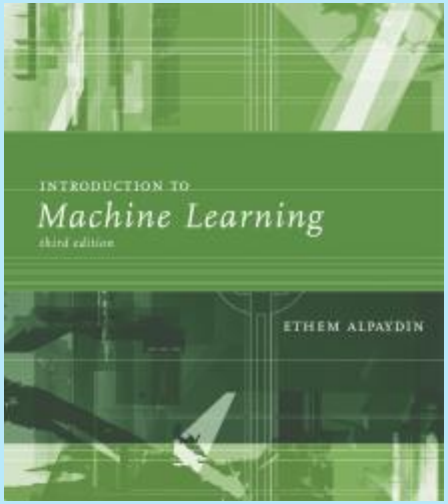


Lecture Slides for

INTRODUCTION TO MACHINE LEARNING

3RD EDITION

ETHEM ALPAYDIN

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml3e

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:

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

- 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:

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

- Higher-order (product) terms:

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

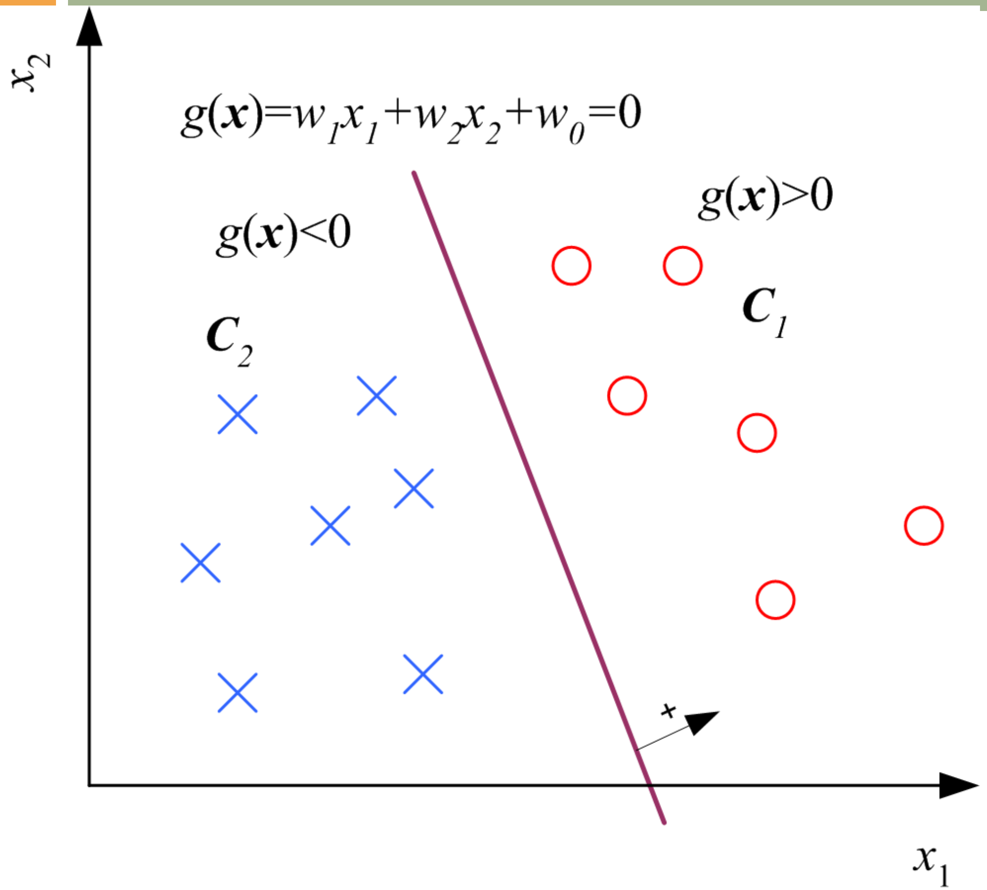
Map from ***x*** to ***z*** using nonlinear basis functions and use a linear discriminant in ***z***-space

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

$$\sin(x_1), \quad \exp\left(-(x_1 - m)^2 / c\right),$$

$$\exp\left(-\|\mathbf{x} - \mathbf{m}\|^2 / c\right), \; \log(x_2), 1\left(x_1 > c\right)$$

Two Classes

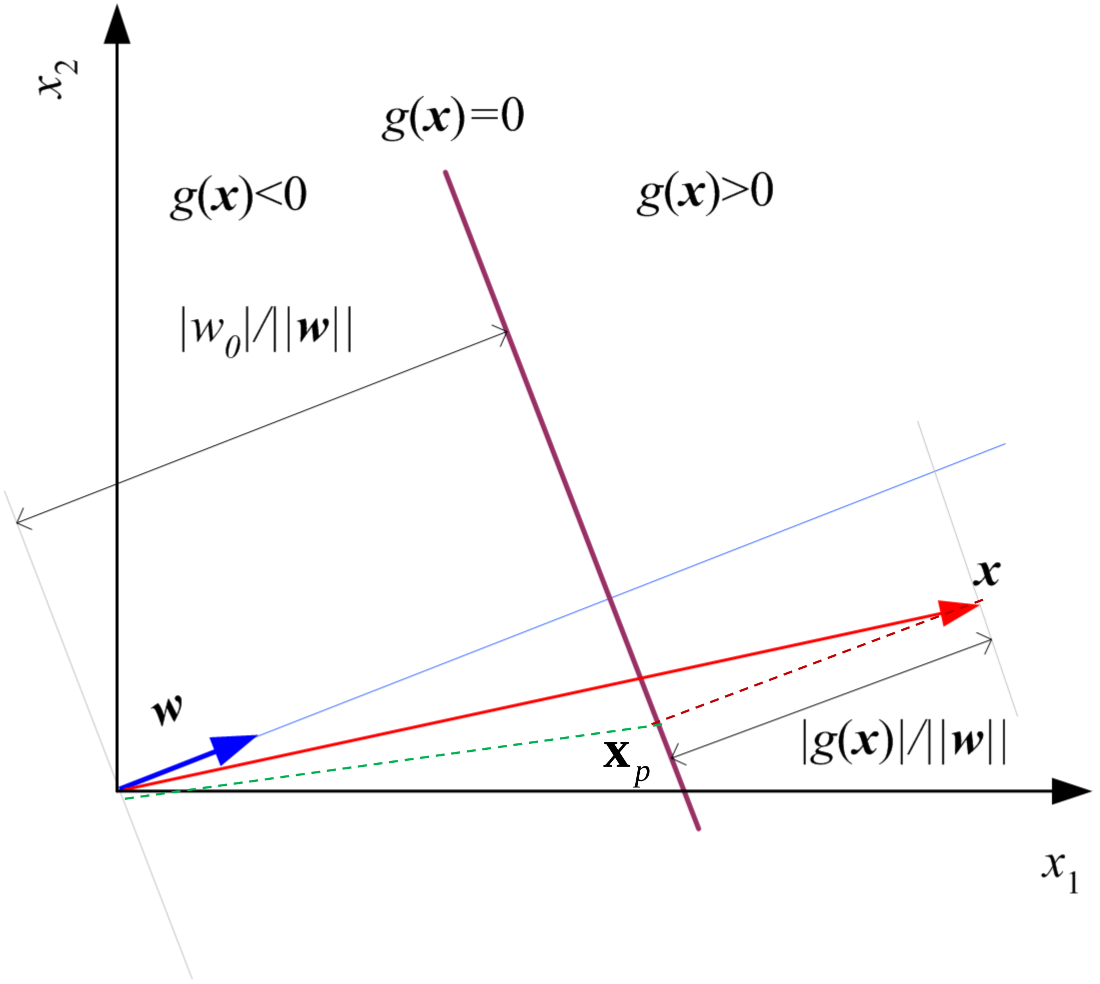


$$
\begin{aligned}
g(\mathbf{x}) &= g_1(\mathbf{x}) - g_2(\mathbf{x}) \\
&= \left(\mathbf{w}_1^T\mathbf{x} + w_{10}\right) - \left(\mathbf{w}_2^T\mathbf{x} + w_{20}\right) \\
&= \left(\mathbf{w}_1 - \mathbf{w}_2\right)^T \mathbf{x} + \left(w_{10} - w_{20}\right) \\
&= \mathbf{w}^T\mathbf{x} + w_0
\end{aligned}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

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

Take two points $\mathbf{x}_1$ and $\mathbf{x}_2$ both on the decision surface; that is, $g(\mathbf{x}_1) = g(\mathbf{x}_2) = 0$, then $\mathbf{w}^T\,\mathbf{x}_1 + w_0 = \mathbf{w}^T\,\mathbf{x}_2 + w_0 \rightarrow \mathbf{w}^T(\mathbf{x}_1 - \mathbf{x}_2)=0$

Geometry



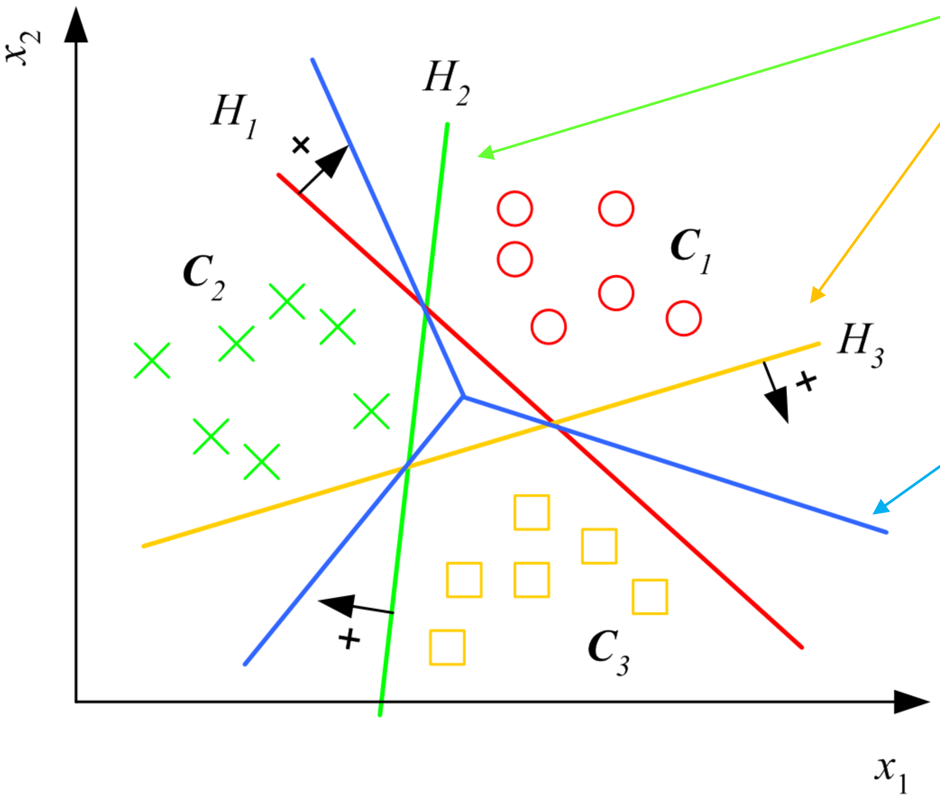
$$\boldsymbol{x} = \boldsymbol{x}_p + r\frac{\boldsymbol{w}}{||\boldsymbol{w}||}$$

$$r = \frac{g(\boldsymbol{x})}{||\boldsymbol{w}||}$$

$$r_0 = \frac{w_0}{||\boldsymbol{w}||}$$

Multiple Classes

$$g_i(\mathbf{x}|\mathbf{w}_i, w_{i0}) = \mathbf{w}_i^T\mathbf{x} + w_{i0} = \begin{cases} > 0 & \text{if } \mathbf{x} \in C_i \\ \leq 0 & \text{otherwise} \end{cases}$$

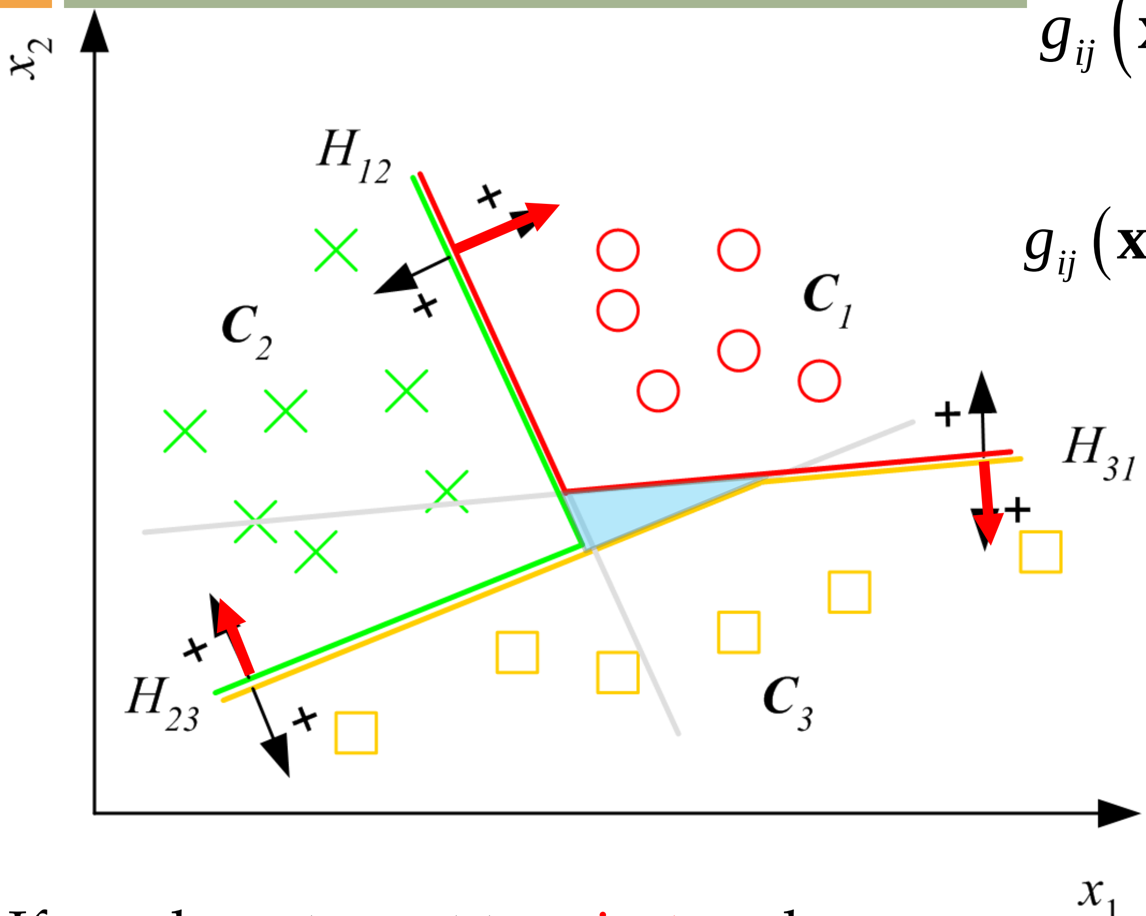


Choose C_i if

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

Classes are
linearly separable

Pairwise Separation

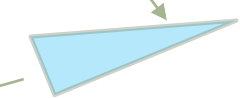


$$g_{ij}\left(\mathbf{x}|\mathbf{w}_{ij}, w_{ij0}\right) = \mathbf{w}_{ij}^T\mathbf{x} + w_{ij0}$$

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

choose C_i if

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



reject

If we do not want to reject such cases, we can relax the conjunction by using a summation and choosing the maximum of

$$g_i(\mathbf{x}) = \sum_{j \neq i} g_{ij}(\mathbf{x})$$

From Discriminants to Posteriors

From Ch5: When $p\,(\boldsymbol{x} \mid C_i) \sim N(\boldsymbol{\mu}_i, \Sigma)$

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

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

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

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

the logit transformation or log odds of y

In the case of two normal classes sharing a common covariance matrix, the log odds is linear

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

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

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

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

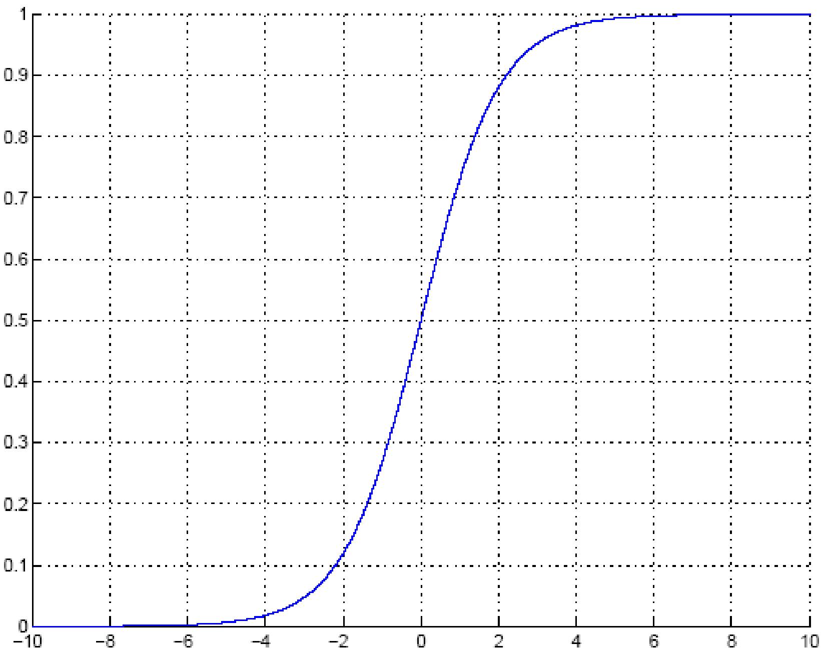
$$\text{where } \mathbf{w} = \Sigma^{-1}(\mu_1 - \mu_2) \quad w_0 = -\frac{1}{2}(\mu_1+\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2) + \log\frac{P(C_1)}{P(C_2)}$$

The inverse of logit:

$$\log \frac{P(C_1|\mathbf{x})}{1-P(C_1|\mathbf{x})} = \mathbf{w}^T\mathbf{x} + w_0$$ is the logistic function, also called the sigmoid function:

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

Sigmoid (Logistic) Function



Calculate $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$ and choose C_1 if $g(\mathbf{x}) > 0$, or

Calculate $y = \text{sigmoid}\left(\mathbf{w}^T\mathbf{x} + w_0\right)$ and choose C_1 if $y > 0.5$

Gradient-Descent

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

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

- Gradient

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

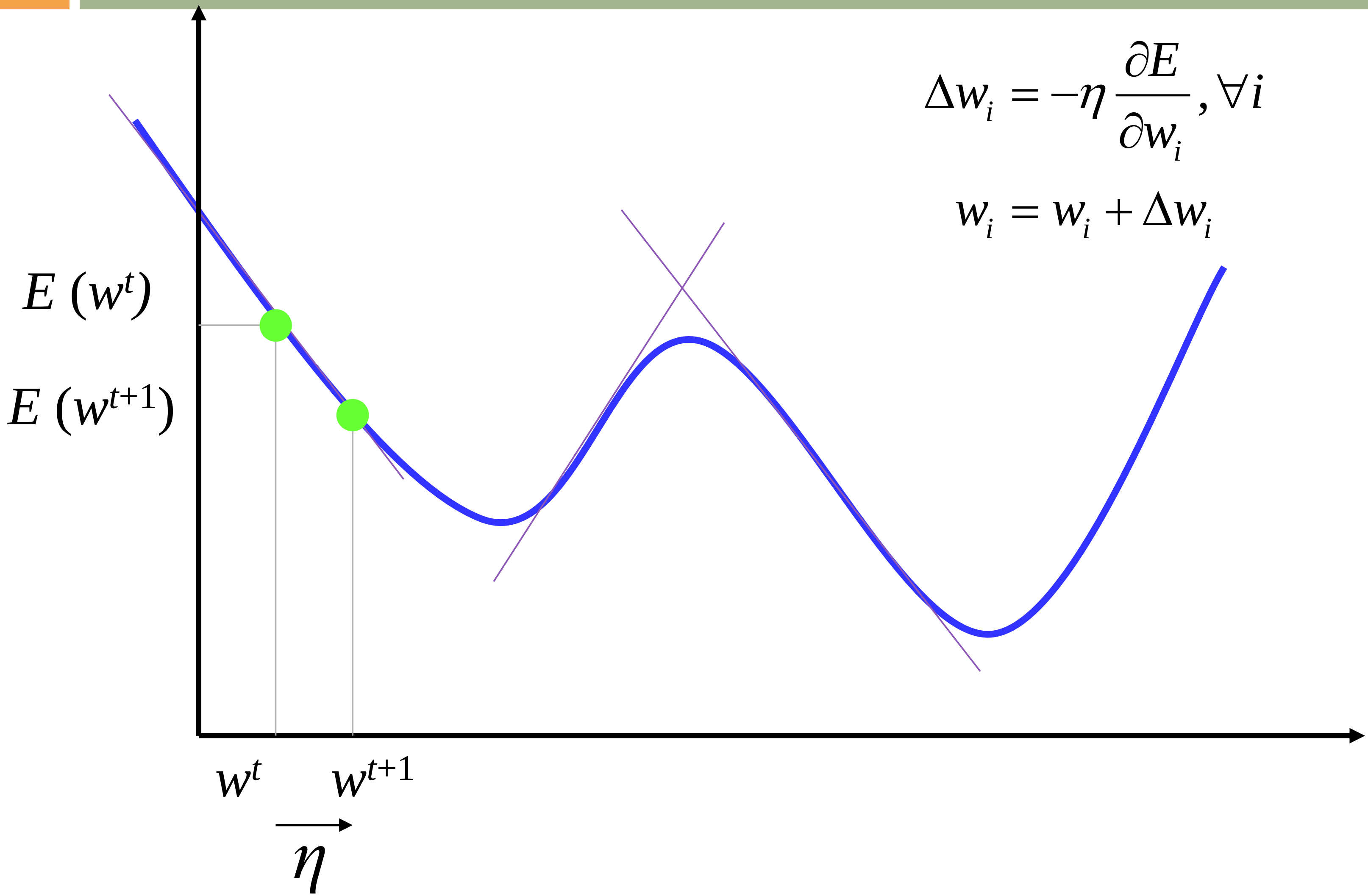
- Gradient-descent:

  Starts from random $\boldsymbol{w}$ and updates $\boldsymbol{w}$ 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

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

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

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

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

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

Training: Two Classes

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

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

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

$E = -\log l$ *cross-entropy* Maximize l ≡Minimize E

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

Training: Gradient-Descent

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

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

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

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

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

For $j = 0, \ldots, d$

  $w_j \leftarrow$rand(-0.01,0.01)

Repeat

  For $j = 0, \ldots, d$

    $\Delta w_j \leftarrow 0$

  For $t = 1, \ldots, N$

    $o \leftarrow 0$

    For $j = 0, \ldots, d$

      $o \leftarrow o + w_j x_j^t$

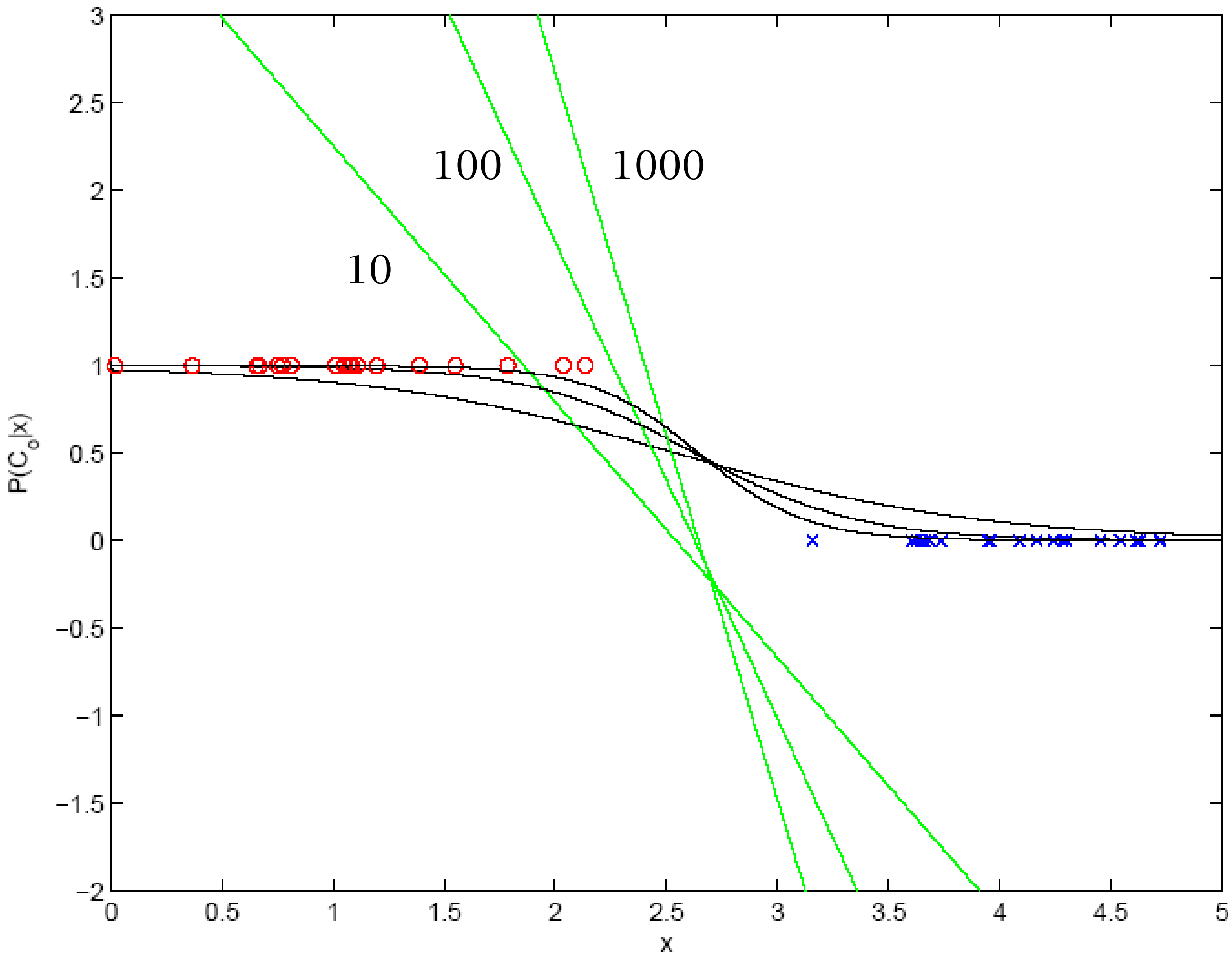
    $y \leftarrow \text{sigmoid}(o)$

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

  For $j = 0, \ldots, d$

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

Until convergence



100
1000
10
P(C_o|x)
x
3
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

Multiple Classes ($K > 2$)

$$\mathcal{X} = \{\mathbf{x}^t, \mathbf{r}^t\}_t \qquad r^t | \mathbf{x}^t \sim \mathrm{Mult}_K\left(1, \mathbf{y}^t\right)$$

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

$$\text{where } w_{i0} = w_{i0}^o + \log \frac{P(C_i)}{P(C_K)}$$

$$\sum_{i=1}^{K-1} \frac{P(C_i|\mathbf{x})}{P(C_K|\mathbf{x})} = \frac{1 - P(C_K|\mathbf{x})}{P(C_K|\mathbf{x})} = \sum_{i=1}^{K-1} \exp\left(\mathbf{w}_i^T \mathbf{x} + w_{i0}\right)$$

$$\Rightarrow P(C_K|\mathbf{x}) = \frac{1}{1 + \sum_{i=1}^{K-1} \exp\left(\mathbf{w}_i^T \mathbf{x} + w_{i0}\right)} = \frac{1}{\sum_{i=1}^{K} \exp\left(\mathbf{w}_i^T \mathbf{x} + w_{i0}\right)}$$

$$\frac{P(C_i|\mathbf{x})}{P(C_K|\mathbf{x})} = \exp\left(\mathbf{w}_i^T\mathbf{x} + w_{i0}\right) \Rightarrow P(C_i|\mathbf{x}) = \frac{\exp\left(\mathbf{w}_i^T\mathbf{x} + w_{i0}\right)}{1 + \sum_{j=1}^{K-1}\exp\left(\mathbf{w}_j^T\mathbf{x} + w_{j0}\right)}$$

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

softmax

$$\sum_i y_i = 1.$$

$$l\left(\{\mathbf{w}_i, w_{i0}\}_i | \mathrm{X}\right) = \prod_t \prod_i \left(y_i^t\right)^{\left(r_i^t\right)}, \quad E = -\log l$$

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

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

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

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 $j = 0, \ldots, d$

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

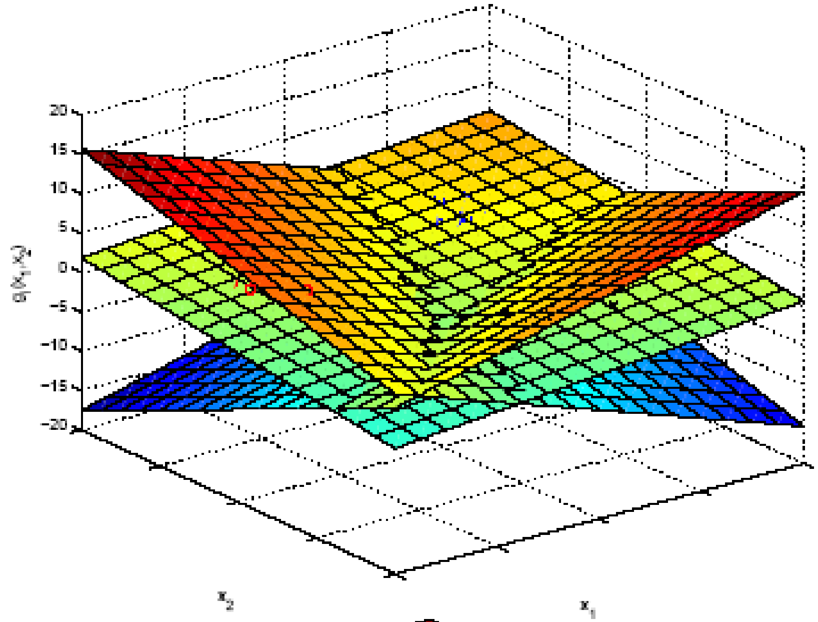
Until convergence

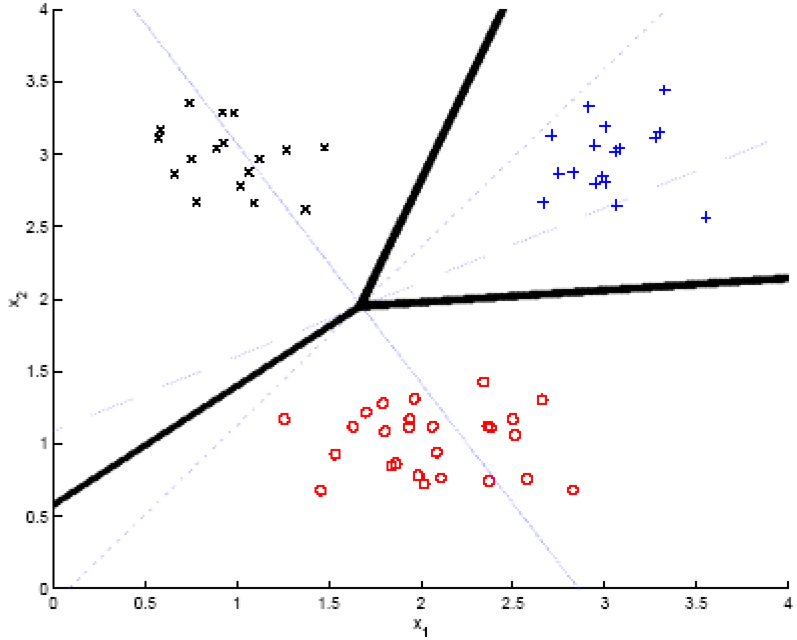
Notes

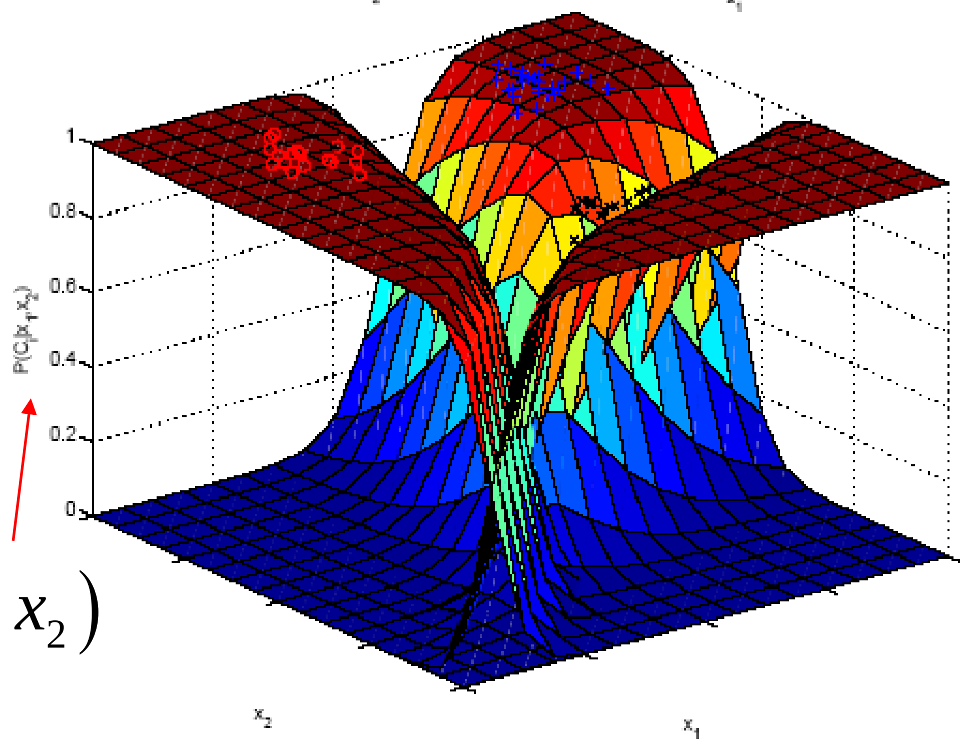
- Note that because of the normalization in softmax, $\boldsymbol{w}_j$ and w_{j0} are affected not only by $\boldsymbol{x}^t \in C_j$ but also by $\boldsymbol{x}^t \in C_i$, $i \neq j$.
- During testing, we calculate all y_k, $k = 1, \ldots, K$ and choose C_i if $y_i = \max_k y_k$.
- We do not need to continue training to minimize cross-entropy as much as possible; we train only until the correct class has the highest weighted sum, and therefore we can stop training earlier by checking the number of misclassifications.

Example

$g_i(x_1, x_2)$







$P(C_i | x_1, x_2)$

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:

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

where $\boldsymbol{\varphi}(\boldsymbol{x})$ are basis functions. Examples:

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

*Discrimination by Regression

- Classes are NOT mutually exclusive and exhaustive

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

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

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

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

$$\Delta \mathbf{w} = \eta \sum_t \left(r^t - y^t\right) y^t \left(1 - y^t\right) \mathbf{x}^t, \quad \Delta w_0 = \eta \sum_t \left(r^t - y^t\right) y^t \left(1 - y^t\right)$$

Learning to Rank

- Ranking: A different problem than classification or regression
- Let us say $\boldsymbol{x}^u$ and $\boldsymbol{x}^v$ are two instances, e.g., two movies.

  We prefer u to v implies that $g(\boldsymbol{x}^u|\boldsymbol{\theta})>g(\boldsymbol{x}^v|\boldsymbol{\theta})$

  where $g(\boldsymbol{x})$ is a score function, here linear:

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

- Find a direction $\boldsymbol{w}$ such that we get the desired ranks when instances are projected along $\boldsymbol{w}$

Ranking Error

- We prefer u to v implies that $g(\mathbf{x}^u)>g(\mathbf{x}^v)$, so error is $g(\mathbf{x}^v)$-$g(\mathbf{x}^u)$, if $g(\mathbf{x}^u)<g(\mathbf{x}^v)$

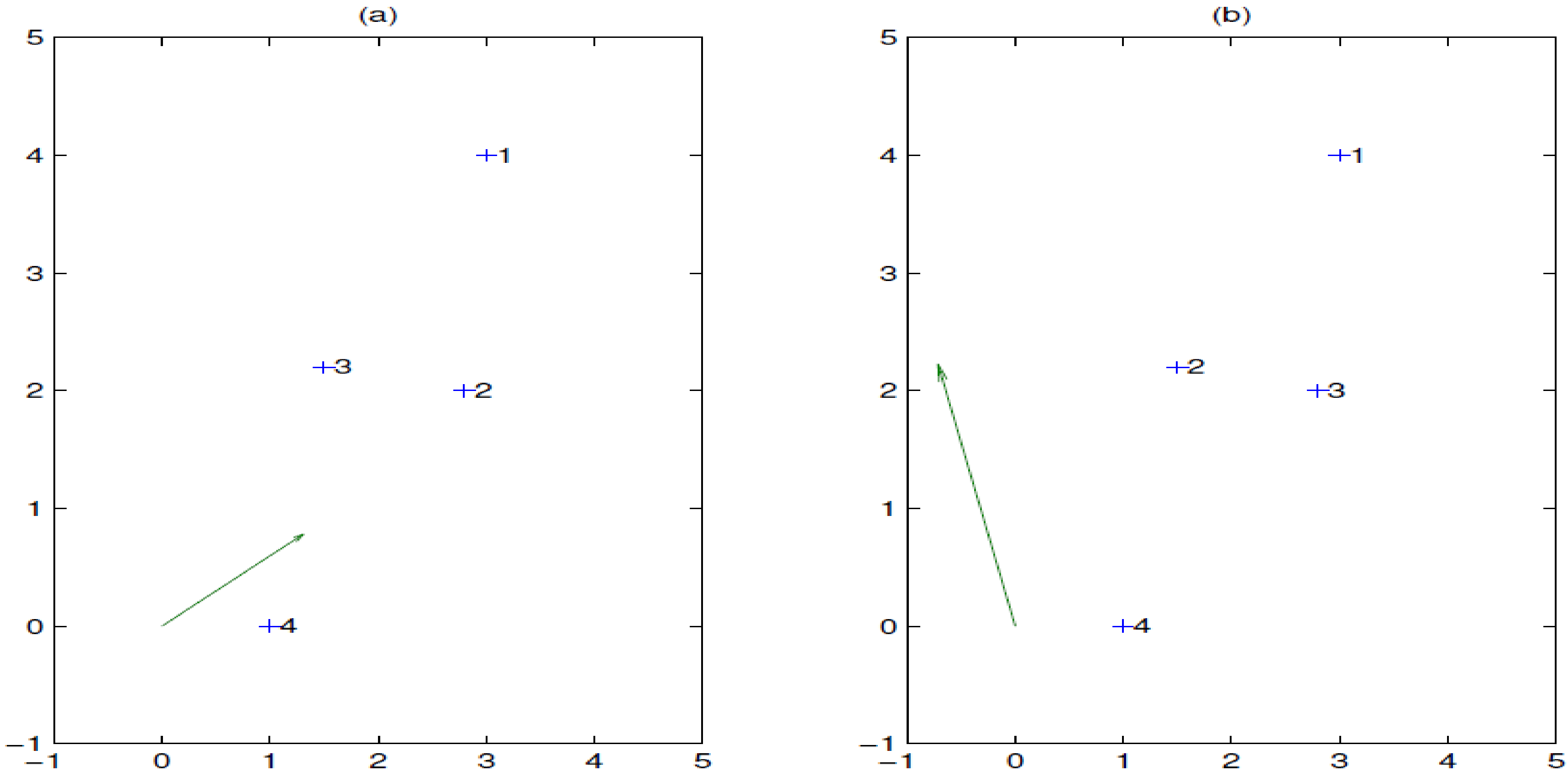
$$E(\mathbf{w}|\{r^u, r^v\}) = \sum_{r^u \prec r^v} [g(\mathbf{x}^v|\theta) - g(\mathbf{x}^u|\theta)]_+$$

where a_+ is equal to a if $a \geq 0$ and 0 otherwise.

$g(\mathbf{x})=\mathbf{w}^T\mathbf{x}$ ⇨ $E(\mathbf{w}|\{r^u, r^v\}) = \sum_{r^u \prec r^v} \mathbf{w}^T(\mathbf{x}^v - \mathbf{x}^u)_+$

For each $r^u \prec r^v$ where $g(\mathbf{x}^v|\boldsymbol{\theta}) > g(\mathbf{x}^u|\boldsymbol{\theta})$, we do a small update:

$$\Delta w_j = -\eta \frac{\partial E}{\partial w_j} = -\eta(\mathbf{x}_j^v - \mathbf{x}_j^u), j = 1, \ldots, d$$



Sample ranking problems and solutions. Data points are indicated by '+' and the numbers next to them indicate the rank where 1 is the highest. We have a full ordering here. The arrow indicate the learned ***w***. In (a) and (b), we see two different ranking problems and the two corresponding solutions.