BLG 454E Learning from Data

FALL 2022-2023

These slides are prepeared by Assoc. Prof. Yusuf Yaslan & Assist. Prof. Ayşe Tosun

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:

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

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

Generalized Linear Model

- Quadratic discriminant:

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

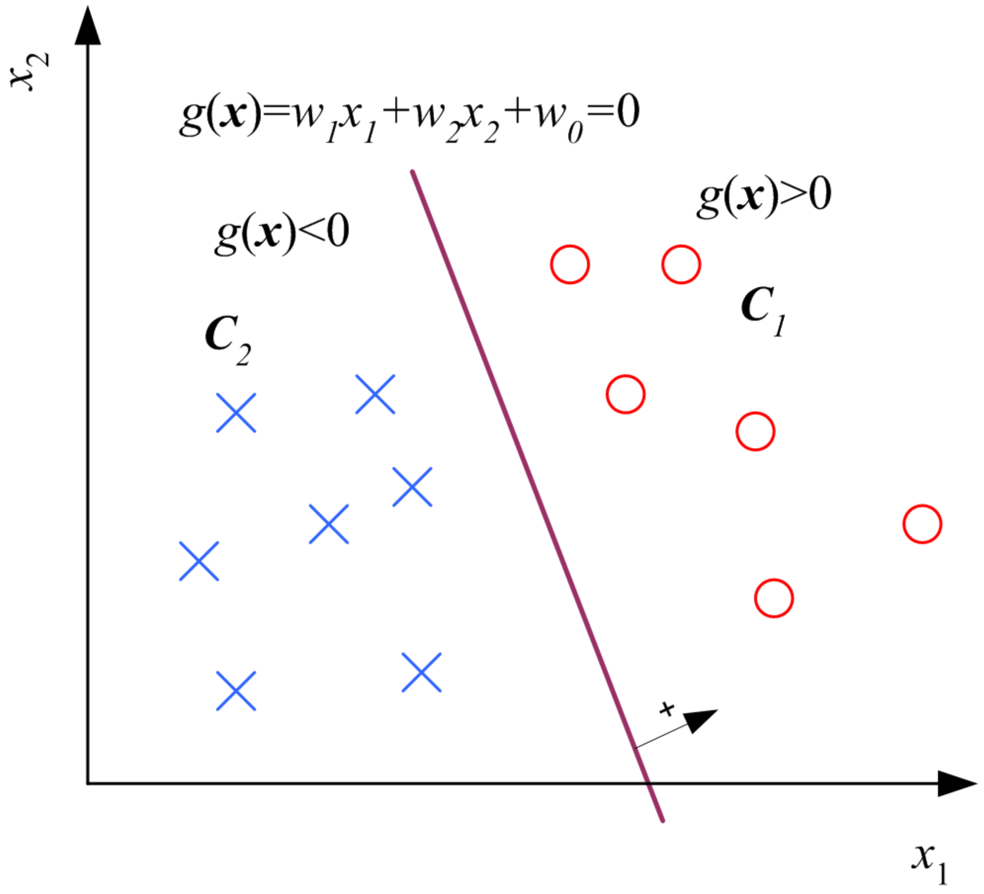
- Higher-order (product) terms:

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

Map from ***x*** to ***z*** using nonlinear basis functions and use a linear discriminant in ***z***-space

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

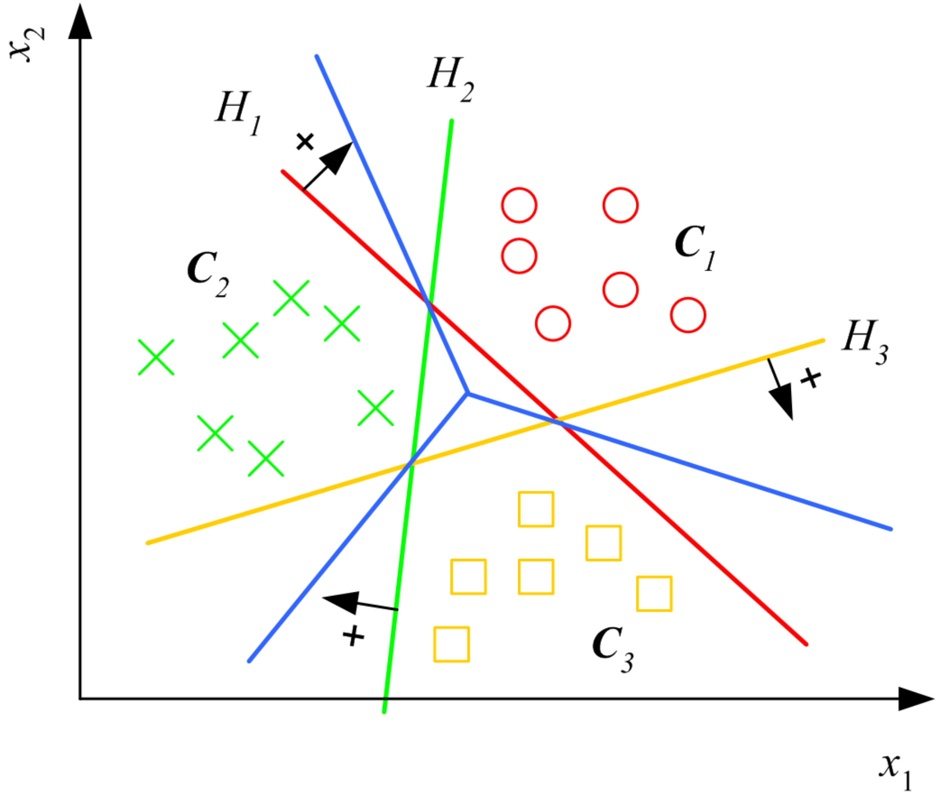
Two Classes



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

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

Multiple Classes



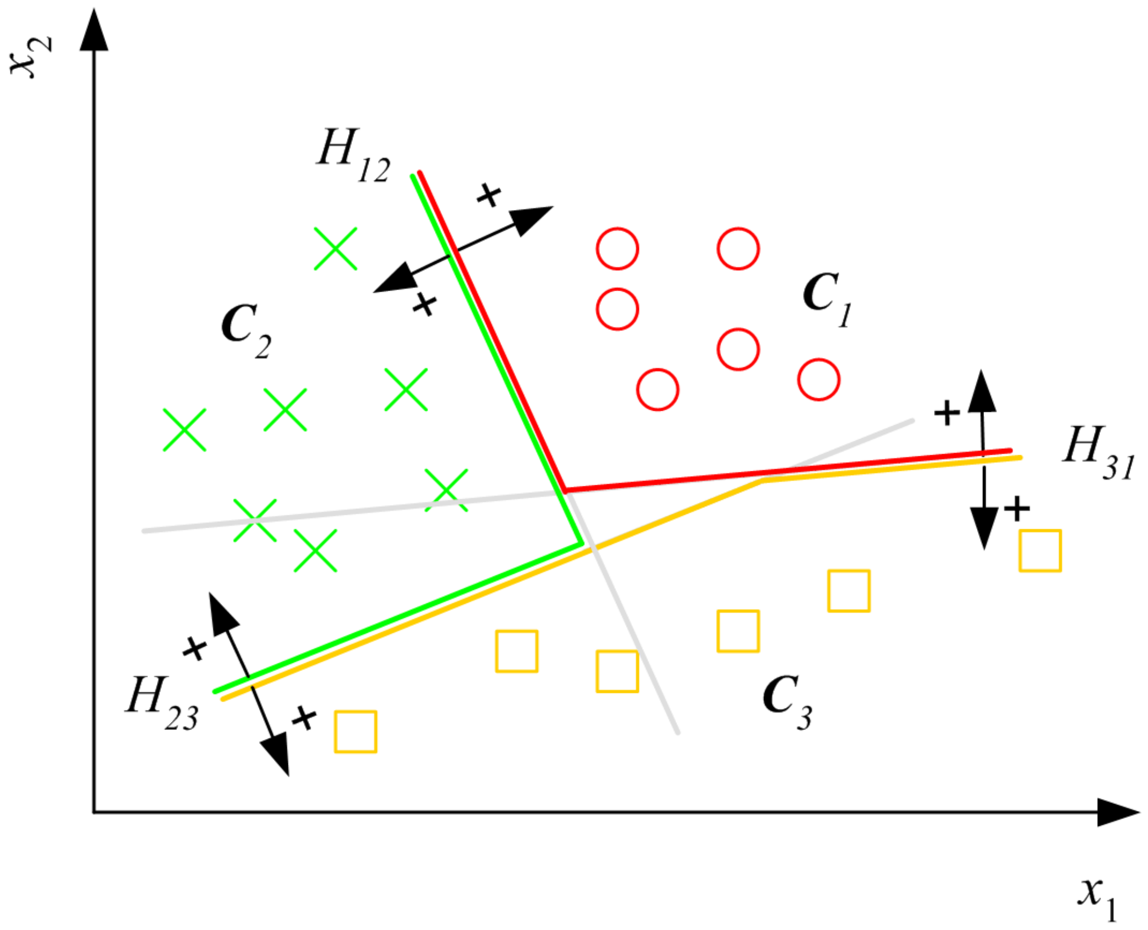
$$g_i(\mathbf{x}|\mathbf{w}_i, w_{i0}) = \mathbf{w}_i^T\mathbf{x} + w_{i0}$$

Choose C_i if

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

Classes are linearly separable
Each hyperplane H_i separates the examples of C_i from the examples of all other classes

Pairwise Separation



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

During training:

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

During testing:

choose C_i if

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

From Discriminants to Posteriors

We saw that when $p\ (\boldsymbol{x} \mid C_i) \sim N\ (\boldsymbol{\mu}_i\ , \Sigma)$ and share a common covariance matrix the discriminant function is linear

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

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

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

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

$$\text{logit}\left(P(C_1 \mid \mathbf{x})\right) = \log \frac{P(C_1 \mid \mathbf{x})}{1 - P(C_1 \mid \mathbf{x})} = \log \frac{P(C_1 \mid \mathbf{x})}{P(C_2 \mid \mathbf{x})}$$

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

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

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

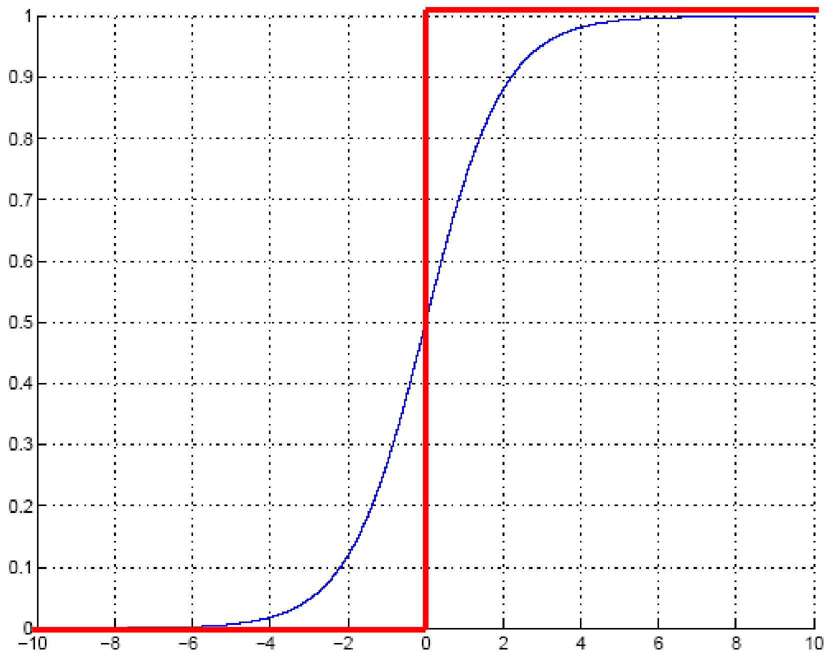
$$\text{where } \mathbf{w} = \Sigma^{-1}(\mu_1 - \mu_2) \quad w_0 = -\frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) \quad + \log \frac{P(C_1)}{P(C_2)}$$

The inverse of logit

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

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

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(***w***|X) is error with parameters ***w*** on sample X

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

- Gradient

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

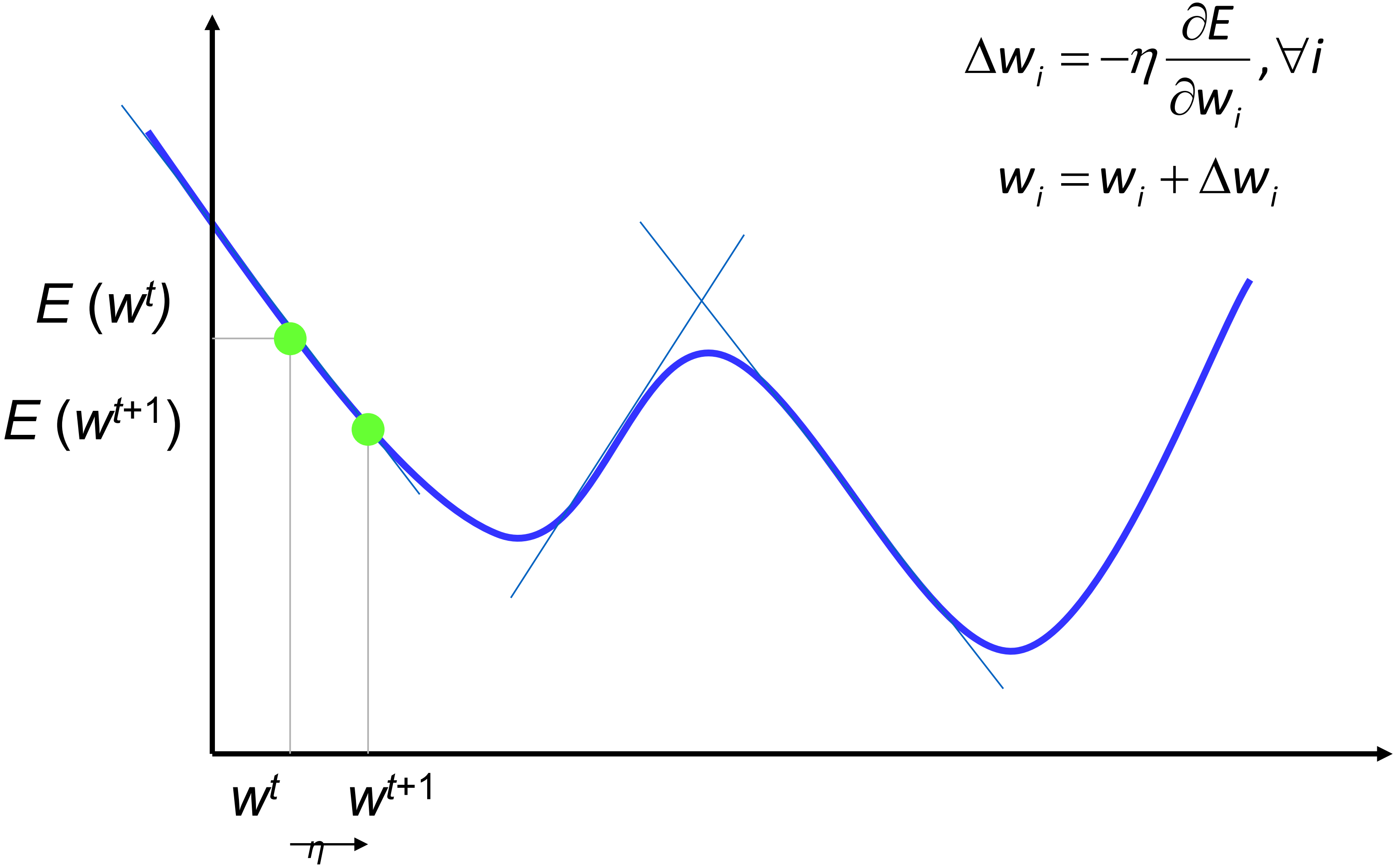
- Gradient-descent:

  Starts from random ***w*** and updates ***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

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

$$\operatorname{logit}(P(C_1|\mathbf{x})) = \log \frac{P(C_1|\mathbf{x})}{1-P(C_1|\mathbf{x})} = \log \frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)} + \log \frac{P(C_1)}{P(C_2)}$$

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

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

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

Training: Two Classes

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

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

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

$$E = -\log l$$

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

Cross Entropy

Training: Gradient-Descent

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

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

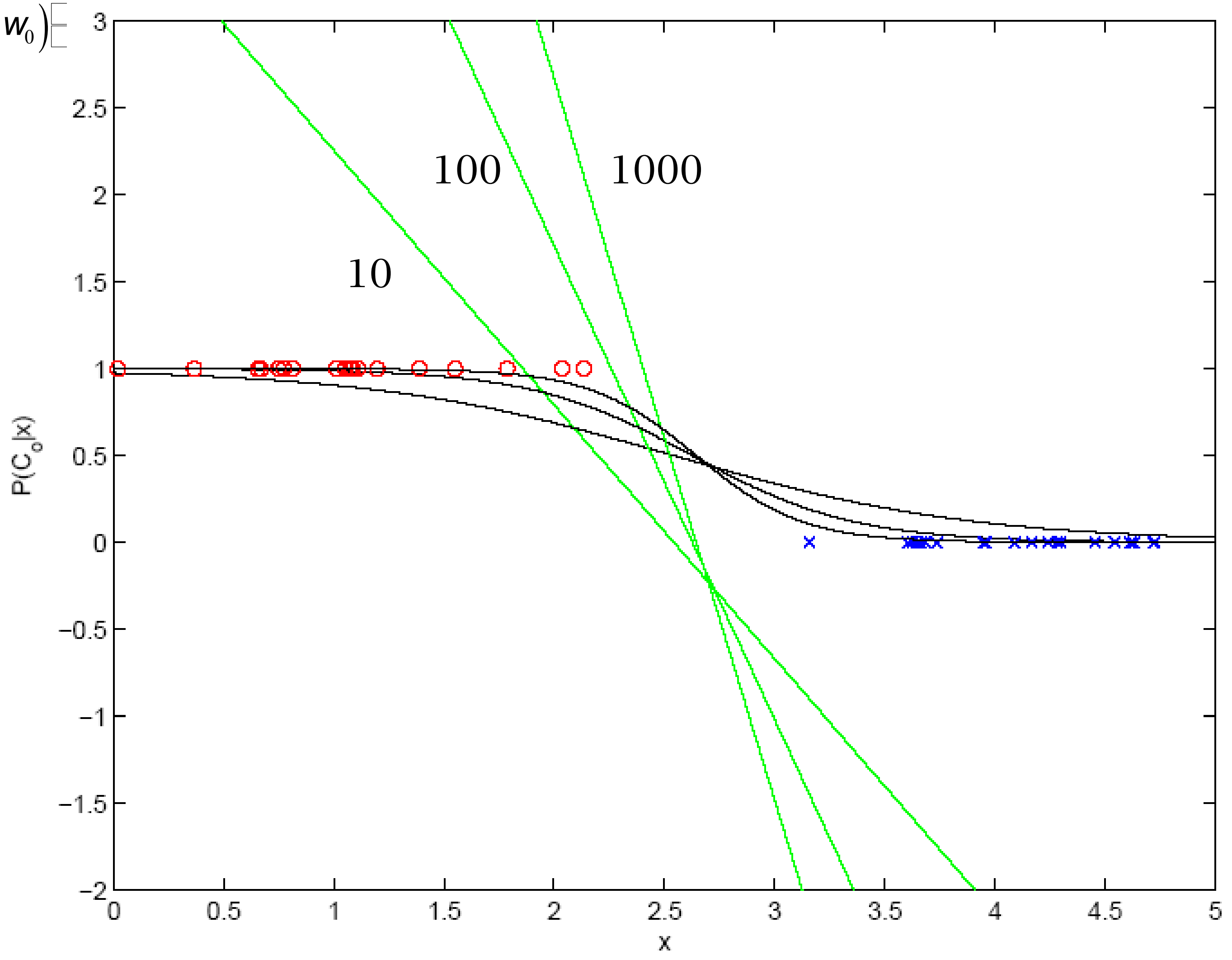
$$\Delta w_j = -\eta \frac{\partial E}{\partial w_j} = \eta \sum_t \left(\frac{r^t}{y^t} - \frac{1 - r^t}{1 - y^t} \right) y^t (1 - y^t) x_j^t$$

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

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

```
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
```

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



K>2 Classes

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

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

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

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

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

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

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 $j = 0, \ldots, d$

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

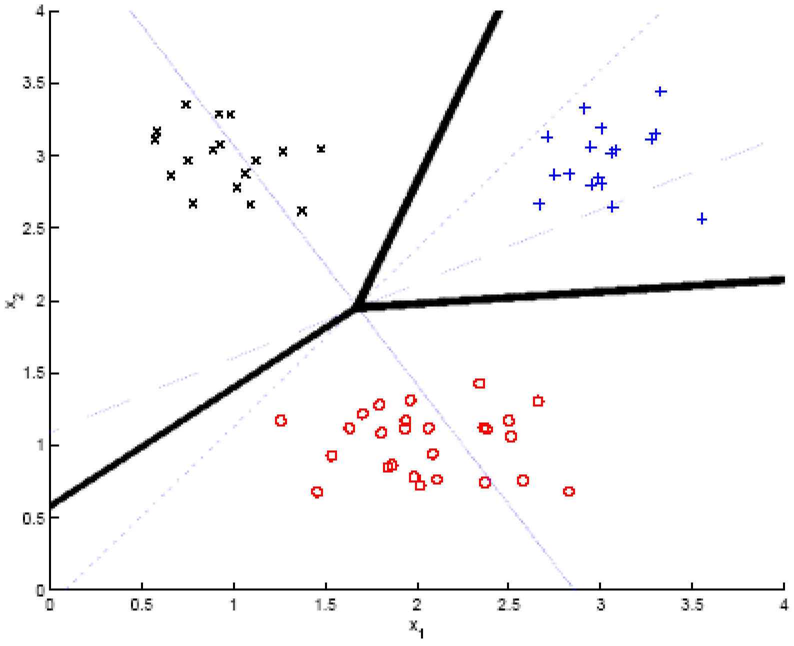
    For $i = 1, \ldots, K$

        For $j = 0, \ldots, d$

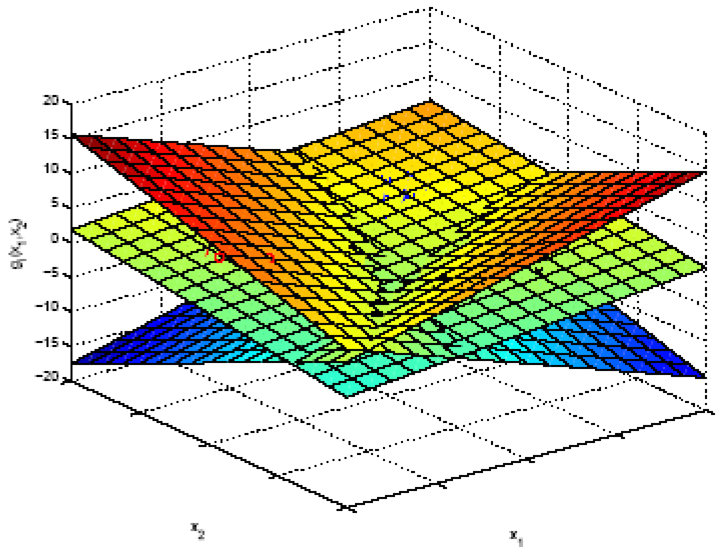
            $w_{ij} \leftarrow w_{ij} + \eta \Delta w_{ij}$

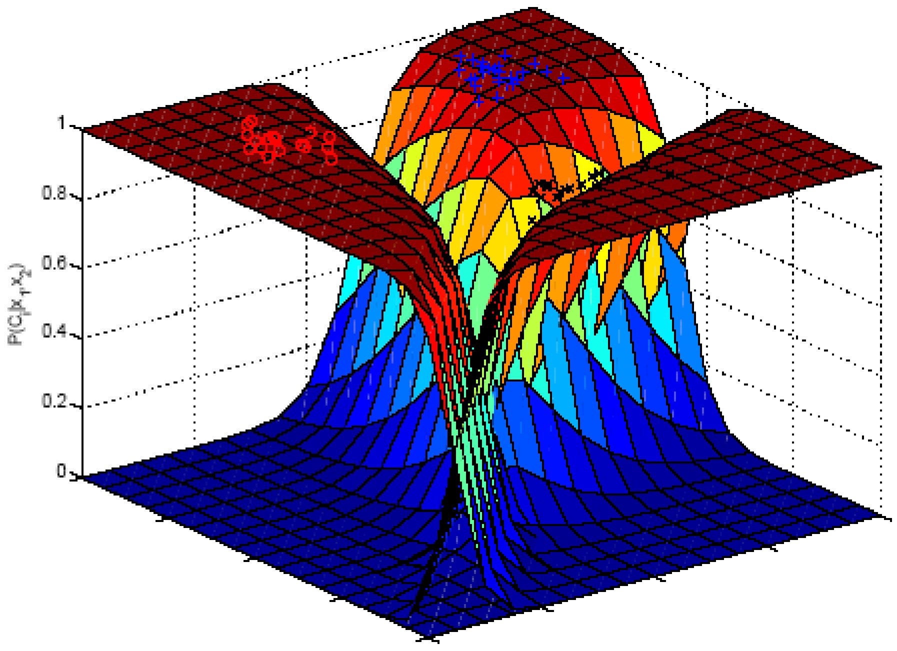
Until convergence

Example



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





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 **φ**(***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

$$r^t = y^t + \varepsilon \text{ where } \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right) \qquad r^t \in \{0,1\}$$

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

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

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

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