

Lecture Slides for

INTRODUCTION TO

Machine Learning

2nd Edition

ETHEM ALPAYDIN

© The MIT Press, 2010

alpaydin@boun.edu.tr

<http://www.cmpe.boun.edu.tr/~ethem/i2ml2e>

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:

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

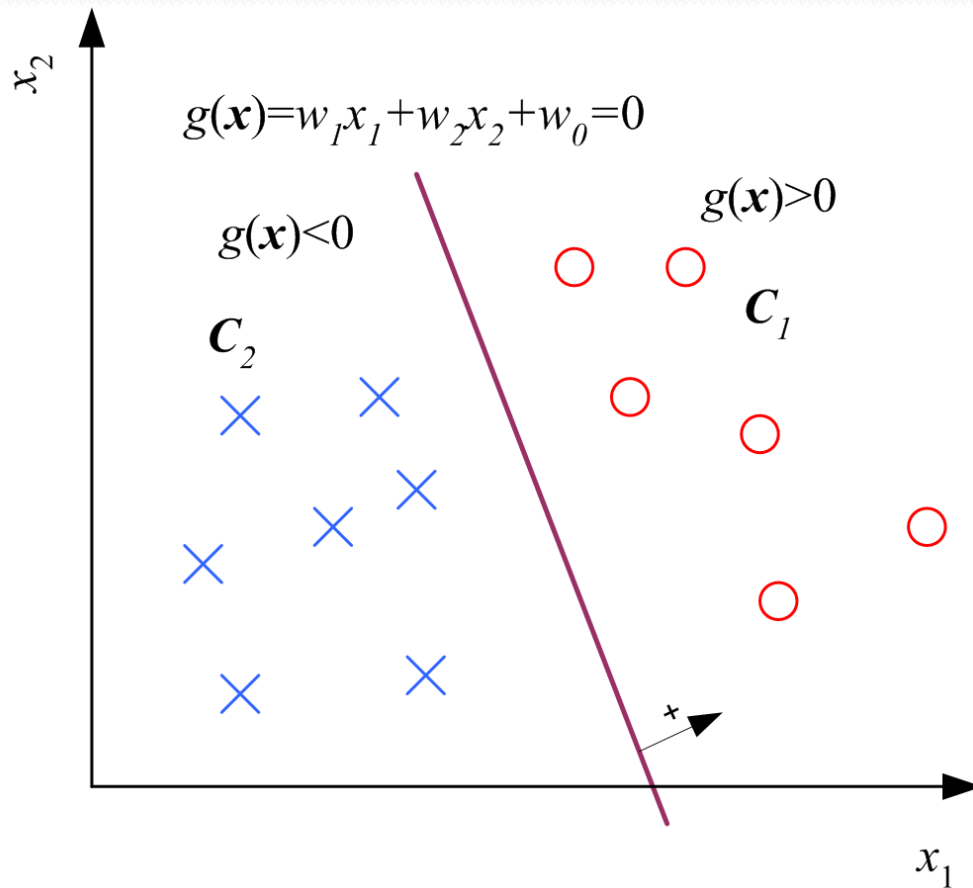
- Higher-order (product) terms:

$$z_1 = x_1, z_2 = x_2, z_3 = x_1^2, z_4 = x_2^2, z_5 = x_1 x_2$$

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

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

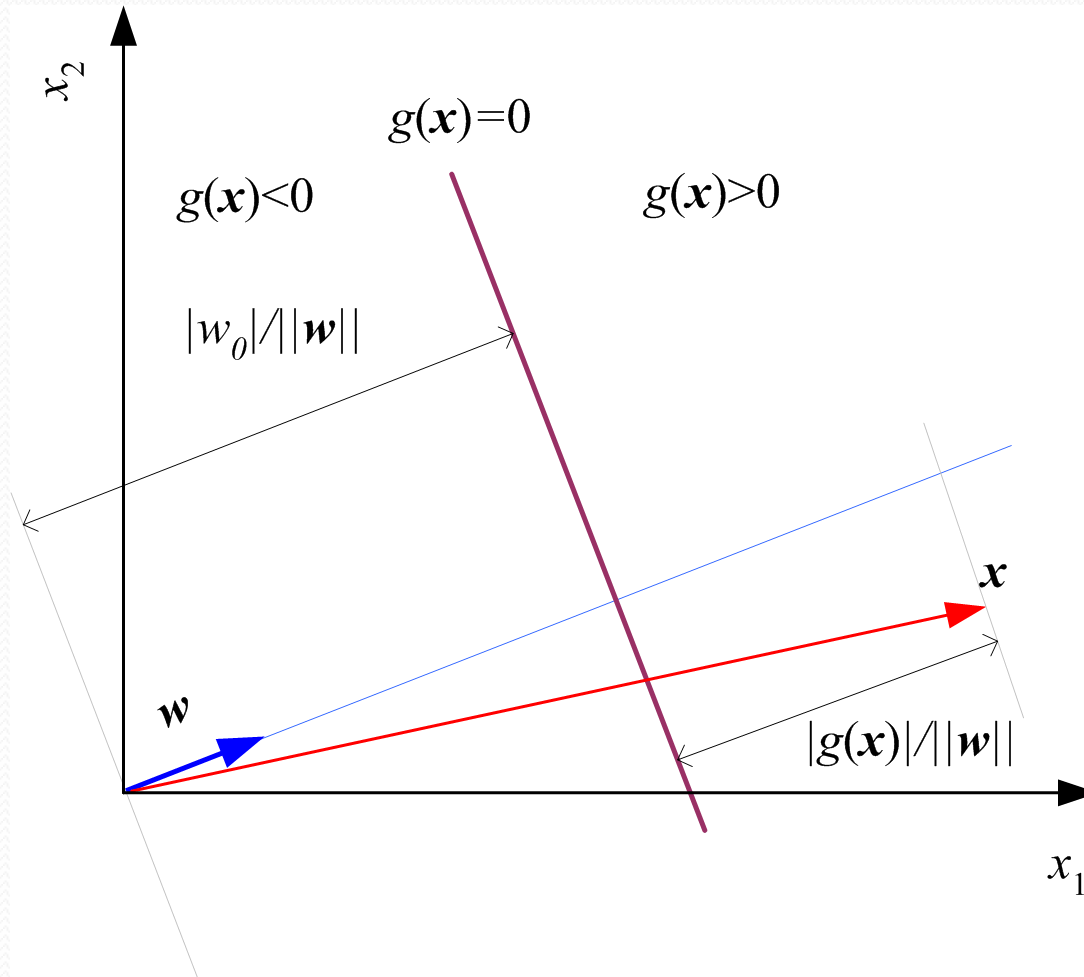
Two Classes



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

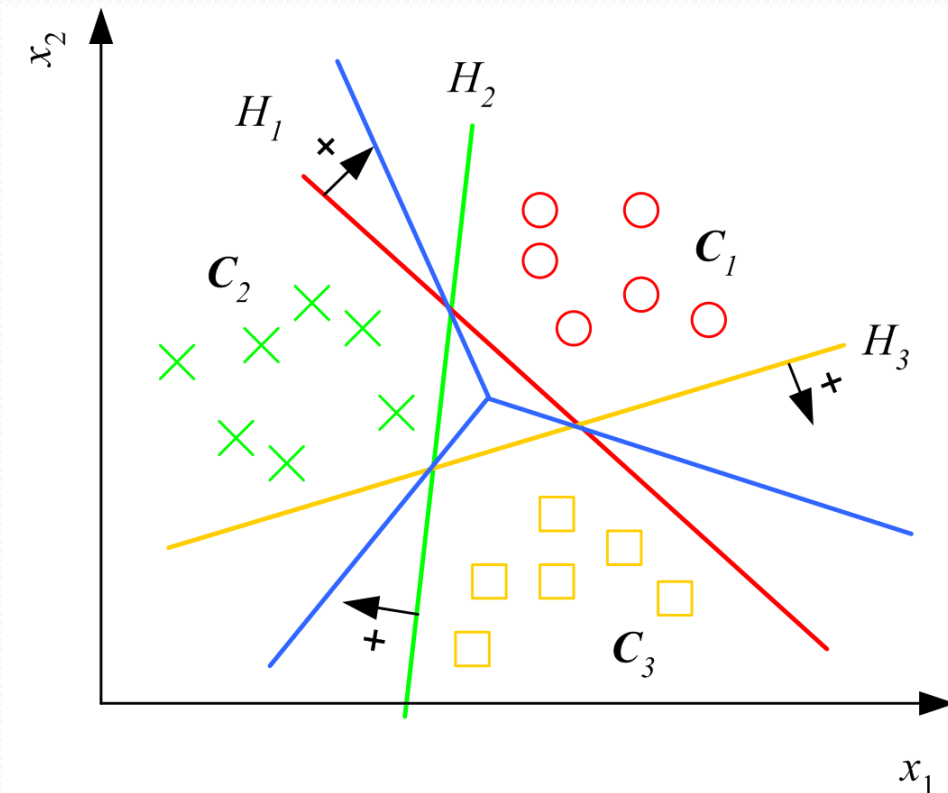
$$\text{choose} \begin{cases} C_1 & \text{if } g(\mathbf{x}) > 0 \\ C_2 & \text{otherwise} \end{cases}$$

Geometry



Multiple Classes

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

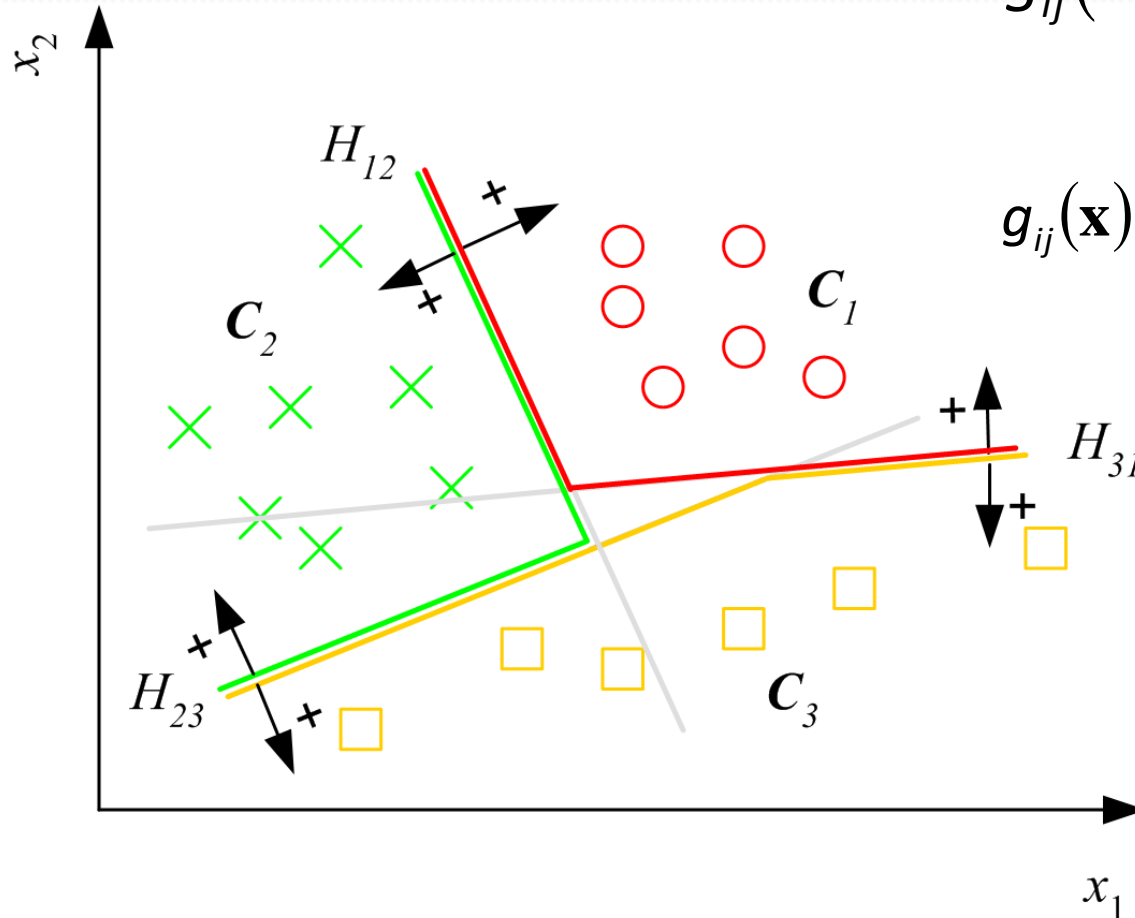


Choose C_i if

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

Classes are
linearly separable

Pairwise Separation



$$g_{ij}(\mathbf{x} | \mathbf{w}_{ij}, w_{ij0}) = \mathbf{w}_{ij}^T \mathbf{x} + w_{ij0}$$

$$g_{ij}(\mathbf{x}) = \begin{cases} > 0 & \text{if } \mathbf{x} \in C_i \\ \leq 0 & \text{if } \mathbf{x} \in C_j \\ \text{don't care} & \text{otherwise} \end{cases}$$

choose C_i if
 $\forall j \neq i, g_{ij}(\mathbf{x}) > 0$

From Discriminants to Posteriors

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

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

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

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

$$\text{choose } C_1 \text{ if } \begin{cases} y > 0.5 \\ y/(1-y) > 1 \\ \log[y/(1-y)] > 0 \end{cases} \text{ and } C_2 \text{ otherwise}$$

$$\begin{aligned}
\text{logit}(P(C_1 | \mathbf{x})) &= \log \frac{P(C_1 | \mathbf{x})}{1 - P(C_1 | \mathbf{x})} = \log \frac{P(C_1 | \mathbf{x})}{P(C_2 | \mathbf{x})} \\
&= \log \frac{p(\mathbf{x} | C_1)}{p(\mathbf{x} | C_2)} + \log \frac{P(C_1)}{P(C_2)} \\
&= \log \frac{(2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left[-(1/2)(\mathbf{x} - \mu_1)^T \Sigma^{-1} (\mathbf{x} - \mu_1)\right]}{(2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left[-(1/2)(\mathbf{x} - \mu_2)^T \Sigma^{-1} (\mathbf{x} - \mu_2)\right]} + \log \frac{P(C_1)}{P(C_2)} \\
&= \mathbf{w}^T \mathbf{x} + w_0
\end{aligned}$$

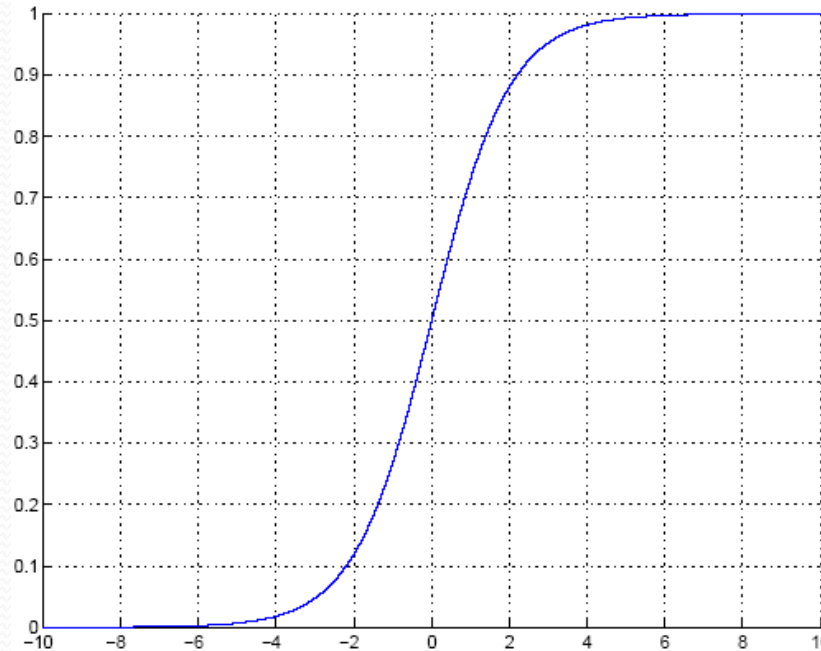
$$\text{where } \mathbf{w} = \Sigma^{-1}(\mu_1 - \mu_2) \quad w_0 = -\frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2)$$

The inverse of logit

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

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

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

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

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

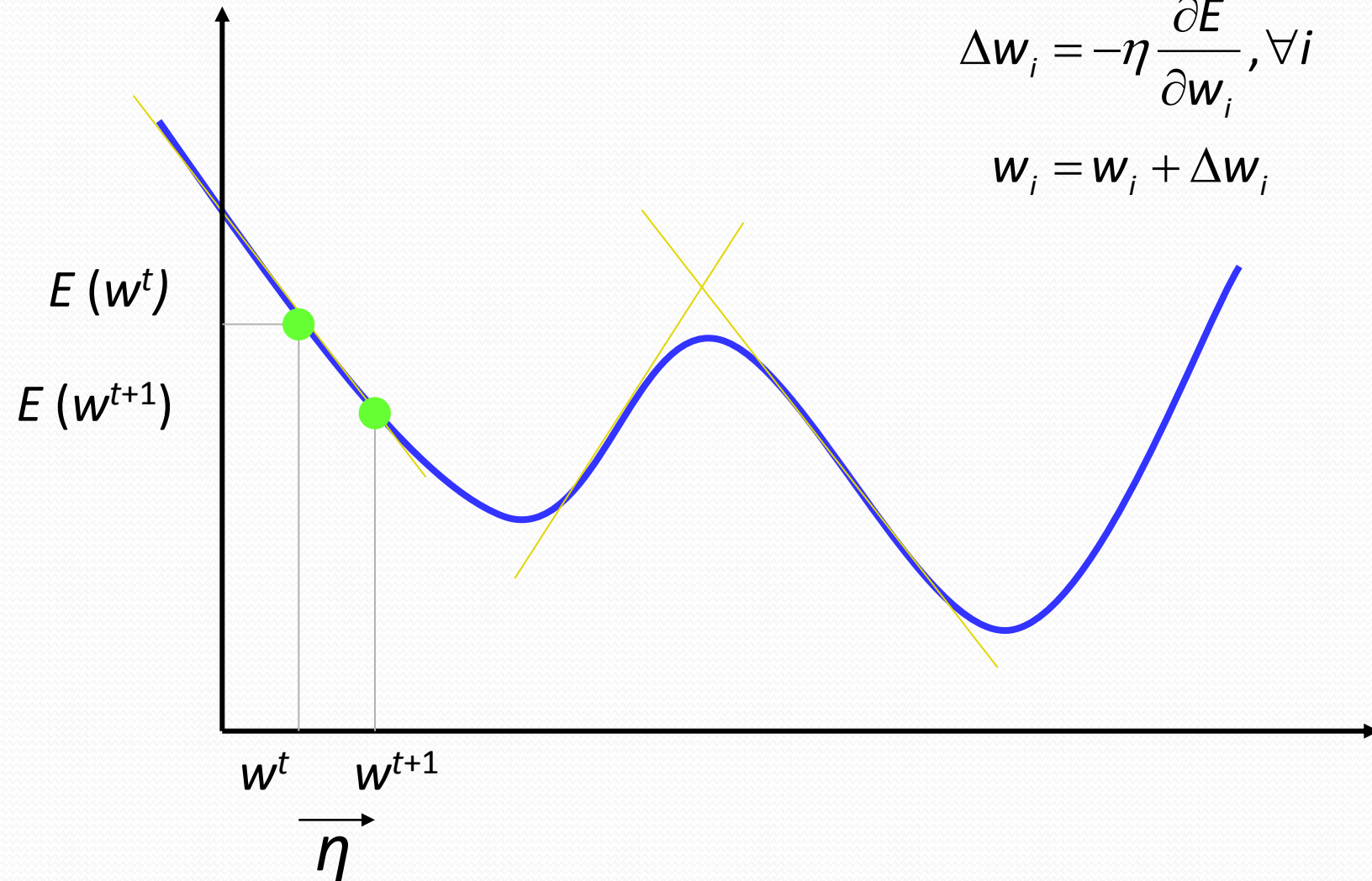
- Gradient
$$\nabla_{\mathbf{w}} E = \left[\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \dots, \frac{\partial E}{\partial w_d} \right]^T$$

- Gradient-descent:

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

Gradient-Descent

$$\Delta w_i = -\eta \frac{\partial E}{\partial w_i}, \forall i$$
$$w_i = w_i + \Delta w_i$$



Logistic Discrimination

- Two classes: Assume log likelihood ratio is linear

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

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

$$\text{where } w_0 = w_0^o + \log \frac{P(C_1)}{P(C_2)}$$

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

Training: Two Classes

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

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

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

$$E = -\log l$$

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

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 $j = 0, \dots, d$

$o \leftarrow o + w_j x_j^t$

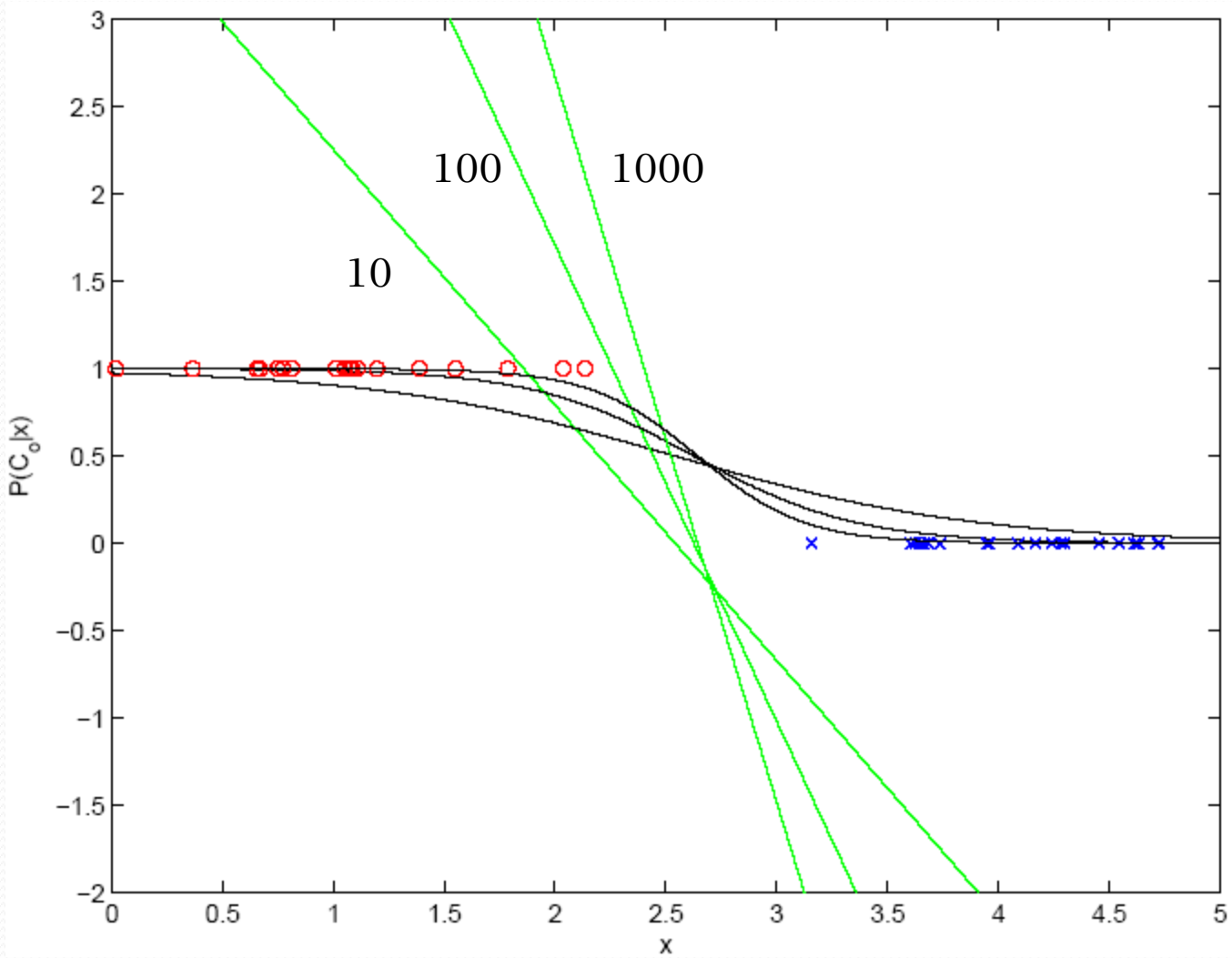
$y \leftarrow \text{sigmoid}(o)$

$\Delta w_j \leftarrow \Delta w_j + (r^t - y)x_j^t$

For $j = 0, \dots, d$

$w_j \leftarrow w_j + \eta \Delta w_j$

Until convergence



K > 2 Classes

$$\mathcal{X} = \{\mathbf{x}^t, \mathbf{r}^t\}_t \quad r^t | \mathbf{x}^t \sim \text{Mult}_K(\mathbf{1}, \mathbf{y}^t)$$

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

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

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

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

$$\Delta \mathbf{w}_j = \eta \sum_t (r_j^t - y_j^t) \mathbf{x}^t \quad \Delta w_{j0} = \eta \sum_t (r_j^t - y_j^t)$$

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 $j = 0, \dots, d$

$\Delta w_{ij} \leftarrow \Delta w_{ij} + (r_i^t - y_i)x_j^t$

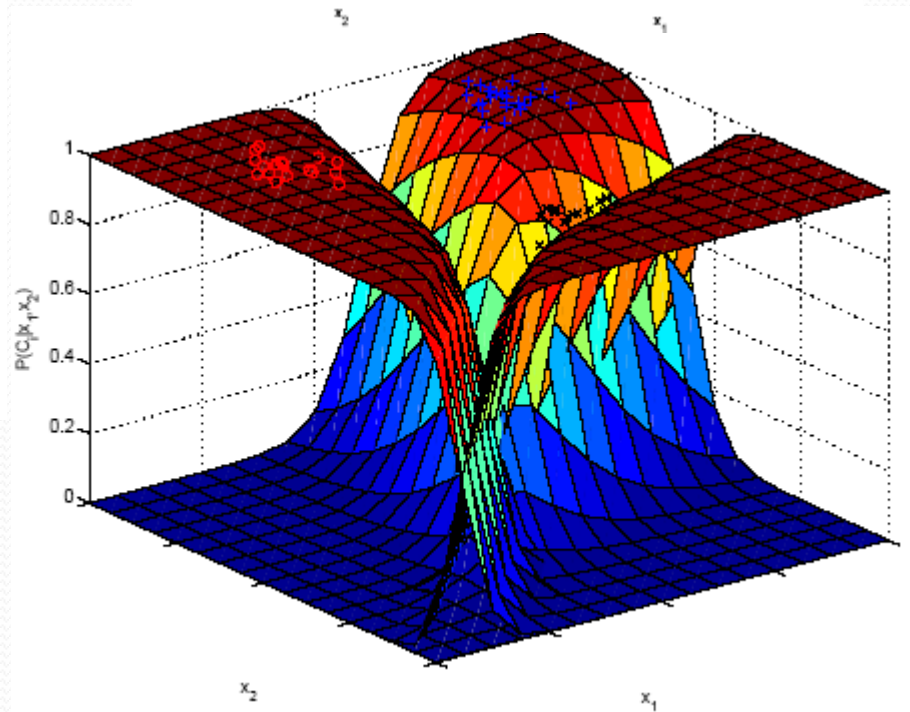
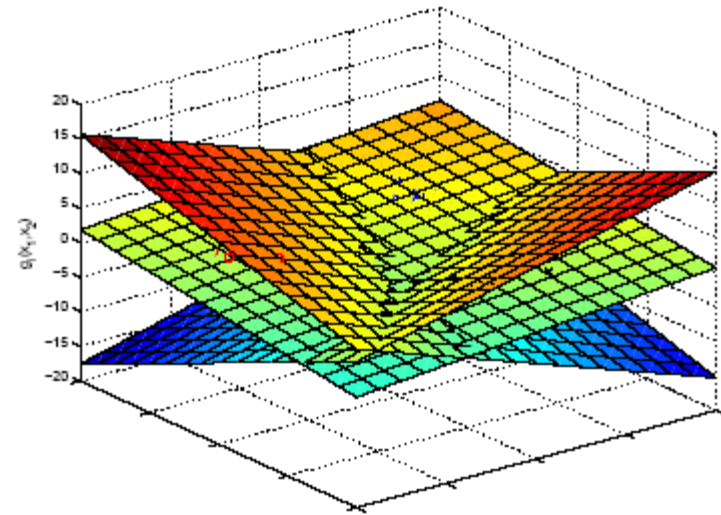
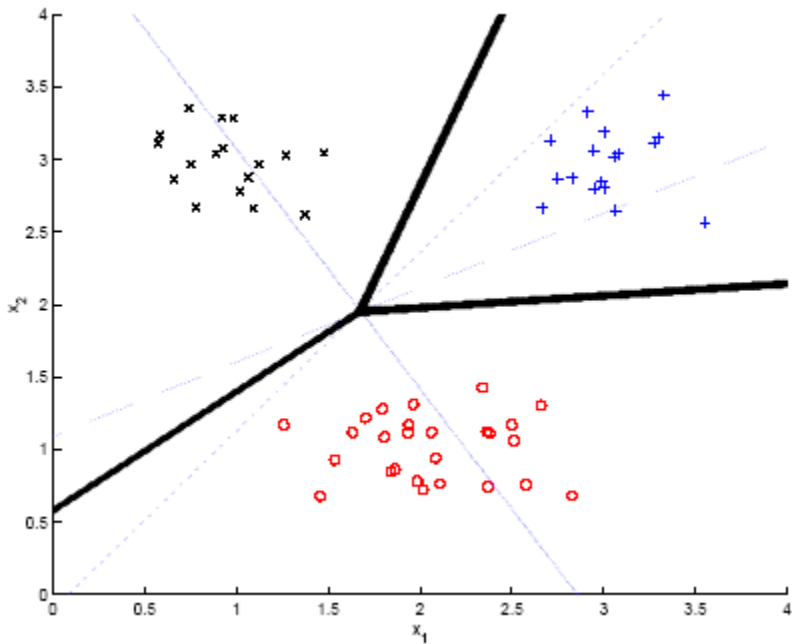
For $i = 1, \dots, K$

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:

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

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

- Hidden units in neural networks (Chapters 11 and 12)
- Kernels in SVM (Chapter 13)

Discrimination by Regression

- Classes are NOT mutually exclusive and exhaustive

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

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

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

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

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