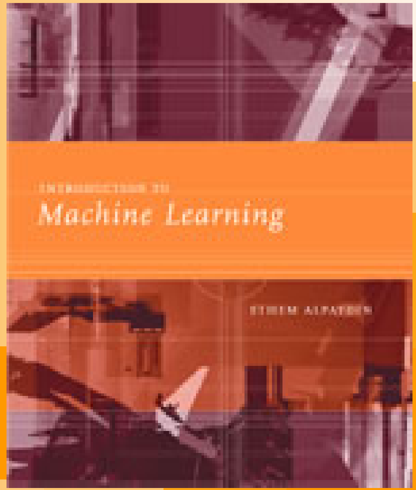


Lecture Slides for

INTRODUCTION TO

Machine Learning

ETHEM ALPAYDIN

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml

CHAPTER 10:

Linear Discrimination

Likelihood- vs. Discriminant-based Classification

- Likelihood-based: Assume a model for $p(\boldsymbol{x}|C_i)$, use Bayes' rule to calculate $P(C_i|\boldsymbol{x})$

  $$g_i(\boldsymbol{x}) = \log P(C_i|\boldsymbol{x})$$

- Discriminant-based: Assume a model for $g_i(\boldsymbol{x}|\Phi_i)$; no density estimation
- Estimating the boundaries is enough; no need to accurately estimate the densities inside the boundaries

Linear Discriminant

- Linear discriminant:

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

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

Generalized Linear Model

- Quadratic discriminant:

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

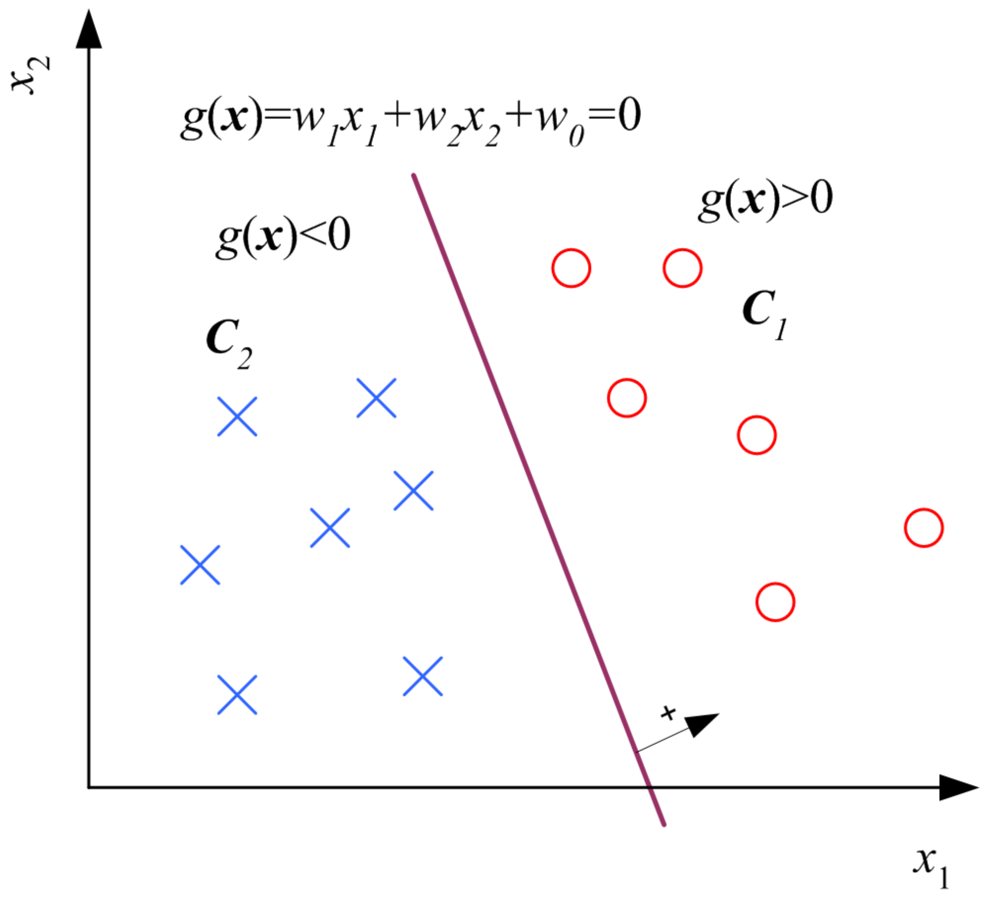
- Higher-order (product) terms:

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

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

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

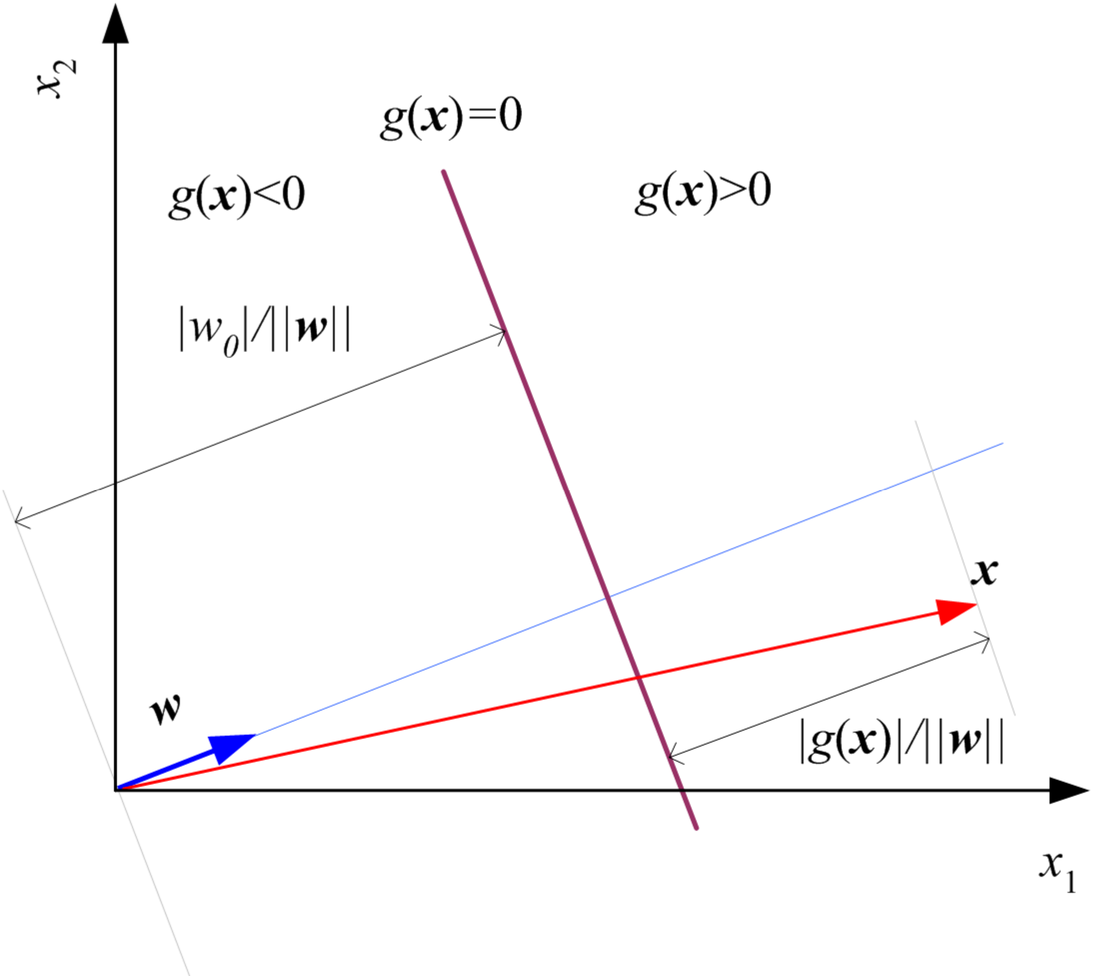
Two Classes



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

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

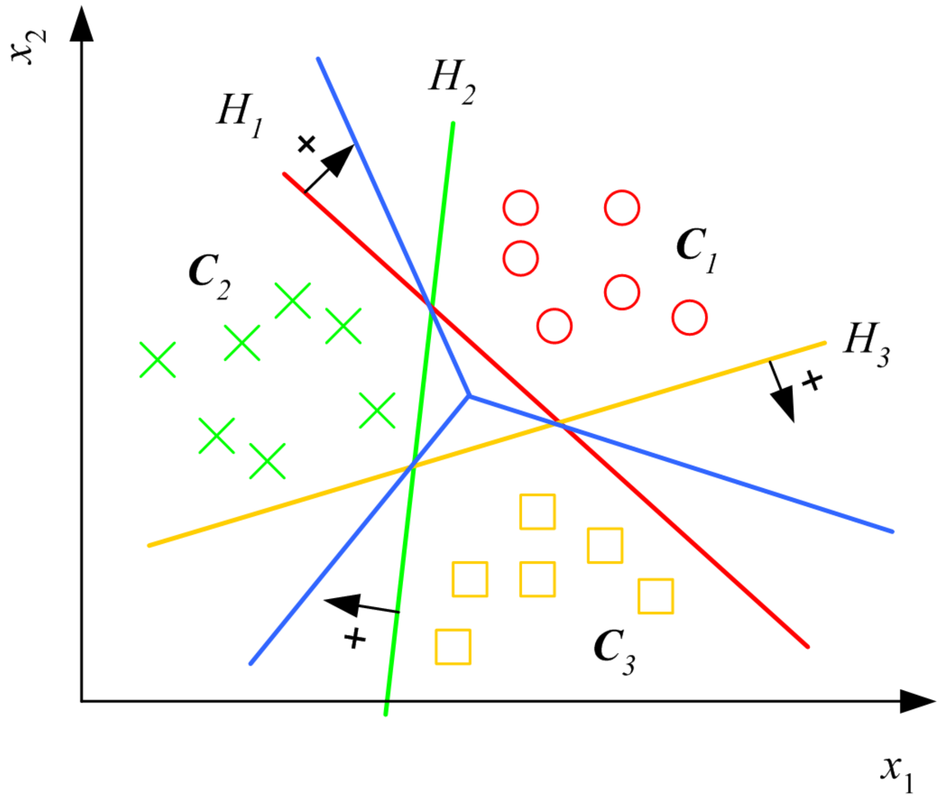
Geometry



x_2
$g(\boldsymbol{x})=0$
$g(\boldsymbol{x})<0$
$g(\boldsymbol{x})>0$
$|w_0|/||\boldsymbol{w}||$
$\boldsymbol{x}$
$\boldsymbol{w}$
$|g(\boldsymbol{x})|/||\boldsymbol{w}||$
x_1

Multiple Classes

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

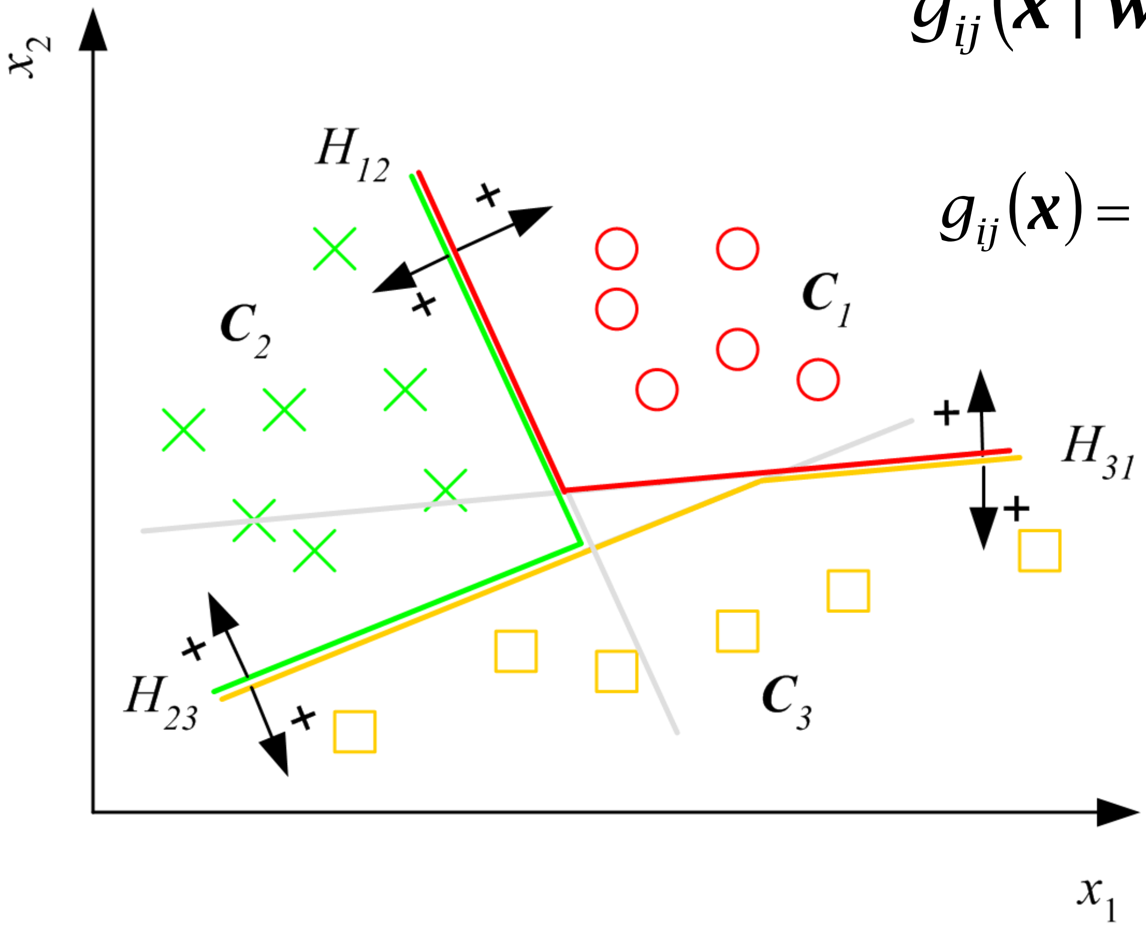


Choose C_i if

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

Classes are
linearly separable

Pairwise Separation



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

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

choose C_i if

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

From Discriminants to Posteriors

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

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

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

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

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

$$\operatorname{logit}(P(C_1 \mid \boldsymbol{x})) = \log \frac{P(C_1 \mid \boldsymbol{x})}{1 - P(C_1 \mid \boldsymbol{x})} = \log \frac{P(C_1 \mid \boldsymbol{x})}{P(C_2 \mid \boldsymbol{x})}$$

$$= \log \frac{p(\boldsymbol{x} \mid C_1)}{p(\boldsymbol{x} \mid C_2)} + \log \frac{P(C_1)}{P(C_2)}$$

$$= \log \frac{(2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left[-(1/2)(\boldsymbol{x} - \mu_1)^T \Sigma^{-1} (\boldsymbol{x} - \mu_1)\right]}{(2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left[-(1/2)(\boldsymbol{x} - \mu_2)^T \Sigma^{-1} (\boldsymbol{x} - \mu_2)\right]} + \log \frac{P(C_1)}{P(C_2)}$$

$$= \boldsymbol{w}^T \boldsymbol{x} + w_0$$

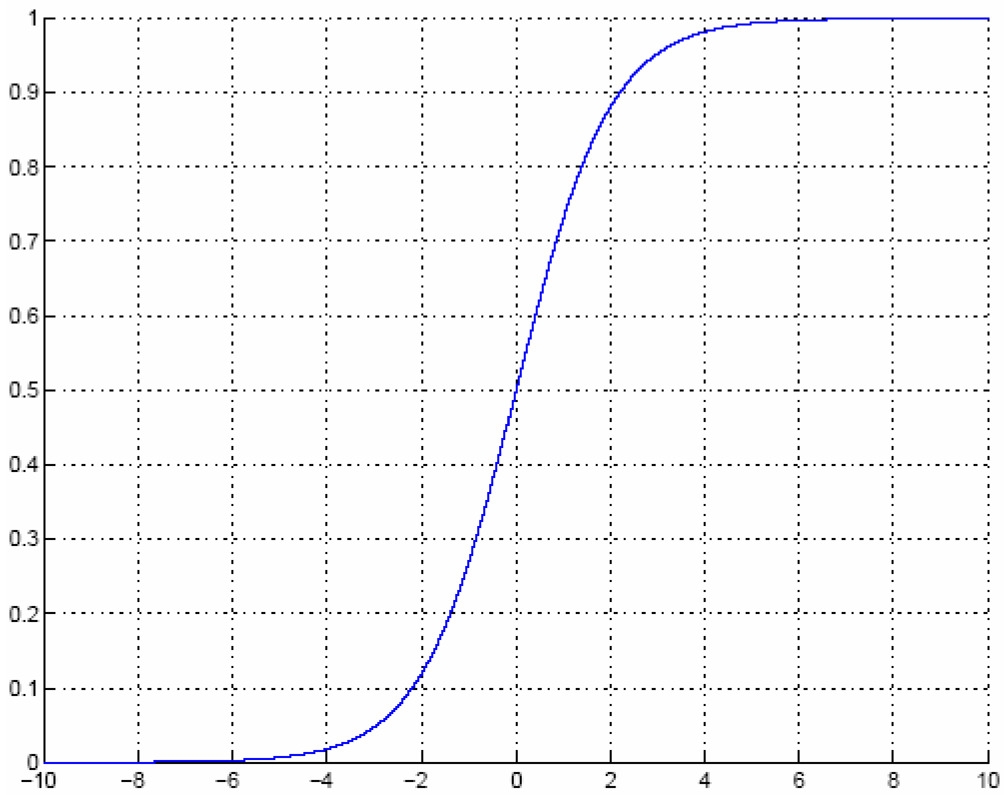
$$\text{where } \boldsymbol{w} = \Sigma^{-1}(\mu_1 - \mu_2) \quad w_0 = -\frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)$$

The inverse of logit

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

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

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

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

- Gradient

$$\nabla_w E = \left[\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \ldots, \frac{\partial E}{\partial w_d} \right]^T$$

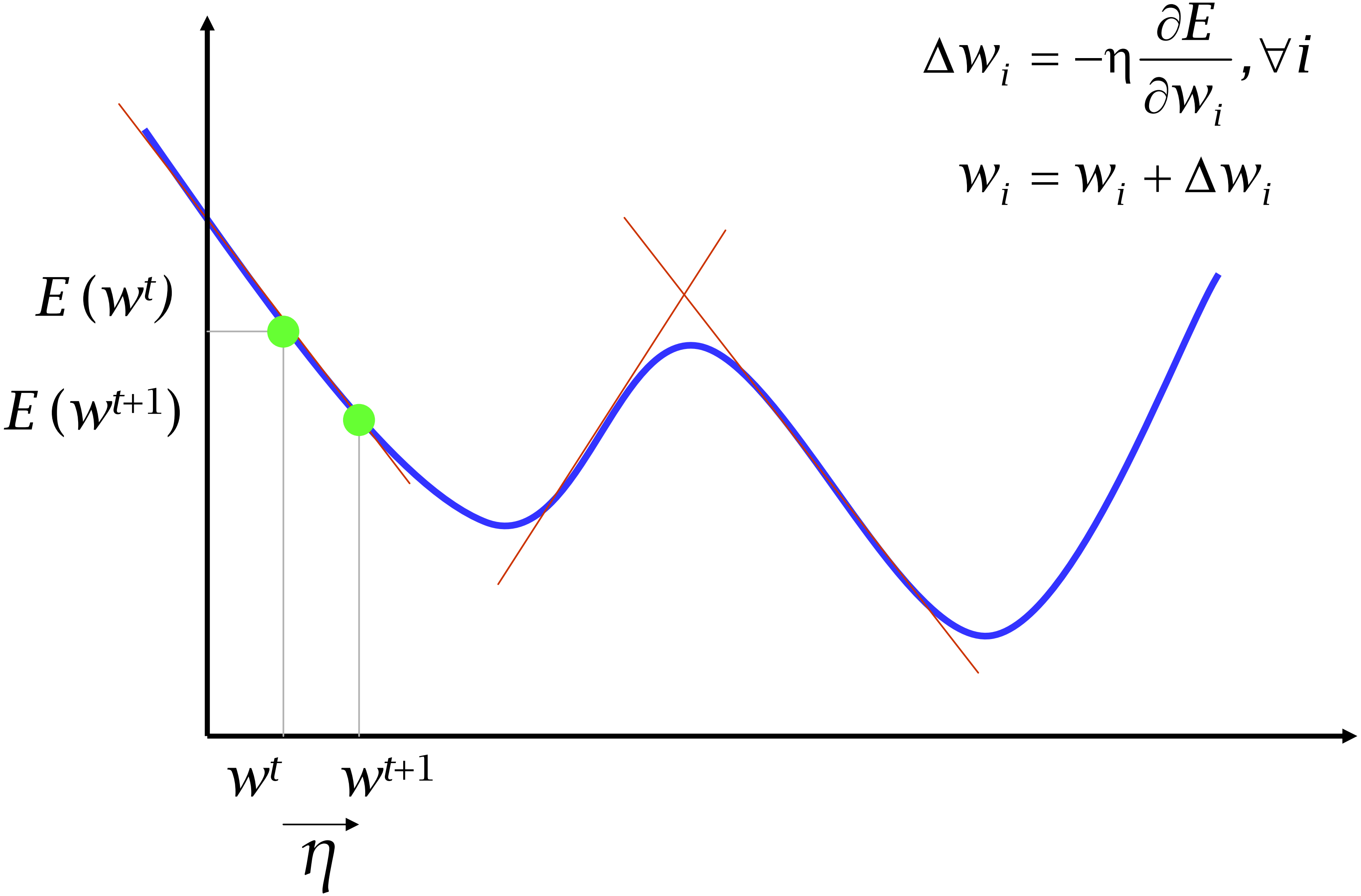
- Gradient-descent:

  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

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

$$\text{logit}(P(C_1 \mid \boldsymbol{x})) = \log \frac{P(C_1 \mid \boldsymbol{x})}{1 - P(C_1 \mid \boldsymbol{x})} = \log \frac{p(\boldsymbol{x} \mid C_1)}{p(\boldsymbol{x} \mid C_2)} + \log \frac{P(C_1)}{P(C_2)}$$

$$= \boldsymbol{w}^T \boldsymbol{x} + w_0$$

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

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

Training: Two Classes

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

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

$$l(\boldsymbol{w}, w_0 \mid \mathcal{X}) = \prod_t \left(y^t\right)^{\left(r^t\right)}\left(1 - y^t\right)^{\left(1 - r^t\right)}$$

$$E = -\log l$$

$$E(\boldsymbol{w}, w_0 \mid \mathcal{X}) = -\sum_t r^t \log y^t + \left(1 - r^t\right)\log\left(1 - y^t\right)$$

Training: Gradient-Descent

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

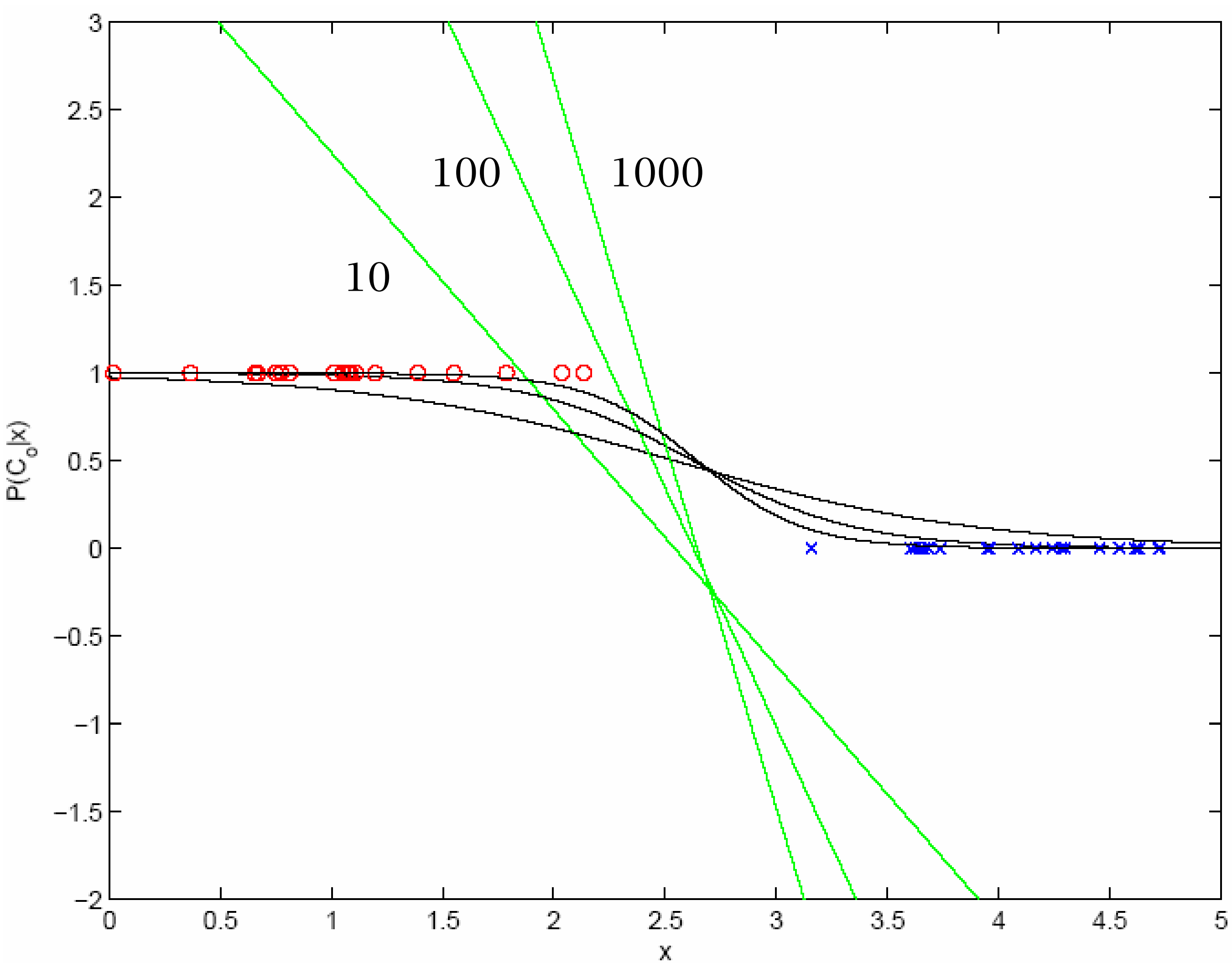
$$\text{If } y = \text{sigmoid}(a) \quad \frac{dy}{da} = y(1 - y)$$

$$\Delta w_j = -\eta \frac{\partial E}{\partial w_j} = \eta \sum_t \left(\frac{r^t}{y^t} - \frac{1 - r^t}{1 - y^t} \right) y^t (1 - y^t) x_j^t$$

$$= \eta \sum_t (r^t - y^t) x_j^t, j = 1, \ldots, d$$

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

```
For j = 0, ..., d
    w_j ← rand(-0.01,0.01)
Repeat
    For j = 0, ..., d
        Δw_j ← 0
    For t = 1, ..., N
        o ← 0
        For j = 0, ..., d
            o ← o + w_j x_j^t
        y ← sigmoid(o)
        Δw_j ← Δw_j + (r^t − y) x_j^t
    For j = 0, ..., d
        w_j ← w_j + η Δw_j
Until convergence
```



100
1000
10
P(C_o|x)
x

K>2 Classes

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

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

softmax

$$y = \hat{P}(C_i \mid \boldsymbol{x}) = \frac{\exp\left[\boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}\right]}{\sum_{j=1}^{K} \exp\left[\boldsymbol{w}_j^T \boldsymbol{x} + w_{j0}\right]}, i = 1,\ldots,K$$

$$l(\{\boldsymbol{w}_i, w_{i0}\}_i \mid \mathcal{X}) = \prod_t \prod_i \left(y_i^t\right)^{\left(r_i^t\right)}$$

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

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

For $i = 1, \ldots, K$, For $j = 0, \ldots, d$, $w_{ij} \leftarrow$ rand$(-0.01, 0.01)$

Repeat

  For $i = 1, \ldots, K$, For $j = 0, \ldots, d$, $\Delta w_{ij} \leftarrow 0$

  For $t = 1, \ldots, N$

    For $i = 1, \ldots, K$

      $o_i \leftarrow 0$

      For $j = 0, \ldots, d$

        $o_i \leftarrow o_i + w_{ij} x_j^t$

    For $i = 1, \ldots, K$

      $y_i \leftarrow \exp(o_i) / \sum_k \exp(o_k)$

    For $i = 1, \ldots, K$

      For $j = 0, \ldots, d$

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

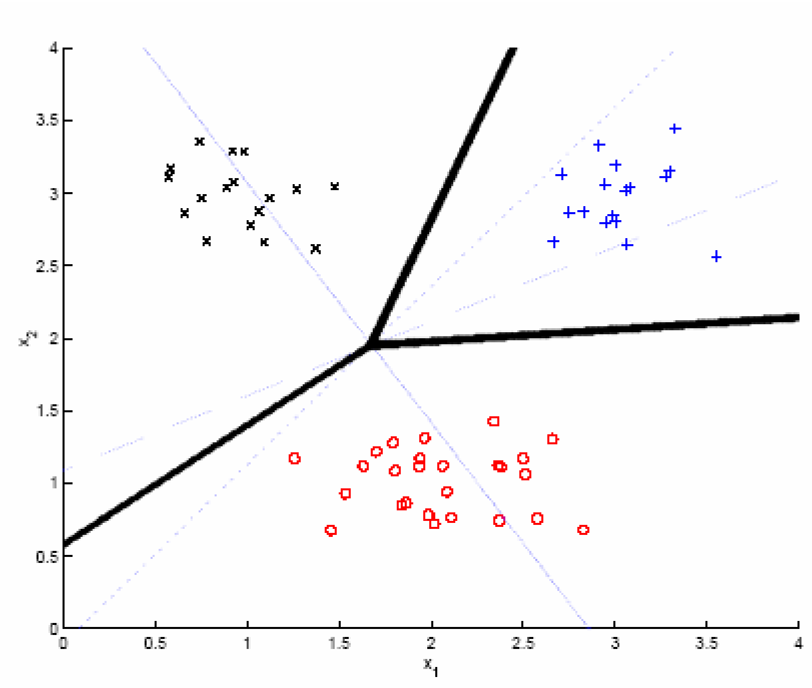
  For $i = 1, \ldots, K$

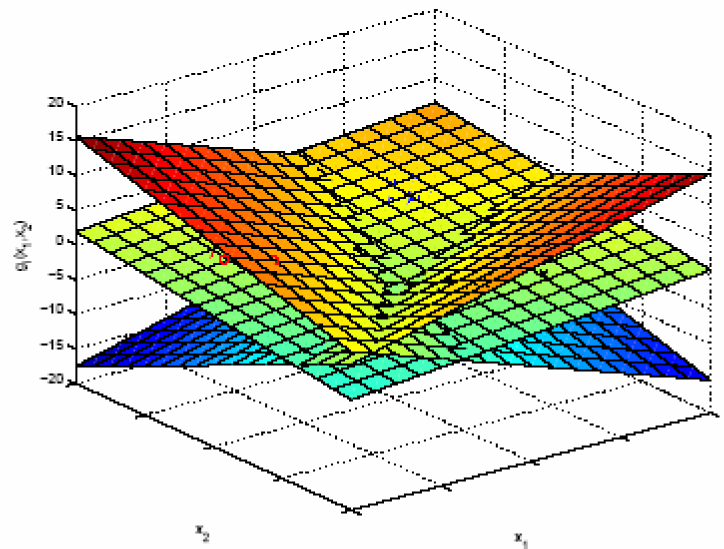
    For $j = 0, \ldots, d$

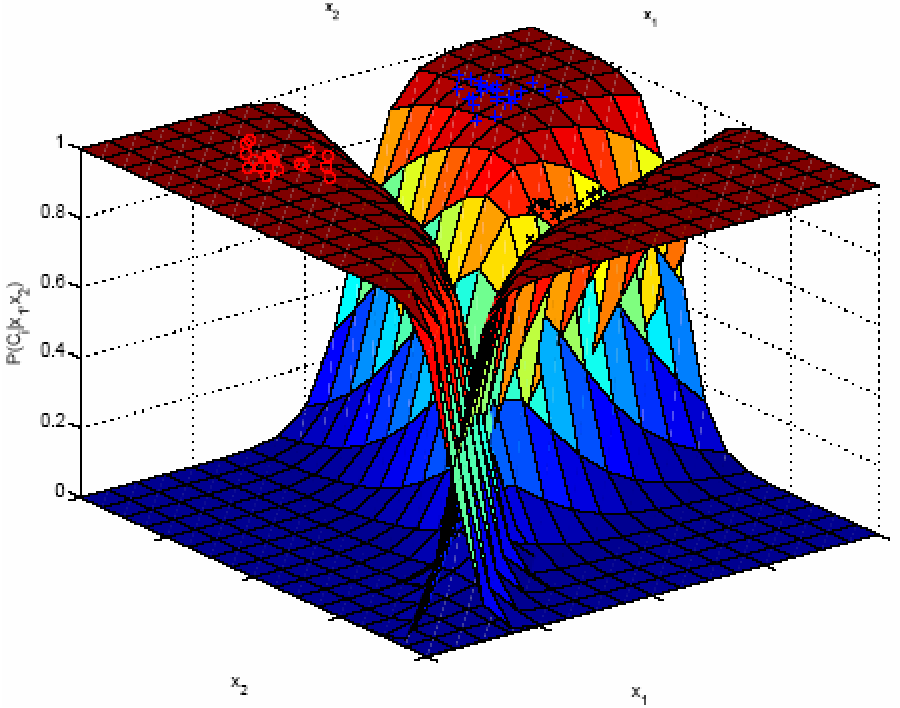
      $w_{ij} \leftarrow w_{ij} + \eta \Delta w_{ij}$

Until convergence

Example







Generalizing the Linear Model

- Quadratic:

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

- Sum of basis functions:

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

  where $\phi(\mathbf{x})$ are basis functions
  - Kernels in SVM
  - Hidden units in neural networks

Discrimination by Regression

- Classes are NOT mutually exclusive and exhaustive

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

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

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

$$E(\boldsymbol{w}, w_0 \mid \mathcal{X}) = \frac{1}{2}\sum_t \left(r^t - y^t\right)^2$$

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

Optimal Separating Hyperplane

$$\mathcal{X} = \{\boldsymbol{x}^t, r^t\}_t \text{ where } r^t = \begin{cases} +1 & \text{if } \boldsymbol{x}^t \in C_1 \\ -1 & \text{if } \boldsymbol{x}^t \in C_2 \end{cases}$$

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

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

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

which can be rewritten as

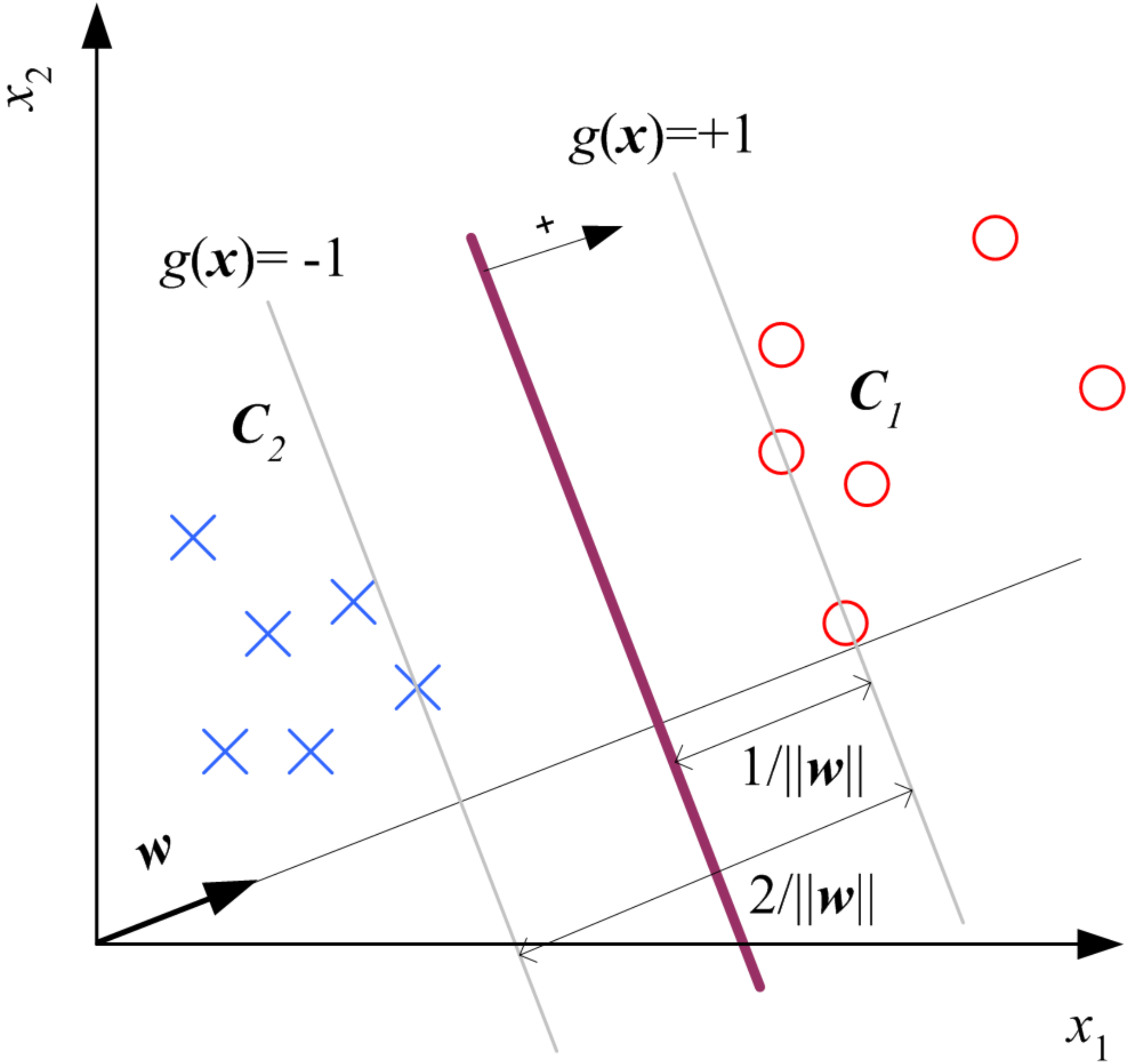
$$r^t \left(\boldsymbol{w}^T \boldsymbol{x}^t + w_0\right) \geq +1$$

(Cortes and Vapnik, 1995; Vapnik, 1995)

Margin

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

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



x_2
$g(\boldsymbol{x})=+1$
$g(\boldsymbol{x})= -1$
+
C_1
C_2
$1/||\boldsymbol{w}||$
$2/||\boldsymbol{w}||$
$\boldsymbol{w}$
x_1

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

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{t=1}^{N} \alpha^t \left[r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) - 1\right]$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{t=1}^{N} \alpha^t r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) + \sum_{t=1}^{N} \alpha^t$$

$$\frac{\partial L_p}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{t=1}^{N} \alpha^t r^t \mathbf{x}^t$$

$$\frac{\partial L_p}{\partial w_0} = 0 \Rightarrow \sum_{t=1}^{N} \alpha^t r^t = 0$$

$$L_d = \frac{1}{2}\left(\mathbf{w}^T\mathbf{w}\right) - \mathbf{w}^T\sum_t \alpha^t r^t \mathbf{x}^t - w_0\sum_t \alpha^t r^t + \sum_t \alpha^t$$

$$= -\frac{1}{2}\left(\mathbf{w}^T\mathbf{w}\right) + \sum_t \alpha^t$$

$$= -\frac{1}{2}\sum_t\sum_s \alpha^t\alpha^s r^t r^s \left(\mathbf{x}^t\right)^T \mathbf{x}^s + \sum_t \alpha^t$$

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

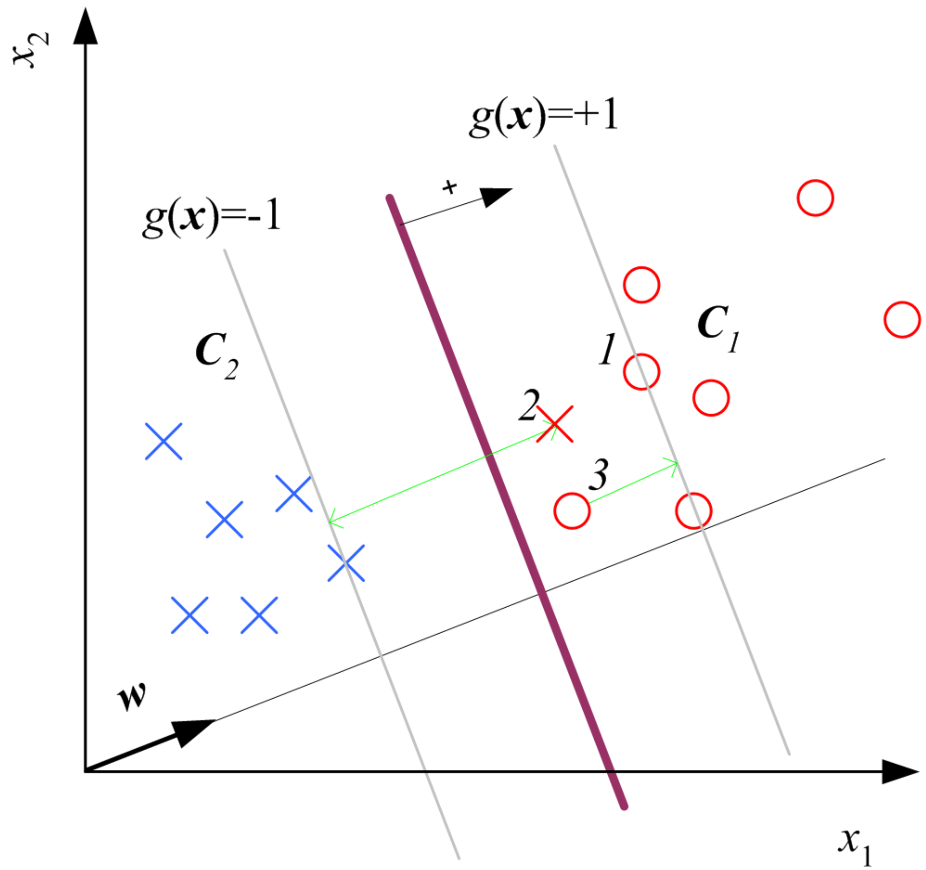
- Not linearly separable

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

- Soft error

$$\sum_t \xi^t$$

- New primal is



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

Kernel Machines

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

$$\boldsymbol{z} = \varphi(\boldsymbol{x}) \qquad g(\boldsymbol{z}) = \boldsymbol{w}^T \boldsymbol{z}$$

$$g(\boldsymbol{x}) = \boldsymbol{w}^T \varphi(\boldsymbol{x})$$

- The SVM solution

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

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

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

Kernel Functions

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

$$K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T \boldsymbol{y} + 1)^2$$
$$= (x_1 y_1 + x_2 y_2 + 1)^2$$
$$= 1 + 2x_1 y_1 + 2x_2 y_2 + 2x_1 x_2 y_1 y_2 + x_1^2 y_1^2 + x_2^2 y_2^2$$
$$\phi(\boldsymbol{x}) = \left[1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2, x_1^2, x_2^2\right]^T$$

- Radial-basis functions: $K(\boldsymbol{x}^t, \boldsymbol{x}) = \exp\left[-\frac{\|\boldsymbol{x}^t - \boldsymbol{x}\|^2}{\sigma^2}\right]$
- Sigmoidal functions: $K(\boldsymbol{x}^t, \boldsymbol{x}) = \tanh(2\boldsymbol{x}^T \boldsymbol{x}^t + 1)$

(Cherkassky and Mulier, 1998)

SVM for Regression

- Use a linear model (possibly kernelized)

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

- Use the ϵ-sensitive error function

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

- $$\min \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_t \left(\xi_+^t + \xi_-^t\right)$$

$$r^t - \left(\boldsymbol{w}^T\boldsymbol{x} + w_0\right) \le \varepsilon + \xi_+^t$$

$$\left(\boldsymbol{w}^T\boldsymbol{x} + w_0\right) - r^t \le \varepsilon + \xi_-^t$$

$$\xi_+^t, \xi_-^t \ge 0$$