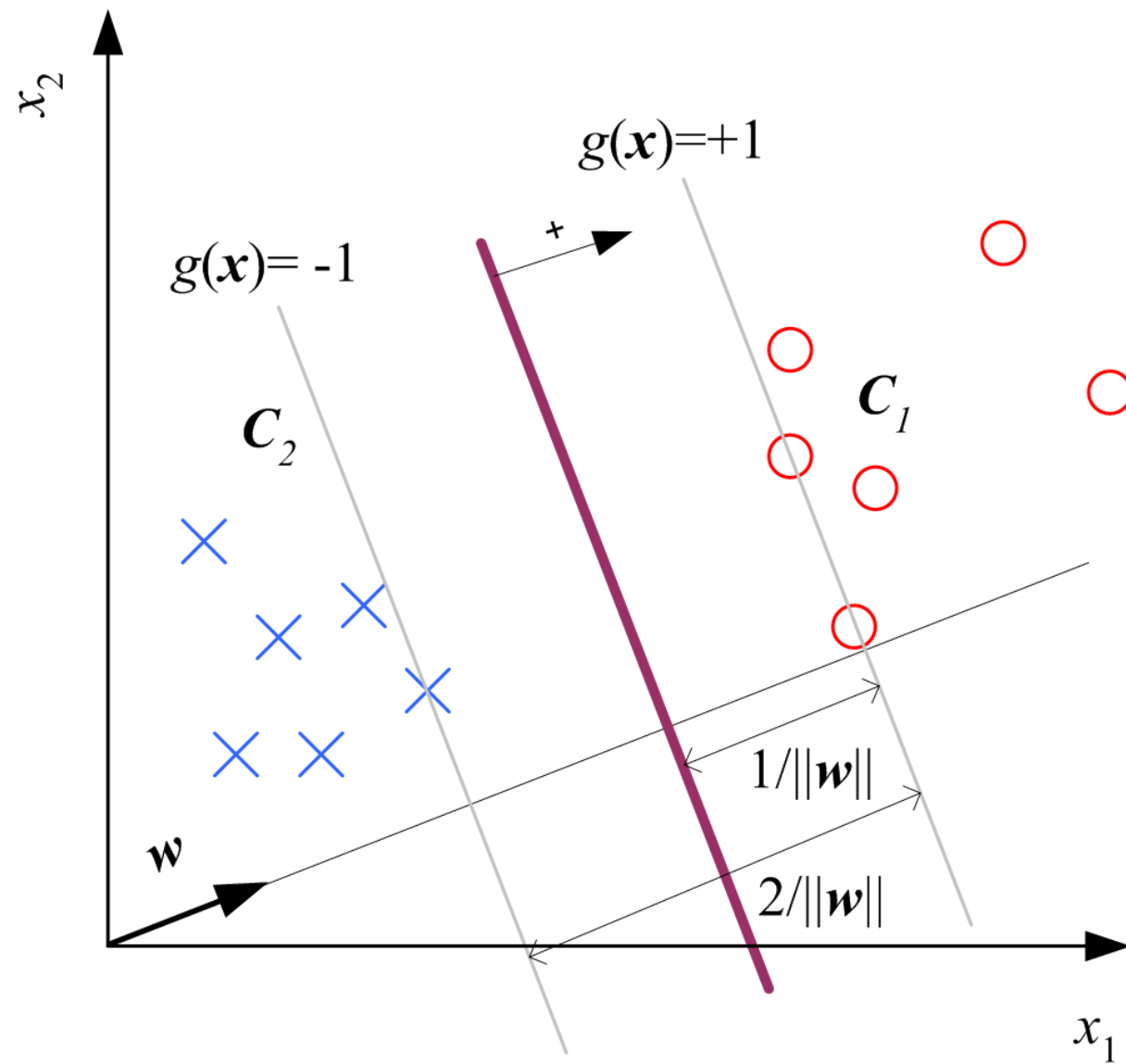
(Cortes and Vapnik, 1995; Vapnik, 1995)



Margin

- Distance from the discriminant to the closest instances on either side
- Distance of $\mathbf{x}$ to the hyperplane is $\frac{|\mathbf{w}^T \mathbf{x}^t + w_0|}{\|\mathbf{w}\|}$
- We require $\frac{r^t (\mathbf{w}^T \mathbf{x}^t + w_0)}{\|\mathbf{w}\|} \geq \rho, \forall t$
- For a unique sol'n, fix $\rho \|\mathbf{w}\| = 1$ and to max margin

$$\min \frac{1}{2} \|\mathbf{w}\|^2 \text{ subject to } r^t (\mathbf{w}^T \mathbf{x}^t + w_0) \geq +1, \forall t$$




$$\min \frac{1}{2} \|\mathbf{w}\|^2 \text{ subject to } r^t(\mathbf{w}^T \mathbf{x}^t + w_0) \geq +1, \forall t$$

$$\begin{aligned} L_p &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{t=1}^N \alpha^t [r^t(\mathbf{w}^T \mathbf{x}^t + w_0) - 1] \\ &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_t \alpha^t r^t(\mathbf{w}^T \mathbf{x}^t + w_0) + \sum_t \alpha^t \end{aligned}$$

$$\frac{\partial L_p}{\partial \mathbf{w}} = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_t \alpha^t r^t \mathbf{x}^t$$

$$\frac{\partial L_p}{\partial w_0} = 0 \quad \Rightarrow \quad \sum_t \alpha^t r^t = 0$$



$$\begin{aligned}
 L_d &= \frac{1}{2}(\mathbf{w}^T \mathbf{w}) - \mathbf{w}^T \sum_t \alpha^t r^t \mathbf{x}^t - w_0 \sum_t \alpha^t r^t + \sum_t \alpha^t \\
 &= -\frac{1}{2}(\mathbf{w}^T \mathbf{w}) + \sum_t \alpha^t \\
 &= -\frac{1}{2} \sum_t \sum_s \alpha^t \alpha^s r^t r^s (\mathbf{x}^t)^T \mathbf{x}^s + \sum_t \alpha^t
 \end{aligned}$$

subject to $\sum_t \alpha^t r^t = 0$, and $\alpha^t \geq 0, \forall t$

Most α_t are 0 and only a small number have $\alpha_t > 0$;
they are the **support vectors**

Soft Margin Hyperplane

- Not linearly separable

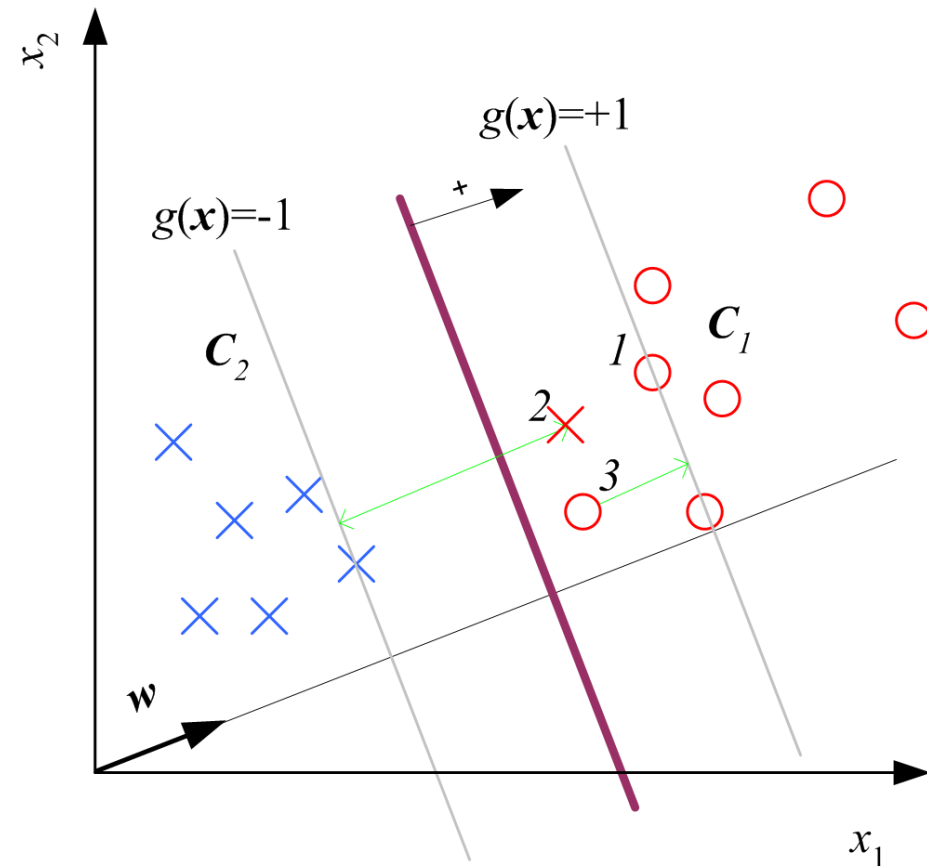
$$r^t(\mathbf{w}^T \mathbf{x}^t + w_0) \geq 1 - \xi^t$$

- Soft error

$$\sum_t \xi^t$$

- New primal is

$$L_p = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_t \xi^t - \sum_t \alpha^t [r^t(\mathbf{w}^T \mathbf{x}^t + w_0) - 1 + \xi^t] - \sum_t \mu^t \xi^t$$



Kernel Machines

- Preprocess input $\mathbf{x}$ by basis functions

$$\mathbf{z} = \boldsymbol{\phi}(\mathbf{x}) \quad g(\mathbf{z}) = \mathbf{w}^T \mathbf{z}$$

- The SVM solution $g(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$

$$\mathbf{w} = \sum_t \alpha^t r^t \mathbf{z}^t = \sum_t \alpha^t r^t \boldsymbol{\phi}(\mathbf{x}^t)$$

$$g(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) = \sum_t \alpha^t r^t \boxed{\boldsymbol{\phi}(\mathbf{x}^t)^T \boldsymbol{\phi}(\mathbf{x})}$$

$$g(\mathbf{x}) = \sum_t \alpha^t r^t \boxed{K(\mathbf{x}^t, \mathbf{x})}$$

Kernel Functions

- Polynomials of degree q : $K(\mathbf{x}^t, \mathbf{x}) = (\mathbf{x}^T \mathbf{x}^t + 1)^q$

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + 1)^2$$

$$= (x_1 y_1 + x_2 y_2 + 1)^2$$

$$= 1 + 2x_1 y_1 + 2x_2 y_2 + 2x_1 x_2 y_1 y_2 + x_1^2 y_1^2 + x_2^2 y_2^2$$

$$\phi(\mathbf{x}) = [1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2, x_1^2, x_2^2]^T$$

- Radial-basis functions: $K(\mathbf{x}^t, \mathbf{x}) = \exp \left[-\frac{\|\mathbf{x}^t - \mathbf{x}\|^2}{\sigma^2} \right]$
- Sigmoidal functions:

$$K(\mathbf{x}^t, \mathbf{x}) = \tanh(2\mathbf{x}^T \mathbf{x}^t + 1)$$

(Cherkassky and Mulier, 1998)

SVM for Regression

- Use a linear model (possibly kernelized)

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

- Use the ϵ -sensitive error function

$$e_{\epsilon}(r^t, f(\mathbf{x}^t)) = \begin{cases} 0 & \text{if } |r^t - f(\mathbf{x}^t)| < \epsilon \\ |r^t - f(\mathbf{x}^t)| - \epsilon & \text{otherwise} \end{cases}$$

- $$\min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_t (\xi_+^t + \xi_-^t)$$
$$r^t - (\mathbf{w}^T \mathbf{x} + w_0) \leq \epsilon + \xi_+^t$$
$$(\mathbf{w}^T \mathbf{x} + w_0) - r^t \leq \epsilon + \xi_-^t$$
$$\xi_+^t, \xi_-^t \geq 0$$

