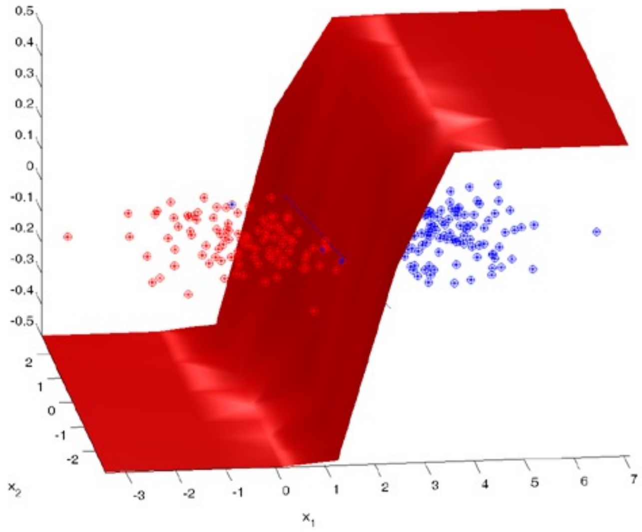
430.457

Introduction to Intelligent Systems

Prof. Songhwai Oh

ECE, SNU

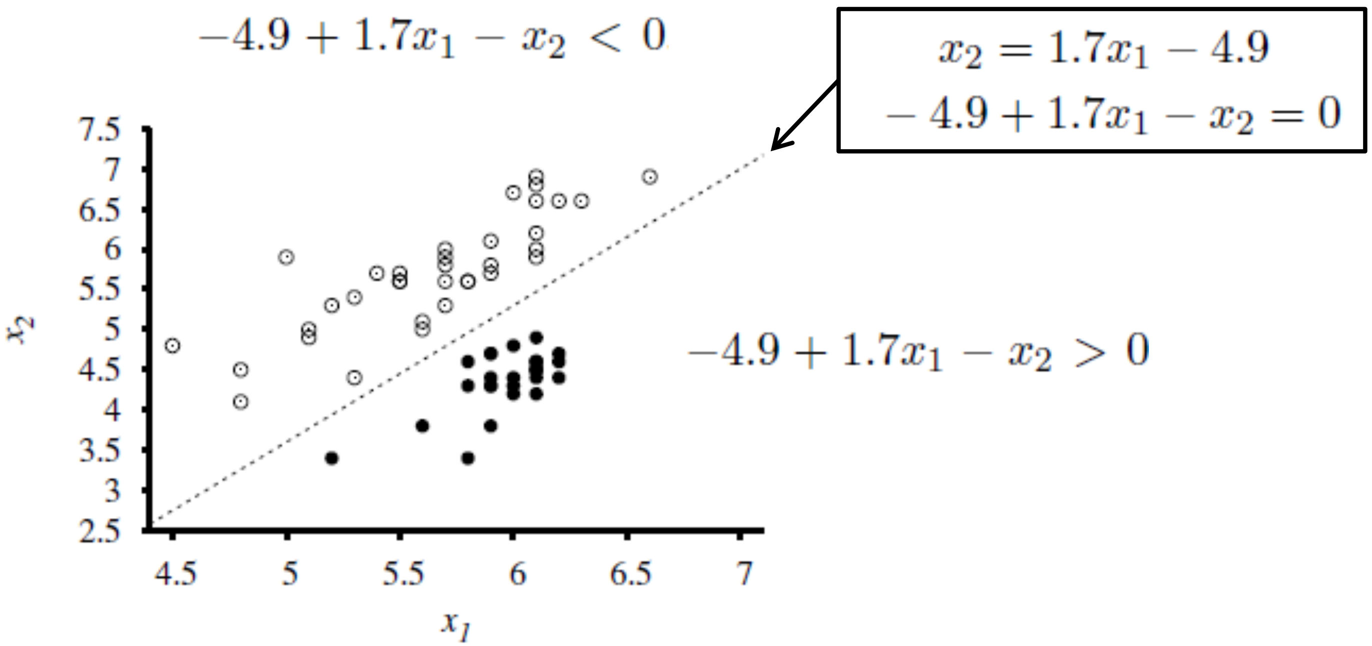
LINEAR CLASSIFICATION



Linear Classifiers

- Example
  - Classes: Earthquakes (0), Underground nuclear explosions (1)
  - Input values: Body wave magnitudes, Surface wave magnitudes
- **Decision boundary**: a line separating two classes.

Linearly Separable Case

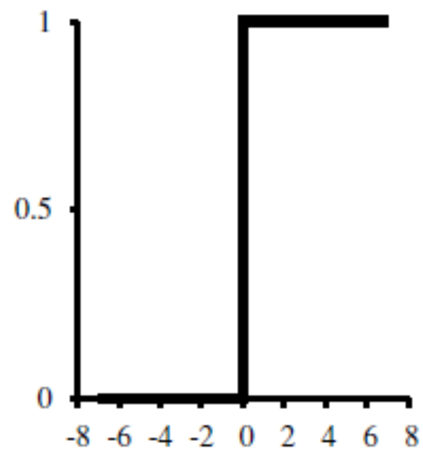


$h_{\mathbf{w}}(\mathbf{x}) = 1$ if $\mathbf{w} \cdot \mathbf{x} \geq 0$ and 0 otherwise.

Perceptron Learning Rule

$h_{\mathbf{w}}(\mathbf{x}) = Threshold(\mathbf{w} \cdot \mathbf{x})$ where $Threshold(z) = 1$ if $z \geq 0$ and 0 otherwise.

Threshold function



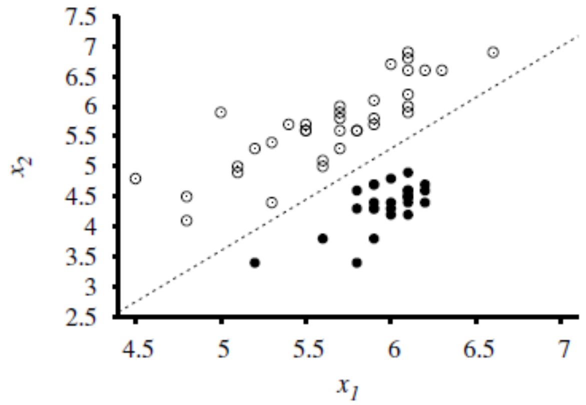
Update rule: $w_i \leftarrow w_i + \alpha\,(y - h_{\mathbf{w}}(\mathbf{x})) \times x_i$

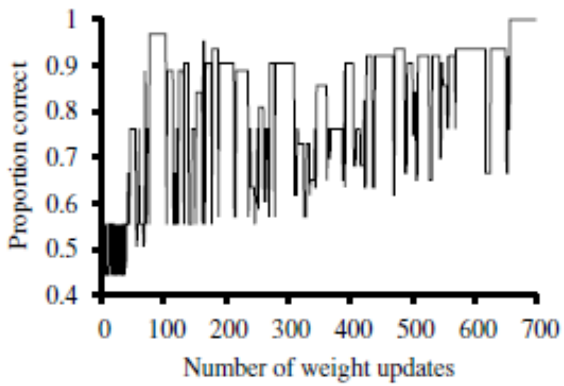
(converges if the problem is linearly separable.)

- If the output is correct, i.e., $y = h_{\mathbf{w}}(\mathbf{x})$, then the weights are not changed.
- If y is 1 but $h_{\mathbf{w}}(\mathbf{x})$ is 0, then w_i is *increased* when the corresponding input x_i is positive and *decreased* when x_i is negative. This makes sense, because we want to make $\mathbf{w} \cdot \mathbf{x}$ bigger so that $h_{\mathbf{w}}(\mathbf{x})$ outputs a 1.
- If y is 0 but $h_{\mathbf{w}}(\mathbf{x})$ is 1, then w_i is *decreased* when the corresponding input x_i is positive and *increased* when x_i is negative. This makes sense, because we want to make $\mathbf{w} \cdot \mathbf{x}$ smaller so that $h_{\mathbf{w}}(\mathbf{x})$ outputs a 0.

Learning Curve

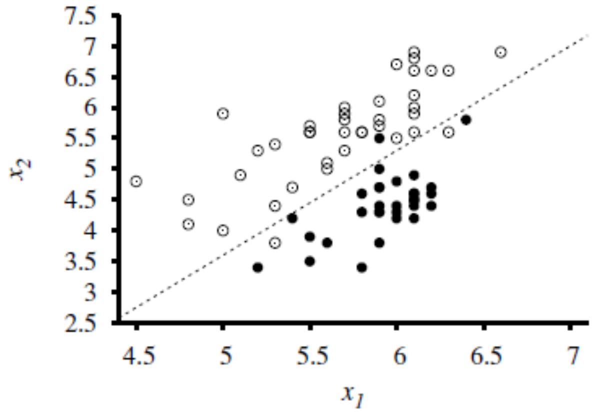
Separable case

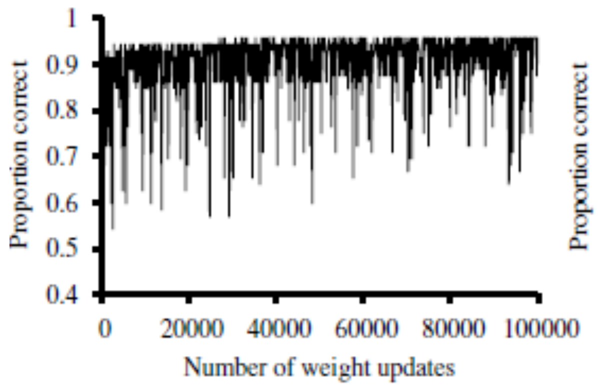




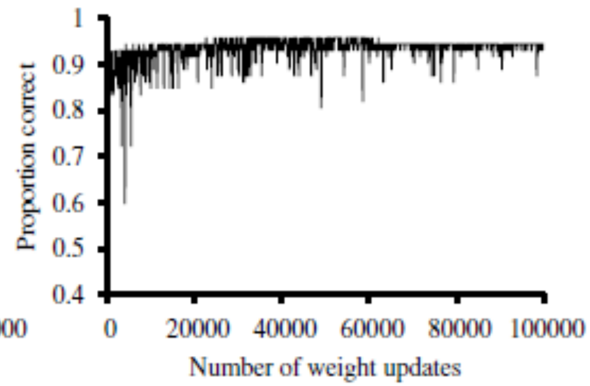
Learning curve

Non-separable case





Learning curve
(constant learning rate)

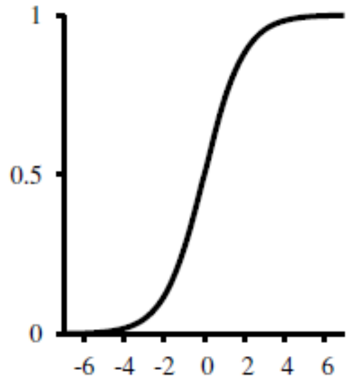


Learning curve
(decreasing learning rate)

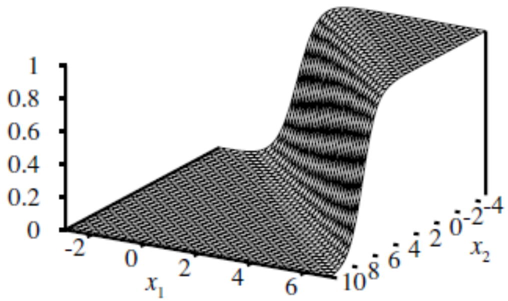
Logistic Regression

Logistic function

$$Logistic(z) = \frac{1}{1+e^{-z}}$$



Soft thresholding



$$h_{\mathbf{w}}(\mathbf{x}) = Logistic(\mathbf{w}\cdot\mathbf{x})$$

Logistic regression

$$h_{\mathbf{w}}(\mathbf{x}) = Logistic(\mathbf{w}\cdot\mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}\cdot\mathbf{x}}}$$

$$\begin{aligned}
\frac{\partial}{\partial w_i}Loss(\mathbf{w}) &= \frac{\partial}{\partial w_i}(y - h_{\mathbf{w}}(\mathbf{x}))^2 \\
&= 2(y - h_{\mathbf{w}}(\mathbf{x})) \times \frac{\partial}{\partial w_i}(y - h_{\mathbf{w}}(\mathbf{x})) \\
\text{(chain rule)} \quad &= -2(y - h_{\mathbf{w}}(\mathbf{x})) \times g'(\mathbf{w}\cdot\mathbf{x}) \times \frac{\partial}{\partial w_i}\mathbf{w}\cdot\mathbf{x} \\
&= -2(y - h_{\mathbf{w}}(\mathbf{x})) \times g'(\mathbf{w}\cdot\mathbf{x}) \times x_i \,.
\end{aligned}$$

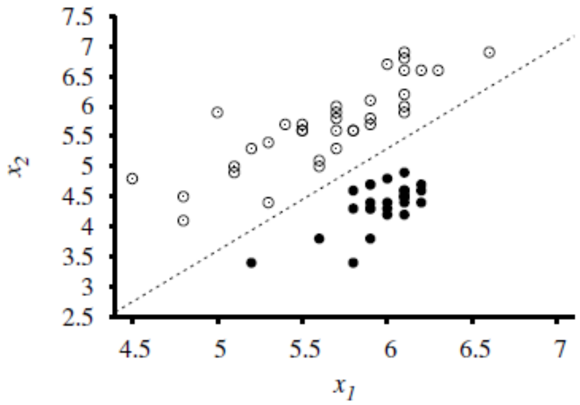
If g is a logistic function, then

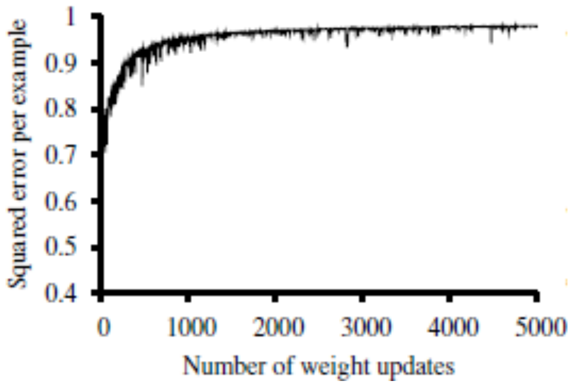
$$g'(z) = \frac{dg}{dz}(z) = g(z)(1-g(z)).$$

$$g'(\mathbf{w}\cdot\mathbf{x}) = g(\mathbf{w}\cdot\mathbf{x})(1-g(\mathbf{w}\cdot\mathbf{x})) = h_{\mathbf{w}}(\mathbf{x})(1-h_{\mathbf{w}}(\mathbf{x}))$$

$$w_i \leftarrow w_i + \alpha\,(y - h_{\mathbf{w}}(\mathbf{x})) \times h_{\mathbf{w}}(\mathbf{x})(1-h_{\mathbf{w}}(\mathbf{x})) \times x_i$$

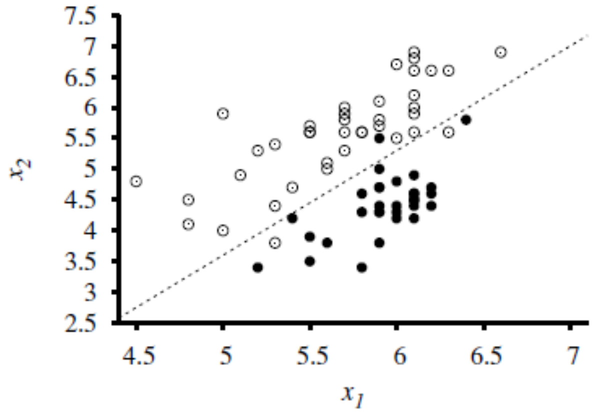
Separable case

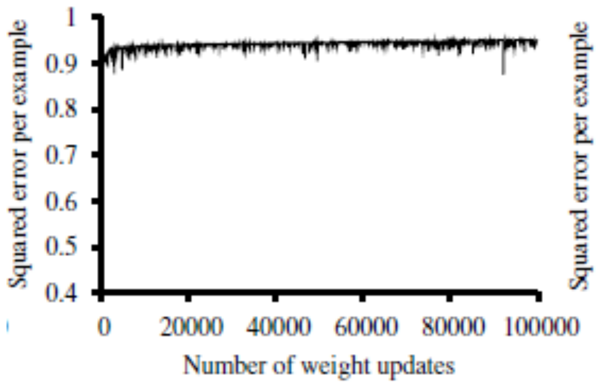




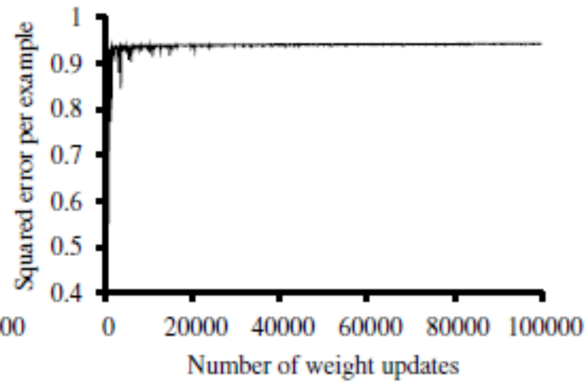
Learning curve

Non-separable case





Learning curve
(constant learning rate)



Learning curve
(decreasing learning rate)